Revealing Stereotypes: Evidence from Immigrants in Schools*

Alberto Alesina[*,**], Michela Carlana[‡], Eliana La Ferrara[§], Paolo Pinotti[¶]

This version: August 2023

Abstract

We study how people change their behavior after being made aware of bias. Teachers in Italian schools give lower grades to immigrant students relative to natives of comparable ability. In two experiments, we reveal to teachers their own stereotypes, measured by an Implicit Association Test (IAT). In the first, we find that learning one's IAT before assigning grades reduces the native-immigrant grade gap. In the second, IAT disclosure and generic debiasing have similar average effects, but there is heterogeneity: teachers with stronger negative stereotypes do not respond to generic debiasing, but change their behavior when informed about their own IAT.

JEL: I24, J15.

Keywords: stereotypes, IAT, bias in grading, immigrants, teachers.

*We thank the Editor, Esther Duflo, two anonymous referees, Daniele Paserman, Sandra Rozo, and seminar participants at various universities and workshops for very useful comments on a previous version of the paper. We are grateful to the schools and teachers that took part in our project, to Gianna Barbieri and Lucia De Fabrizio from MIUR, and to Patrizia Falzetti and Paola Giangiacomo from INVALSI for giving us access to the administrative data used in this paper. Elena De Gioannis, Isabela Duarte, Gaia Gaudenzi, Giulia Tomaselli, and Cristina Perricone provided invaluable help with data collection and data analysis. Carlana acknowledges financial support from the "Policy Design and Evaluation Research in Developing Countries" Initial Training Network (PODER), financed under the Marie Curie Actions of the EU's Seventh Framework Programme (Contract no. 608109). La Ferrara acknowledges financial support from the ERC Advanced Grant "Aspirations, Social Norms and Development" (ASNODEV, Contract no. 694882) under the European Unionâs Horizon 2020 research and innovation programme. The experiment was registered in the AEA RCT Registry (AEARCTR-0003647).

**Harvard University, IGIER, NBER, and CEPR

‡Harvard Kennedy School, IZA, LEAP, NBER, and CEPR (email: michela_carlana@hks.harvard.edu)

§Harvard Kennedy School, LEAP, NBER and CEPR (email: elaferrara@hks.harvard.edu)

¶Bocconi University, DONDENA, CReAM, and CEPR (email: paolo.pinotti@unibocconi.it)

1 Introduction

Economists have studied discrimination toward minority groups since at least Becker (1957) and have more recently discussed how biased judgment toward specific individuals may be induced by stereotypes (Bordalo et al., 2016; Bertrand and Duflo, 2017).[1] Stereotypes can be thought of as over-generalized representations of characteristics of certain groups that allow for an easy and efficient processing of information, but they may also lead to self-fulfilling prophecies by influencing the behavior of stigmatized groups. Individuals exposed to negative stereotyping toward their own group may experience reduced effort, self-confidence, and productivity (Glover et al., 2017; Carlana, 2019). Several organizations —including universities, corporations, and police departments (especially in the U.S. and Canada)— are currently promoting interventions to mitigate discriminatory behavior by increasing employees' awareness of their implicit stereotypes.[2] However, there is limited causal evidence on the success of these policies (Bohnet, 2016; Lai et al., 2013).

We study this problem in a context where the detrimental effects of stereotyping are particularly serious: children exposed to teachers' stereotypes may be discouraged from investing in human capital (Rosenthal and Jacobson, 1968; Papageorge et al., 2020; Carlana, 2019). Immigrant students in Italian schools receive lower grades from their teachers compared to native students with the same performance in standardized tests, and we relate this gap in grades to teachers' implicit stereotypes. We experimentally evaluate the effects of revealing to teachers their own stereotypes and find that this leads to a change in their grading behavior, reducing the immigrant-native gap compared to a group of teachers who do not receive any information (pure control). In a second experiment, we delve deeper into the mechanisms behind the change to understand the importance of learning about *one's own* bias relative to learning about bias in general. We do not find differences on average, but the effects are heterogeneous: teachers with stronger negative stereotypes adjust their behavior when they are informed about their own bias, but do not react to a generic debiasing message informing them of the presence of bias toward immigrants in society and in schools.

The case of Italy is interesting for at least two reasons. First, mass immigration is a relatively recent phenomenon, and Italy has experienced one of the highest increases in the share of immi-

[1] A non-exhaustive list of papers addresses discrimination by employers (Bertrand and Mullainathan, 2004), police officers (Fryer Jr, 2019; Coviello and Persico, 2015; Knowles et al., 2001), referees (Price and Wolfers, 2010), courts (Dobbie et al., 2018; Alesina and La Ferrara, 2014), and teachers (Figlio, 2005; Botelho et al., 2015). For a review of theoretical and experimental results, see Altonji and Blank (1999) and Bertrand and Duflo (2017).

[2] For example, employees are advised to take an Implicit Association Test to increase awareness about one's implicit associations on race and gender. Among others, Harvard University strongly encourages "every search committee member to take at least one Implicit Association Test (IAT)" (https://faculty.harvard.edu/recruitment-best-practices), and Starbucks has promoted a "racial bias training" for all employees (https://stories.starbucks.com/press/2018/starbucks-to-close-stores-nationwide-for-racial-bias-education-may-29/).

grants over the past few years, which has fueled anti-immigrant sentiments (Alesina et al., 2022). Second, in the Italian education system, middle school is a critical juncture, at the end of which students get tracked into different types of high schools, which affects their future education and work prospects (Carlana et al., 2022a). This type of tracking is similar to that of most other European countries. The attitudes of middle school teachers could thus have important long-term effects on students' educational and professional careers.

We use two unique datasets. The first combines administrative data on students with original survey data from a sample of over 1,300 teachers in Northern Italy, collected in person during a field experiment. The second dataset includes survey data from a sample of around 200 middle school teachers, collected online and embedded into a lab-in-the-field experiment where teachers were required to evaluate students' tests.

In both datasets, we measure teachers' stereotypes using the Implicit Association Test (IAT). This is a computer-based tool developed by social psychologists, designed to minimize the risk of social desirability bias in self-reported answers (Greenwald and Banaji, 1995). Previous research in social psychology has highlighted a number of limitations including weak predictive power and potential manipulation (Blanton et al., 2009; Oswald et al., 2013; Olson and Fazio, 2004; Fiedler and Bluemke, 2005; Cvencek et al., 2010). Despite that, it is increasingly used by social scientists to measure stereotypes both in the lab and in the field (Rooth, 2010; Bertrand and Duflo, 2017; Corno et al., 2022; Glover et al., 2017; Reuben et al., 2014). Our IAT data shows that teachers generally hold strong, negative stereotypes toward immigrant students. According to the metrics proposed by Greenwald et al. (2009), around 70 percent of teachers in our field experiment and 80 percent in our online experiment exhibit "moderate to severe" stereotypes, and almost all of them exhibit some degree of negative stereotypes toward immigrants.

As a first step in the analysis, we establish that, holding constant the performance on standardized blindly graded tests, immigrant students receive lower grades than natives when they are graded by their teachers in a non-blind way. Furthermore, we correlate bias in grading with teachers' IAT and find that higher IAT scores —indicating more negative stereotypes toward immigrants— are associated with lower grades to immigrant students at the high end of the distribution, i.e., for high-performing immigrants. Teachers' IATs are uncorrelated with the grades given to native students.

We then move to the main contribution of the paper, that is, the impact of revealing stereotypes. We conducted two experiments. The first is a field experiment in around 100 schools, where we randomized the timing of the feedback on own IAT. In half of the schools (randomly selected), teachers were informed of their IAT score shortly before end-of-semester grading, while in the remaining half they were informed shortly after. We find that teachers who received their IAT score

'early' gave higher grades to immigrant students relative to native ones.

Our second experiment was run online with a different group of Italian middle school teachers and complements the first one in two ways. First, it allows us to test whether learning about one's own bias has any additional effect relative to learning about generic bias in society; second, we gain a better understanding of individuals' updating process regarding their own bias. As in the field experiment, teachers in the online sample took an IAT measuring stereotypes against immigrants. Immediately after they took the test, we asked them to predict their own stereotypes, and then provided feedback on IAT scores to only half of the teachers, randomly selected. Different from our field experiment, all teachers received a generic debiasing message. A few weeks after this initial session, we asked teachers to grade 10 tests, randomly ascribed to native-sounding or immigrant-sounding student names. The grade given to these tests is used as the main outcome for the online experiment. We find that, on average, the personalized feedback does not increase grades assigned to immigrants vs. natives, relative to generic debiasing. However, there are significant heterogeneous treatment effects: teachers with stronger implicit bias do not react to generic debiasing, but they decrease their gap in grading when provided information on their own IAT. The effect is driven by teachers who did not expect the feedback they received, hence those who update based on new information. This is consistent with the possibility that at least part of the bias in grading was due to people being unaware of their implicit bias.

Overall, the results of our experiments suggest two things. First, from the field experiment, we learn that providing individualized feedback works on average compared to the status quo (no feedback). This is important for policy and it could be applied at scale. Second, our online experiment suggests that IAT feedback does not work better than a generic debiasing message for teachers with mildly negative stereotypes, but it works significantly better for teachers with strongly negative stereotypes. Depending on whether the policymaker's objective is to correct the strongest biases or to work on other parts of the distribution, the IAT feedback may or may not be preferred to generic debiasing.

Our work is related to several strands of literature. First, we contribute to the recent economics literature emphasizing the importance of considering implicit bias when analyzing discriminatory behavior (Avitzour et al., 2020; Bursztyn et al., 2021; Van den Bergh et al., 2010; Guryan and Charles, 2013; Corno et al., 2022; Bertrand and Duflo, 2017). Glover et al. (2017) provide evidence that exposure to managers with stronger implicit bias negatively affects the work performance of minorities. Reuben et al. (2014) show in a lab experiment that the gender-science IAT predicts employers' biased expectations against women, while Carlana (2019) shows that teachers' stereotypes affect the gender gap in math, track choice, and self-confidence in math for girls in middle school. Research in social psychology and medicine has examined individuals' emotional responses when

provided feedback about their own implicit associations, showing that people tend to react defensively —for instance, by questioning the validity of the IAT (O'Brien et al., 2010; Howell et al., 2015; Sukhera et al., 2018). However, none of these papers investigates whether revealing people's own stereotypes to themselves has an impact on discriminatory behavior toward others. In this respect, our paper also differs from Pope et al. (2018), who show that racial bias among professional basketball referees disappears after the media calls attention to the results of an academic study highlighting bias. In their study, referees learn about existing bias in the profession, not about their own.

We also contribute to the literature on teacher bias, which finds that teachers' expectations are often biased against minority students. This behavior may lead to a self-fulfilling prophecy, with students internalizing negative expectations and ultimately behaving in the direction predicted by the biased beliefs (Papageorge et al., 2020; Jussim and Harber, 2005; Rosenthal and Jacobson, 1968). A few previous studies compare teacher-assigned (non-blind) grades and standardized (blind) test scores across minority and non-minority students (Botelho et al., 2015; Burgess and Greaves, 2013; Hanna and Linden, 2012; Van Ewijk, 2011) and across genders (Lavy and Sand, 2018; Lavy, 2008; Lavy and Megalokonomou, 2019; Terrier, 2020). We add to this literature by using the IAT as a direct measure of teachers' stereotypes, which allows us to trace a stronger link between grading gaps and bias even in the presence of unobserved characteristics that may lead students to perform differentially in blindly versus non-blindly graded tests. Furthermore, none of the above papers tests the effectiveness of remedial interventions to mitigate bias in the schooling context.

Finally, our results speak to a recent literature on how to reduce bias. Lai et al. (2014) underline that interventions providing counter-stereotypical exemplars and strategies to override biases are the most effective in reducing implicit racial prejudice. The impact of diversity training on behavior change is discussed by Chang et al. (2019), while recent awareness-raising campaigns on gender bias have been studied by Boring and Philippe (2021), Mengel (2021), Carnes et al. (2015), and Devine et al. (2017). The interventions on bias mitigation toward immigrants have mainly focused on providing information about immigrants and have estimated the effect on attitudes (Grigorieff et al., 2018; Hopkins et al., 2019) and support for immigration policies (Facchini et al., 2022; Alesina et al., 2022). An additional group of interventions focuses on promoting inter-group contact (Allport, 1958; Lowe, 2021; Corno et al., 2022). As suggested in a recent meta-analysis by Paluck et al. (2018), "the absence of studies addressing adults' racial or ethnic prejudices [is] an important limitation for both theory and policy."

The remainder of the paper is organized as follows. In Section 2 we provide background information on the grading system in Italian middle schools. Section 3 describes our data and the

experimental design, and Section 4 contains descriptive evidence on implicit stereotypes and grading. Section 5 presents our results, and the last section concludes.

2 Institutional background

2.1 The Italian schooling system

Education in Italy is free for all children and is compulsory between the ages of 6 to 16. The schooling system is organized as five years of primary school, three years of middle school, and five years of high school. Students are assigned to the same class for all subjects, and they interact with the same set of peers within each type of school. In middle school, which comprises grades 6 to 8, students are usually taught by the same teachers for all three years, and they spend at least six hours per week with the math teacher and five hours with the literature and grammar teacher. Teachers are assigned to schools by the Italian Ministry of Education, and their allocation is determined by seniority: teachers with more experience can teach at schools that are higher in their preference ranking and tend to work close to their hometown and away from disadvantaged areas (Barbieri et al., 2011). Students are assessed continuously with written and oral exams in each subject, and they receive end-of-semester grades in January and June. These "final" grades are discrete variables ranging between 3 and 10, with 6 being the pass grade. Thus, end-of-semester grades may incorporate significant discretion of the teachers.

In addition to teacher evaluations, standardized tests in math and reading are administered by the National Institute for the Evaluation of the Education and Training System (INVALSI) to all Italian students at the end of middle school (grade 8). INVALSI tests mainly consist of multiple choice questions or short answers, which are blindly graded following a precise evaluation grid.

At the end of middle school, students must choose between three high school tracks: academic oriented (*liceo*), technical, and vocational. Academic and technical schools offer significantly better educational and employment prospects than vocational schools (Carlana et al., 2022a).

2.2 Immigrants in Italian schools

In the last two decades, the share of immigrant children (i.e., children without an Italian citizenship) in Italian schools has increased from less than 1% in 1998 to 10% in 2018, with a higher concentration in the northern part of the country and big cities. Immigrant students come from diverse geographic backgrounds, with the most represented nationalities being Romanian, Albanian, Moroccan, Chinese, Filipino, and Indian (see Appendix Table A.1). Currently, about 65%

of immigrant children are born in Italy, but they can obtain Italian citizenship only after turning 18 and are subject to rather stringent conditions.[3] Throughout the paper, immigrant students are defined according to their citizenship: they include first-generation students born abroad and second-generation students born in Italy from parents who are not Italian citizens.

Immigrant students have, on average, lower performance than native students in Italian schools (Carlana et al., 2022a), and the same is true in most other destination countries (OECD, 2014). Of course, this may at least in part reflect language barriers and parental investment given that, on average, they typically come from a lower socioeconomic background, but it may also partly reflect discrimination by teachers.

3 Data and experimental design

3.1 The IAT

We measure implicit stereotypes toward immigrants using the Implicit Association Test (IAT). The test requires categorizing words to the left or to the right of a computer screen, and it measures the strength of the association between two concepts based on response times. The underlying idea, as conceived by Donders (1969) and Greenwald et al. (1998), is that the easier the mental task, the faster the response production.

The version of the IAT that we developed for our study requires associating immigrant and native names (e.g., Fatima and Francesca) with positive and negative adjectives in the schooling context (e.g., smart and lazy). Labels and categories are in the top corners of the screen, names and adjectives randomly appear at the center of the screen, and subjects are asked to categorize the words as quickly as possible. If respondents hold negative stereotypes against immigrants, they should react more slowly when the label "immigrant" is associated with positive adjectives compared to when it is associated with negative ones, because positive associations are less natural to them. The IAT measures stereotypes using the difference in reaction times between rounds in which native-sounding names and negative adjectives appear on the same side of the screen and rounds in which immigrant-sounding names and negative adjectives appear on the same side.

Starting from the continuous IAT score produced by the test, one can define a categorical measure based on conventional thresholds recommended by Greenwald et al. (2009). In particular, the

[3] Like most other European countries—and unlike the United States—Italy follows the principle of *ius sanguinis*; i.e., citizenship is determined by the nationality of one's parents. There is a limited time window (one year) to apply for Italian citizenship after turning 18, and the candidate citizen must be able to prove continuous residence in Italy during the previous years.

negative association with immigrant names is absent when the IAT score is positive but below 0.15, "slight" when it is between 0.15 and 0.35, and "moderate to severe" when it is above 0.35. Negative values of these same thresholds define the strength of positive associations.

In the field experiment, each teacher in our survey completed two immigrant-native IATs, one using male names and one using female names, and the order of the IAT with male and female names was randomized at the individual level. In the online experiment, we administered the IAT using a mix of male and female names of immigrant and native students. This allowed us to minimize the duration of the baseline survey (a key aspect given the online setting) and to calculate only one IAT score per teacher. Further details on the IATs that we administered are available in Online Appendix B.1.

While the IAT is widely used (Green et al., 2007; Arcuri et al., 2008; Nosek et al., 2009; Monteith et al., 2001), previous research in social psychology has highlighted a number of limitations (Olson and Fazio, 2004). First, some argue that the IAT has a weak predictive power (Blanton et al., 2009; Oswald et al., 2013; Meissner et al., 2019) and, in particular, that it does not predict behavior better than explicit measures (Axt et al., 2020; Schimmack, 2021). However, most of these studies refer to experiments with a limited number of subjects and do not have information outside the lab on whether individuals with stronger implicit associations are actually biased in their interactions. Recent papers in economics have shown correlations of IAT scores with real-world behavior, including call-back rates of job applicants (Rooth, 2010), job performance of minorities (Glover et al., 2017), and teachers' track recommendations (Carlana et al., 2022b).

The second main concern with the IAT is that subjects may fake the test by voluntarily slowing down or speeding up on specific blocks or strategically increasing errors (Fiedler and Bluemke, 2005; Cvencek et al., 2010). However, this type of manipulation would require a deep knowledge of the test, which is unlikely within our sample of teachers, as the IAT is not widely known in Italy. Furthermore, the improved scoring algorithm that we use (Greenwald et al., 2003) discards observations characterized by abnormal reaction times.

Third, some researchers argue that the IAT measures social constructs such as salience of attributes (Rothermund and Wentura, 2004), familiarity with the concepts it quantifies, and, more generally, cultural stereotypes rather than "personal animus" (Arkes and Tetlock, 2004; Karpinski and Hilton, 2001; Mitchell and Tetlock, 2017; Tetlock and Mitchell, 2009). However, these possibilities have been addressed empirically (Nosek and Hansen, 2008; Olson and Fazio, 2004; Ottaway et al., 2001; Rudman et al., 1999; Dasgupta and Greenwald, 2001), and past research in social psychology suggests there is no reason why familiarity and attitudinal evaluation should be unrelated since familiarity breeds liking (Jost, 2019).

A fourth concern is that the IAT may capture unstable characteristics that vary over time (Das-

gupta and Greenwald, 2001; Bar-Anan and Nosek, 2014; Gawronski et al., 2017). However, social psychology theory establishes that attitudes are intrinsically dynamic (Banaji, 2004; Hardin and Banaji, 2013). Moreover, the IAT exhibits a higher (within-person) test-retest reliability than other response-latency measures commonly used in psychological research, including Stroop and priming tasks (Bar-Anan and Nosek, 2014; Jost, 2019).

Overall, we acknowledge that the IAT may be a noisy measure of stereotypes, but it has the advantage of (i) avoiding social desirability bias present in explicit responses on socially sensitive topics (Greenwald et al., 2009) and (ii) capturing implicit associations that may be unknown to the individual but may nevertheless affect their interaction with stigmatized groups (Bertrand et al., 2005).

3.2 The field experiment

In our first experiment, we administered an IAT to a large sample of teachers of grade 8 students and revealed to half of them their own IAT score just before end-of-semester grading, while the other half received the same information shortly after end-of-semester grading. We then compared the grades given to immigrants and natives between the two groups of teachers.

The experiment took place in five large cities of Northern Italy —Milan, Brescia, Padua, Genoa, and Turin— during the first part of the 2016/2017 school year. In September 2016, all middle schools in these cities enrolling at least 20 immigrant students in grade 6 (as of 2012) were invited to participate to a survey titled "The role of teachers in high school track choice." We intentionally avoided mentioning immigrants and immigration-related issues to prevent sample selection on attitudes toward immigration. Out of 145 schools invited, 102 accepted to take part in the project.[4]

The survey was addressed to all math and literature teachers in grade 8, and it consisted of two parts. In the first part, teachers completed two immigrant-native IATs, one with male names and one with female names, as described in the previous section. In what follows, we use the average of the two.[5] The second part of the questionnaire elicited information on respondents'

[4]It is useful to discuss if and how these 102 schools differ in terms of student and teacher characteristics. We cannot provide balance tables of the characteristics of students in the 102 schools compared to the 43 schools that did not participate, as we do not have the code to identify those 43 schools from the pseudo-anonymized dataset of Italian schools. However, in Appendix Table A.2 we compare students in our experimental sample with all students in Italian schools (column 1) and all other students in the selected provinces (column 2). Schools in our sample are comparable in terms of gender composition but have a higher share of immigrants than other schools, as should be expected given the selection criteria for our study. This also implies some differences in socioeconomic characteristics correlated with immigrant status; however, the standardized differences are very small for all variables.

[5]The correlation between the two continuous IAT scores is 0.28. However, based on the categories we communicated to teachers, 76% of them received a consistent message (either biased in both or unbiased in both): despite the noise in the measurement, the test is accurately capturing individuals' implicit associations between immigrants/natives

socioeconomic characteristics, teaching experience, explicit bias toward immigrants, and criteria followed to advise students on high school track choice.[6]

On average, 80% of the teachers in our 102 schools completed the survey, yielding a sample of 1,384 teachers. This is the main sample used for estimating the relationship between teachers' IAT and grading of immigrant students. To this purpose, we obtained both teacher-assigned grades and standardized test scores in grade 8 for all students taught by these teachers between school years 2011/12 and 2016/17.

Within the 102 schools, 65 schools comprising 533 teachers in grade 8 were surveyed before the end of the first semester (i.e., end of January) due to logistical reasons and are therefore included in the experimental sample, while others were interviewed after the end of January.[7] This is the experimental sample used for estimating the effect of revealing IAT scores on grading behavior. We offered to all teachers in this sub-sample of schools the possibility of receiving feedback on their IAT score, and more than 80% of teachers chose to receive it. Appendix Table A.5 shows that there is no significant correlation between the decision to receive the feedback and several teacher characteristics, including implicit or explicit biases against immigrants.[8]

Feedback was provided over email. Teachers received a brief description of the IAT and were told whether their association between immigrant names and good/bad adjectives was "slight," "moderate," or "strong" based on the thresholds identified by Greenwald et al. (2009) and discussed in Section 3.1. Each teacher received their score from two IATs: one using male names of natives and immigrants and one using female names. Teachers were assured that these results would not be shared with anyone. The detailed text of the email is reported in Appendix B.3.

We randomized the timing of the feedback across schools. In half of the schools ("treatment") teachers received the feedback before end-of-semester grading, i.e., by the end of January 2017.

and positive/negative adjectives. Teachers also completed a gender-science IAT (see Carlana, 2019) that we do not use in this paper. The order of the IATs was randomized across individuals, but the two immigrant-native IATs were always presented one after the other. The correlation between each immigrant-native IAT and the gender-science IAT is lower (0.06) compared to the correlation between the two immigrant-native IATs (0.28), suggesting that the IAT is not merely capturing the ability to complete the test.

[6]The questionnaire was administered during meetings held in school buildings. Our enumerators gave each teacher one tablet to complete the survey autonomously but remained available in the room to answer questions or help with tablets if requested. Teachers who agreed to take part in the survey gave written informed consent. The time to complete the survey was around 30 minutes, and participants did not receive any compensation.

[7]Appendix Table A.3 compares the two sets of teachers, while Appendix Table A.4 compares the characteristics of their students. In both cases, the two groups are comparable.

[8]Instead, there is a significant correlation with how much "in a hurry" the respondent was. A survey completion time of more than 20 minutes (33% above the mean) is associated with a 5 percentage point increase in the probability of consenting to receive the feedback. Similarly, those who completed only the IAT and not the rest of the survey were almost 10 percentage points less likely to request feedback. These correlations do not survive when including school fixed effects, which explains a substantial share of the variation in the choice of receiving feedback, as shown by the R-squared in Table A.5.

In the remaining schools ("control") teachers received the feedback within two weeks after end-of-semester grading. This implies that all teachers (in both the treated and control groups) learned about their IAT by mid-February, which prevents us from studying the long-term impact of our intervention. We chose to randomize at the school level, rather than at the teacher level, to avoid contamination.

Leveraging the randomization, we can estimate the effect of revealing IAT on grading behavior by comparing the grades assigned by teachers in the treated and control groups to immigrant and native students. The grades given by teachers at the end of the first semester are normally the arithmetic mean of previously assigned scores in written and oral exams, where teachers have substantial power to decide whether to round the score up or down. We expect that our intervention may affect this discretionary choice of the teacher. We used grades available from administrative registries so that teachers were unaware that we could observe their grades, thus reducing the risk of experimenter demand effects.

Figure 1 illustrates the timeline of the survey and experiment, as well as the periods covered by the data on standardized test scores and teacher-assigned grades. Note that when we study the role of teacher stereotypes in grading, we use *end-of-year* grades (i.e., in June) as these are contemporaneous to the (blindly graded) INVALSI test scores in grade 8, which are essential for this type of analysis.[9] In contrast, when we estimate the effect of revealing IAT scores, we use *end-of-semester* grades (i.e., in January), for ethical reasons: these grades are not decisive for students' careers, and hence we minimize the possibility of our intervention harming students' outcomes. Unfortunately, INVALSI tests are *not* administered mid-year, which implies that in analyzing the impact of our experiment we cannot control for the INVALSI score. This, however, does not affect our ability to estimate the effects of the intervention, given randomization.

[Insert Figure 1]

3.3 The online experiment

Our second experiment aims to isolate the effect of the *unexpected* component of bias revelation and to compare the effects of revealing one's own bias to those of a more generic debiasing message. From December to January 2021, we invited 595 teachers to an online survey, which was completed by 179 teachers from 74 different schools.[10] This baseline survey included the

[9] Note that knowledge of our study could not affect the behavior of teachers toward the cohorts of children used for this part of the analysis given that they graduated from middle school before our data collection.

[10] The pool of teachers we invited were part of a separate data collection for the Tutoring Online Program (TOP) described in (Carlana and La Ferrara, 2021). Teachers received the invitation upon completing the endline survey

immigrant-native IAT described in Section 3.1, together with a short questionnaire collecting basic demographic characteristics. After having completed the IAT, participants were asked whether they expected to have no bias against immigrants or a "slight," "moderate," or "strong" bias. We classify respondents as *underestimating* their own bias whenever this self-assessment is lower than the classification based on their actual IAT score, using the thresholds defined by Greenwald et al. (2009) and discussed in Section 3.1.

After teachers completed the baseline survey, we randomized them into two groups, at the school level. The first group ("active control") comprised of 88 teachers who received a generic debiasing message, with information on implicit biases in society and their potential negative impact on students. The second group ("treatment") comprised of 91 teachers who received the generic debiasing message plus information on their own IAT score.[11] The detailed content of the two messages is reported in Online Appendix B.3.2.

Teachers received the debiasing message and, if applicable, their IAT score by email at the end of January 2021. Approximately three weeks later, we contacted them again and asked them to grade 10 short tests in their subject (alternatively, math, literature, or English). We randomized across teachers the name of the student reported in each answer, between typical immigrant names (two tests with female names, two with male ones) and typical native names (three tests with female names, three with male ones). The tests were prepared by consultants hired by our team who were teachers in middle schools outside our sample. The same consultants provided sample answers of varying quality, corresponding to different test grades. These answers are the ones that were submitted for grading to the teachers in our online experiment.[12] Online Appendix Figure A.1 shows that there is a very high correlation between the intended grade according to the consultants who prepared the tests and the average grade assigned by the teachers who participated in our experiment.

of TOP, with the following recruitment message: "Are you interested in completing a survey for another research project and getting a thank you voucher of 40 euros? This is for a project completely independent from TOP, aimed at understanding the way in which teachers grade assignments. Your participation in this second research will not affect the participation in the TOP program in the future. We expect that the second research project will require a total of 45 minutes, divided in two moments."

[11] Ideally, we would have liked to implement the experiment with three arms: Generic debiasing + IAT feedback, Generic debiasing, and Pure control. However, given the difficulties in recruiting a large enough number of teachers for the online experiment during the pandemic, we decided to prioritize the comparison between generic debiasing and IAT revelation and we included only two treatment arms.

[12] Some examples of test questions and answers are available in Online Appendix B.4. The incomplete disclosure of the fictitious exams and names during the experiment did not have more than minimal risk for teachers. After the experiment, following the IRB protocol, teachers were informed with a debriefing message on the detailed purpose and incomplete disclosure of the experiment.

3.4 Descriptive statistics

3.4.1 IAT score and teacher characteristics

Figure 2 plots the distribution of the IAT score across teachers in the field and in the online experiments (Panels A and B, respectively). The vast majority of teachers have negative stereotypes toward immigrants (i.e., an IAT score greater than 0.15), with no relevant differences between literature and math teachers. About 80% of teachers in the online experiment exhibit strong stereotypes (i.e., IAT score greater than 0.35), compared to 67% in the field experiment.

[Insert Figure 2]

In addition, the last row of Table 1, Panel B shows that 80% of teachers in the online experiment underestimate their biases.[13] In general, teachers in the online experiment exhibit a higher average IAT compared to participants to the field experiment: 0.70 and 0.48, respectively.[14] One potential explanation for this difference is related to the different timing and implementation of the test —in person before the COVID-19 pandemic for the field experiment and remotely during the pandemic for the online experiment.

[Insert Table 1]

Most importantly, for our purposes, the first row of Panels A and B in Table 1 shows that average IAT scores are balanced between the treated and control groups within each experiment, as one would expect given randomization. The remaining rows of the table show that other observable characteristics are also balanced between the two groups.[15] Not only are the differences not statistically significant, but in all cases the normalized difference (column 5) remains below the threshold of 0.25, as recommended by Imbens and Rubin (2015).[16]

Table 2 shows the correlation between the IAT score and other teacher characteristics, both for the field experiment (Panel A) and for the online experiment (Panel B). The correlation between gender, place of birth, and working experience is small and generally non-significant (columns

[13] In the field experiment we did not elicit teachers' priors about their IAT.

[14] The mean IAT score in our experiments is slightly higher than the mean of 0.41 in the sample of Italians who decided to take the race IAT online on the website `https://implicit.harvard.edu`.

[15] In table A.7, there is only one student-level observation if both math and literature teachers of the same student participate in the experiment. For this reason, the sample includes 6,050 students, while in the analysis we will include student by teacher observations. The sample is balanced also when considering separately the students whose math teacher participate in the experiment (85% of the student sample) and the students whose literature teacher participate in the experiment (85% of the student sample).

[16] The formula for the normalized difference is $\Delta = \frac{\overline{X}_T - \overline{X}_C}{(\sqrt{S_T^2 + S_C^2})/2}$, where $\overline{X}_T$ and $\overline{X}_C$ are the means of covariate X in the treated and control group, respectively, and S_T^2 and S_C^2 are the corresponding sample variances of X.

1–3). On the other hand, there is a significant correlation between IAT and explicit beliefs about immigrants, as measured by a question asking whether immigrants and natives should have equal opportunities to access available jobs (the variable "WVS Immigrants' Rights to Job" in the table, as a similar question is routinely included in the World Values Survey). Column 4 shows that respondents who agree with this statement have significantly less negative implicit stereotypes against immigrants.

[Insert Table 2]

In columns 5 and 6 of Panel A, we test whether teachers' stereotypes reflect the relative ability of native and immigrant students to whom teachers were previously exposed. For this purpose, we collected the standardized test scores (INVALSI) of the students taught by teachers in our sample during the five years before our analysis. We could recover previous students' test scores for 779 out of 1,384 teachers, which explains the reduced sample size in columns 5–8 of Panel A.[17] We find no meaningful correlation between teachers' IAT score and the share of immigrant students they taught in the past (column 5), nor with the difference in the average test scores of past native and immigrant students (column 6). This suggests that stronger stereotypes toward immigrant students may not reflect statistical discrimination based on objective information on average group ability. The results remain qualitatively similar (with the exception of the *Northern* dummy in Panel A) when we introduce all regressors at the same time and when we include school fixed effects (columns 7 and 8).

In addition, Appendix Table A.6 shows no significant correlation between IAT score and characteristics such as having children, parents' education, and the beliefs on the reasons underlying the gap in high school track choice between native and immigrant students (e.g., ability, economic conditions, language differences, prejudice).[18]

3.4.2 Grades and student characteristics

Figure 3 shows the distribution of teacher-assigned grades (left graph) and standardized test scores (right graph) for native and immigrant students at the end of the school year, compiled using data for all schools in our field experiment sample over the school years 2011–12 to 2015–16 (i.e., before our experiment). The two measures have different scales, with teacher-assigned grades ranging from 3 to 10 and INVALSI scores from 0 to 100.

[Insert Figure 3]

[17] We include teachers who had at least three immigrant (and native) students.

[18] The detailed questions are reported in Appendix B.2.

The leftmost graph shows that at the end of school year, there is a substantial bunching in teacher-assigned grades at 6 ("pass"), for about 60% of immigrant students and 35% of native students. The distribution of both teacher grades and standardized test scores for native students first-order stochastically dominates that for immigrants. This gap may reflect differences in academic performance between native and immigrant students as well as other factors (e.g., gaps in diligence, behavioral issues, teacher bias in grading).[19] Appendix Table A.7 confirms that past grades and all other student characteristics are balanced between the treated and control group.

4 Implicit stereotypes and grading

In this section, we compare teacher grades between immigrant and native students, holding constant standardized test scores, and we relate differences in grading to teachers' IAT scores. Figure 4 plots the average grades assigned by teachers to immigrant and native students (on the vertical axis) by quintile of the standardized test score (on the horizontal axis), with the associated 95% confidence intervals. Not surprisingly, students with a higher standardized test score receive, on average, a higher grade from their teacher, with a correlation of 0.56. However, conditional on obtaining the same standardized test score, immigrant students receive significantly lower grades from teachers, particularly in the upper part of the test score distribution.[20] The average gap is 0.14, comparable in magnitude to the difference explained by maternal education: controlling for the quintiles of the standardized test score, students whose mothers have less than a high school diploma receive a grade that is on average 0.21 points lower compared to children of mothers with a high school diploma or university degree.

[Insert Figure 4]

The difference highlighted in Figure 4 may reflect teacher bias against immigrant students (see, e.g., Botelho et al., 2015; Burgess and Greaves, 2013; Hanna and Linden, 2012; Lavy, 2008). However, there may also be other reasons why immigrant students perform better in standardized tests than in teacher-graded assignments. For instance, teachers could place greater emphasis on multidimensional competence (e.g., oral expression, behavior in class) that is not easily captured by standardized tests. To corroborate the role of teachers' implicit stereotypes, we relate differences in grading to teachers' IAT scores.

[19] Appendix Figure A.3 reports the distribution of teacher-assigned grades separately for math and literature. The pattern is very similar.

[20] Appendix Figure A.4 provides separate figures for math and literature. The gap is found in both subjects.

Figure 5 shows the association between teachers' implicit stereotypes, as measured by their IAT score, and the grading of native and immigrant students. The black and blue solid lines represent the residuals from regressions of grades assigned to native and immigrant students, respectively, on teacher fixed effects, a cubic polynomial in the INVALSI test score, and cohort fixed effects (dashed lines represent the associated 95% confidence intervals). Higher values of the IAT score are associated with significantly lower grades to immigrant students, while they do not correlate with the grades assigned to native students (the black line remains flat around zero over the entire distribution of the IAT score). Table 3 quantifies the effects shown in Figures 4 and 5. Even controlling for teacher fixed effects and the cubic polynomial of the INVALSI test score, immigrant students receive on average a 0.097 lower teacher-assigned grade than native students (Panel A, column 1), which corresponds to 0.09 standard deviations. In column 2, the gap between natives and immigrants is about one-half when teachers do not have implicit bias against immigrants ($IAT = 0$) compared to highly biased teachers ($IAT = 1$). However, the difference is not statistically significant, likely due to two factors: measurement error and bunching at the low end of the grade distribution. We next discuss these two issues in order.

First, for each teacher t, we calculate a standard measure of bias in grading (θ_t) obtained as the gap between native (n) and immigrant (i) students (n) in the difference between "non-blind" teacher-assigned grades (NB) and "blind" standardized test scores (B).

$$\theta_t = (NB_{nt} - B_{nt}) - (NB_{it} - B_{it}) \tag{1}$$

In Appendix Figure A.2, we correlate teachers' IAT with the above measure of bias in grading, calculated in two alternative ways. In the leftmost panel, we use a "naive" measure that does not adjust for sample variation. This measure is positively but not significantly correlated with teachers' implicit bias (consistent with columns 2 and 3 of Table 3). In the rightmost panel of Figure A.2, to reduce estimation error arising from sample variation, we calculate an empirical Bayes estimate of the bias in grading following Kane and Staiger (2002), Chetty et al. (2014), and Terrier (2020).[21] This estimate shows a stronger positive and significant correlation with teachers' IAT score, suggesting that the measurement error, stemming from the fact that we only have data on a limited number of students for each teacher, may have large impacts on the results.

[Insert Figure 5]

Second, the empirical relationship between bias in grading and the IAT score is mitigated by the bunching in end-of-semester grading at the pass grade (score 6), with more than 60% of immigrant

[21] Details on how we calculate the empirical Bayes estimate of the bias in grading are reported in Appendix C.

students getting the pass grade in teacher-assigned evaluations (see Figure 3). This bunching makes it difficult to detect potential bias at the low end of the grade distribution. To gain more insights, we plot teacher-assigned grades by quintiles of the INVALSI score in Figure 6, separating teachers into high- and low-IAT groups (using 0.6 as the threshold for high bias, as in the literature). The figure shows that while teachers with low and high IAT scores give similar grades to native students throughout the test score distribution (right panel), teachers with stronger stereotypes give lower scores to high-performing immigrant students (left panel).

[Insert Figure 6]

Columns 4 and 7 of Table 3 show that the average gap in grading is three times as large for high-ability than for low-ability students. Furthermore, high-ability immigrant students get relatively lower grades than comparable native students when they are assigned to teachers with higher implicit stereotypes (columns 5 and 6), while the gap is small and insignificant for low-ability immigrant students (columns 8 and 9). Appendix Table A.8 shows that the results are qualitatively and quantitatively very similar when using the first difference between the teacher-assigned grades and test scores as an outcome.[22]

5 Main results

5.1 Field experiment

In the first experiment, we evaluate the effect that revealing to teachers their own stereotypes has on their grading behavior at the end of the first semester. Appendix Figure A.5 compares the distribution of grades assigned to immigrant and native students (left and right panel, respectively) by teachers in the treated (colored bar) and control groups (white bar). As explained in Section 3.2, teachers randomized into the treated group were offered feedback on their IAT score before end-of-semester grading, while teachers randomized into the control group could receive the same information only after grading. The leftmost graph in Appendix Figure A.5 shows that the distribution of grades assigned to immigrant students by teachers in the treated group shifts to the right compared to the distribution of those in the control group. The rightmost graph shows an opposite effect on grades assigned to native students.

[22] In the field experiment, we collected two IAT score with male and female names of natives and immigrants. Appendix Table D.1 reports the results using the gender-specific IAT for each students. The results are unaffected as gender is not a focal characteristic of those IAT tests: the key categories are Immigrant and Native, and the brain is focused on those when completing the IAT (Greenwald and Banaji, 1995).

In Table 4 we quantify the above effects by regressing the grade assigned by a teacher to a student on a treatment indicator for the teacher, a dummy for whether the student is immigrant, and the interaction between the two. Standard errors are clustered at the school level (the unit of randomization). Panel A shows the intention-to-treat effect: teachers offered the early IAT feedback exhibit a 0.35 point lower gap in grades between native and immigrant students (or 0.27 standard deviations) compared to teachers in the control group (column 1). The effect is generated by 0.2 point higher grades to immigrant students and 0.15 point lower grades to native students.

We interpret this result as driven by the implicit standardization of grades within each class and by the nature of the information provided to teachers. In fact, the IAT feedback compares the association with positive/negative attributes of native versus immigrant students. By virtue of randomization, the results are robust to controlling for student and teacher characteristics and the interaction of these characteristics with the *Immigrant* dummy (columns 2 and 3). As shown in Appendix Figure A.6, the results are also robust to a permutation test that replicates specification (1) in Table 4 after randomly assigning the treatment variable *IAT Feedback* across teachers 1,000 times. In only 6 out of 1,000 cases we find a coefficient larger than the one observed in Table 4.

[Insert Table 4]

A visual inspection of Appendix Figure A.5 suggests that the effect may be particularly large around the margin that separates passing and failing students (i.e., between scores 6 and 5). This is confirmed in columns 4–6 of Table 4, where the dependent variable is the probability of failing. Early IAT feedback decreases the probability of failing immigrant students by about 6 percentage points, whereas failing rates of native students remain unaffected (the coefficient on the standalone *IAT Feedback* dummy is not significantly different from zero).

In Panel B of Table 4, we rescale the intention-to-treat effect by the take-up rate of early IAT feedback, which was above 80%, to compute the treatment effect of stereotypes revelation. The variable *Email* in Panel B of Table 4 equals 1 if the teacher *actually received* the feedback and 0 if they did not receive any feedback. The coefficient on the interaction between the treatment and immigrant status increases in magnitude to about $+0.45$ for teacher-assigned grades (columns 1–3) and -0.07 for the probability of failing a school year.

Note that the magnitude of the treatment effect in Table 4 is not comparable to the magnitude of the bias in grading (i.e., the difference between teacher grades and standardized test scores) shown in Table 3. The reason is that the experiment was done at the end of the first semester (when standardized tests are not administered), while the bias in grading was measured at the end of the second semester (when we had information on both standardized test scores and teacher-assigned grades, hence we can control for the standardized test score). Also, the grading policy of teachers

typically differs between the first and the second semester, especially around the pass grade. Failing in the first semester represents a "warning" with no immediate consequences, while students may be retained in the same grade if they fail more than one subject in the second semester. For this reason, teachers are usually more reluctant to fail students in the second than in the first semester: indeed, the average fraction of students failing either literature or math (or both) is 21 percent in the first semester and only 2 percent in the second semester.[23]

Moreover, a lower propensity to fail students in the second semester sets a floor to teacher-assigned grades. For this reason, teachers' stereotypes likely induce a larger grade penalty for immigrant students in the first than in the second semester. To better compare the magnitude of the effects, we calculated a transition matrix between end-of-first-semester grades and end-of-second-semester grades for natives and immigrants.[24] We then estimated the impact of our field experiment using the transformed grade as the outcome. Appendix Table A.9 shows that, compared to our main result in Panel A of Table 4, the magnitude of the intention-to-treat effect on the immigrant-native gap is reduced by 36%: the effect of revealing stereotypes to teachers is around 0.23 grade points, or 0.18 standard deviations.

Heterogeneous effects

We next investigate the heterogeneity in teacher response by the strength of the signal received, as measured by a variable equal to average of the two IAT scores obtained by teachers. As explained in Section 3.1, each teacher in our field experiment received two pieces of feedback: one for the IAT using male names of natives and immigrants and one for the IAT using female names. The information provided in our field experiment is therefore less precise compared to the online experiment, in which teachers received only one signal on their implicit stereotypes.

[Include Figure 7]

Figure 7 plots the local polynomial smooth of the native-immigrant grades on the IAT score of teachers.[25] Panel A refers to the field experiment, while Panel B refers to the online experiment.

[23] Among immigrant students, failure rates in the first and second semester reach 31% and 4%, respectively.

[24] Using the control schools, we calculate the transition matrix between end-of-first-semester grades and end-of-second-semester grades, separately for immigrants and natives. Then, for each student, we calculate the "transformed" grade as the average score of their in-group at the end-of-second-semester, conditional on end-of-first-semester grade.

[25] More precisely, the figure plots the difference in the mean residuals of natives and the mean residuals of immigrants for the teachers. First, we calculate the residuals of the grade for each student absorbing our simple set of baseline controls (gender, education and occupation of parents). Then, for each teacher we calculate the difference in the weighted mean of residuals of natives and the weighted mean of residuals of immigrants.

The rightmost graph in Panel A shows that in the control group, which received no additional information before grading, native students receive higher grades compared to immigrants. The leftmost graph shows that the difference in the residuals is close to zero for the teachers that received the IAT feedback. Although imprecisely estimated, the slope in the relationship between the residualized grade gap and the IAT score is positive for control teachers and negative for treated ones. This qualitatively suggests that the most biased teachers are more generous toward natives compared to immigrants, but they change their behavior more when receiving the information on their own IAT score.

[Include Table 5]

In Table 5 we analyze the same heterogeneity in regression format. Columns 1-2 refer to the field experiment and columns 3-4 to the online one. Column 1 reports the baseline estimated effects, using the most stringent specification in column 3 of Table 4. Column 2 quantifies the evidence in Panel A of Figure 7. The coefficient on the triple interaction (*IAT Feedback*Immigrant*IAT Score*) in column 2 of Table 5 is positive, suggesting that the treatment induced more generous grading toward immigrants for teachers with stronger negative stereotypes, but the result is imprecisely estimated.

A higher average IAT encompasses two features. First, other things equal, teachers with a higher IAT receive a stronger signal about their implicit biases, which should in principle induce a greater reaction. Second, these teachers may be less willing to adjust their behavior, precisely because they are more biased to start with. The results we discussed capture both these effects. The online experiment will allow us to delve deeper into the differential effect of signal strength using a measure of the unexpected component of bias revelation. In addition, participants in the online experiment conducted only one IAT and received thus only one feedback, which may lead to a more precise reaction to their own IAT score. Teachers in the field experiment, by contrast, conducted two IAT tests (one with male and one with female names, and it is ex-ante unclear whether they would respond to the average of the two scores or to only one of them.[26]

In Appendix Table A.10, we explore the heterogeneity of the (intention-to-treat) effect along other teacher characteristics. The first is teachers' explicit bias against immigrants. Learning that one holds negative (implicit) stereotypes against immigrants conveys more information to teachers who are unaware of such stereotypes, hence we should expect a stronger reaction from teachers who

[26] We also tried using the gender-specific IAT score in the regression, depending on the gender of the student, but we found that the results presented in column 2 of Table 5 were unaffected (see column 2 Appendix Table D.2). This is not necessarily surprising, because by construction the IAT score should depend only on the categories (Immigrant-Native) and not on other non-focal characteristics (e.g., the gender in our case).

reported no explicit bias in the baseline survey. To test this hypothesis, in column 2 of Table A.10 we include a triple interaction between the indicator for immigrant student, early IAT feedback and the dummy variable '*WVS*', which equals 1 for teachers who agree with the statement that "immigrants and natives should have equal opportunities to access available jobs." The positive and significant coefficient on the triple interaction confirms that this group is more responsive to the intervention, consistent with the fact that they may have been less aware of their (implicit) stereotypes before our treatment.

We next explore the role played by awareness of anti-immigrant bias in society. Carlana et al. (2022b) show that teachers with stronger implicit bias are more likely to recommend vocational tracks and less likely to recommend top-tier tracks to immigrant students. In our survey we asked teachers whether they believed that bias against immigrant students may be why they enroll disproportionately into less demanding high school tracks compared to natives with the same performance. Twenty percent of teachers answered that it was "likely" or "extremely likely" that prejudice affected the choice of immigrant students. Column 3 in Appendix Table A.10 shows that these same teachers react more strongly to receiving information on their own implicit bias.

5.2 Online experiment

As explained in Section 3.3, we conducted a second experiment in which a different group of teachers was asked to grade 10 tests, with randomly assigned native- or immigrant-sounding student name to each test. As in the first experiment, teachers took an IAT at baseline, and we provided feedback on the IAT result only to a random group. Different from the first experiment, however, both the treated and the (active) control group received a generic debiasing message.

We start by visually showing the results of this experiment in Panel B Figure 7. The two graphs plot the average difference in grades assigned to native and immigrant students against teacher IAT scores, controlling for the quality of the answer, exam order, and subject.[27] The leftmost graph shows that receiving feedback on one's own IAT, in addition to the generic debiasing message, reduces the native-immigrant gap in grades for teachers who display relatively high levels of implicit bias, and the reduction is larger the higher their bias (IAT score). This is consistent with the interpretation that teachers who receive a more negative signal react by helping immigrants more. In contrast, there is a weakly positive —but insignificant— relationship between the gap in grades and IAT scores across teachers who receive only the debiasing message (rightmost graph).

[27]Recall from Section 3.3 that the answers to the tests that the teachers graded were prepared by consultants who also provided a score for each potential answer. This is the variable we include among the regressors to control for the "quality "of the answer.

Column 3 of Table 5 reports the main results of the online experiment in regression format. On average, receiving personalized feedback leads to a small decrease in the grades assigned across the board (significant at the 10% level), but it does not lead to a relative increase in grades assigned to immigrant students *compared to the debiasing message*. Recall that in the field experiment we observed a decrease in grading bias against immigrants as a consequence of the IAT feedback *compared to no information* in the control group. The two results thus provide interesting, complementary evidence: the two policies —generic debiasing message and personalized IAT feedback— have similar effects on average, but Figure 7 clearly shows heterogeneous effects depending on the feedback the teachers received.

Column 4 quantifies the results from Panel B of Figure 7. Teachers with no stereotypes ($IAT\ Score = 0$), who receive a generic debiasing message but remain unaware of their own IAT ($Feedback = 0$), assign a grade 0.42 points higher to immigrant students than to native ones. The fact that raising teachers' general awareness may reduce biased behavior is consistent with previous evidence on debiasing interventions (Boring and Philippe, 2021). The positive correlation on the grade of immigrant students disappears for teachers with an IAT equal to one (the coefficient on $Immigrant \times IAT\ Score$ is -0.426, with a standard error of 0.157). Teachers with IAT greater than one assign lower grades to immigrant students compared to native ones in the 'active control' group.

What happens to teachers' grading when they receive information on their own implicit stereotypes on top of the general debiasing message? The effect varies along the distribution of the IAT score. Teachers essentially assign the same grades, on average, to immigrant students and native students when the feedback reveals the absence of stereotypes (i.e., when $IAT\ Score = 0$, the gap in grading for immigrant students relative to native students is $0.420 - 0.580 = -0.16$ points, statistically indistinguishable from zero), but they significantly increase grades to immigrant students when the feedback reveals strong implicit stereotypes. The higher the IAT score —and therefore the signal received about one's own implicit bias— the stronger the teachers' response.

When receiving feedback on their IAT score, teachers are informed on whether they have "no/ slight/ moderate/ severe" stereotypes against immigrants. Column 1 of Table 6 replicates column 3 of Table 5 using a dummy variable taking value 1 if the teacher is "moderately or severely" biased, instead of the continuous IAT score.[28] The results are confirmed: teachers have a significantly stronger positive reaction in favour of immigrants when receiving the information that they have moderate/severe implicit stereotypes.

[28] In the online experiment, 80.8% of teachers have an IAT score above 0.35 and therefore are "moderately or severely" biased.

[Include Table 6]

Finally, we create an indicator for teachers underestimating their own bias, which equals 1 if the feedback received by the teacher is more negative than their prior (elicited at baseline using the same ordinal scale).[29] Column 2 of Table 6 shows that the effect of revealing stereotypes is driven by teachers who *underestimate* their own IAT and were thus 'surprised' by the information received. The coefficient of the triple interaction ($IAT\,Feedback \times Immigrant \times Underestimate\ Own\ IAT$) is unaffected when we include teachers' IAT score among the regressors (column 3) to account for the mechanical correlation due to more biased teachers being more likely to underestimate their own IAT.

In the last column of Table 6, we include all interactions from columns 1 and 2 to jointly investigate the role of the severity of the IAT feedback received and the belief on own IAT. The two variables are highly correlated (correlation coefficient of 0.71): 94% of the teachers with moderate or severe IAT also underestimate the feedback they receive. The coefficients on the two triple interactions in column 4 remain positive, but are less precisely estimated.[30] To sum up, teachers with higher IAT and teachers who update their beliefs more as a result of the signal change their grading behavior to a larger extent, but the data do not allow us to perfectly disentangle the two driving factors, as the two variables are highly correlated.

6 Conclusions

Immigrant students receive lower teacher-assigned grades than native students after controlling for their performance on blindly graded standardized tests. The gap is substantially wider for high-achieving immigrant students. We acknowledge that there may be characteristics that differentiate immigrant students from native students that are observable to teachers but unobservable to the econometrician (e.g., disciplinary problems or differences in performance on standardized multiple choice questions versus open ended ones). We show that for high-ability students, the difference in the grading of native and immigrant students is systematically correlated with teachers' stereotypes against immigrants, a pattern strongly suggestive of bias.

We conduct two novel experiments to test whether informing teachers about their own stereotypes may be an effective policy to reduce discrimination in grading. Our main treatment consists of receiving feedback on one's own IAT score. In the first experiment ('field experiment'), we share

[29] In the online experiment, 81.3% of teachers underestimate their own bias.

[30] As shown in Appendix Table A.11, the results are very similar when considering the continuous measures — instead of dummies— for IAT score and the difference between the own score and the expected score.

the IAT feedback with teachers in the treated group just before end-of-term grading, i.e., in time to adjust the grade given to students. The control group receives feedback right after end-of-term grading, i.e., too late to adjust grades. We find that teachers (randomly) assigned to the treatment react to the information by increasing the grades they give to immigrant students and decreasing the grades they give to native students. The effect is particularly strong around the threshold that determines whether a student passes or fails a subject.

In the second experiment ('online experiment'), teachers in the (active) control group receive a generic debiasing message, while teachers in the treatment group receive the debiasing message plus information on their own IAT score. Three weeks later, both groups are asked to grade tests randomly assigned to native- or immigrant-sounding names. We find that, on average, informing teachers of their own stereotypes does not increase grades given to immigrant students relative to natives, compared to the active control group. However, we find important heterogeneity based on teachers' baseline IAT. When teachers receive only generic debiasing, the higher their implicit stereotypes, the lower the grade assigned to immigrant students compared to native ones. When teachers receive information on their own stereotypes, they significantly increase the grades given to immigrant students compared to those given to natives only when their feedback suggests they hold negative views against immigrants. Furthermore, thanks to the elicitation of teachers' priors about their own IAT, we can show that the effect is driven by teachers who did not expect to receive negative feedback. This suggests that the effect of revealing stereotypes may come mainly from people being unaware of their own implicit bias.

Our findings can help inform an active policy debate regarding recent efforts by corporations, universities, schools, and other institutions to increase awareness about implicit bias by encouraging search committee members or new employees to take an IAT. In the context of schooling, the IAT is simple to implement and it would not cost much to ask every teacher to take it, say, at the beginning of the academic year.[31] This may help counteract negative stereotypes about certain groups. However, the implications of such a policy are not straightforward. By making teachers aware of their 'implicit' biases, their evaluation of students becomes fairer if they were acting upon their stereotypes by giving lower grades to immigrants. But it is possible that teachers whose negative stereotypes do not translate into discriminatory behavior may also react, thus inducing positive discrimination toward immigrant children. Further research on this point is warranted.

[31]The Ministry of Education in Peru has recently implemented a government educational program including the gender-science IAT for all teachers (Martínez, 2023), collected through the platform http://www.oportunidadesparatodos.pe.

References

Alesina, A. and La Ferrara, E. (2014). A test of racial bias in capital sentencing. *The American Economic Review*, 104(11):3397–3433.

Alesina, A., Miano, A., and Stantcheva, S. (2022). Immigration and redistribution. *Review of Economic Studies*.

Allport, G. W. (1958). *The nature of prejudice: Abridged*. Doubleday.

Altonji, J. G. and Blank, R. M. (1999). Race and gender in the labor market. *Handbook of labor economics*, 3:3143–3259.

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., and Amadori, A. (2008). Predicting the Vote: Implicit Attitudes as Predictors of the Future Behavior of Decided and Undecided Voters. *Political Psychology*, 29(3):369–387.

Arkes, H. R. and Tetlock, P. E. (2004). Attributions of Implicit Prejudice, or "Would Jesse Jackson 'Fail' the Implicit Association Test?". *Psychological Inquiry*, 15(4):257–278.

Avitzour, E., Choen, A., Joel, D., and Lavy, V. (2020). On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes. NBER Working Papers 27818, National Bureau of Economic Research, Inc.

Axt, J. R., Bar-Anan, Y., and Vianello, M. (2020). The Relation Between Evaluation and Racial Categorization of Emotional Faces. *Social Psychological and Personality Science*, 11(2):196–206.

Banaji, M. R. (2004). The opposite of a great truth is also true: Homage of Koan #7. In *Perspectivism in social psychology: The yin and yang of scientific progress.*, APA science series. APA decade of behavior series., pages 127–140. American Psychological Association, Washington, DC, US.

Bar-Anan, Y. and Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior research methods*, 46(3):668–688.

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.

Becker, G. S. (1957). *Economics of Discrimination*. University of Chicago Press.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). Implicit discrimination. *American Economic Review*, pages 94–98.

Bertrand, M. and Duflo, E. (2017). Field experiments on discrimination. *Handbook of Economic Field Experiments*, pages 309–393.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *The American Economic Review*, 94(4):991–1013.

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., and Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, 94(3):567.

Bohnet, I. (2016). *What Works: Gender Equality by Design*. Harvard University Press.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Boring, A. and Philippe, A. (2021). Reducing discrimination in the field: Evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching. *Journal of Public Economics*, 193:104323.

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.

Bursztyn, L., Chaney, T., Hassan, T. A., and Rao, A. (2021). The Immigrant Next Door: Exposure, Prejudice, and Altruism. Working Paper 28448, National Bureau of Economic Research.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

Carlana, M. and La Ferrara, E. (2021). Apart but connected: Online tutoring and student outcomes during the covid-19 pandemic. *EdWorkingPaper No. 21-350.*

Carlana, M., La Ferrara, E., and Pinotti, P. (2022a). Goals and gaps: Educational careers of immigrant children. *Econometrica*, 90(1):1–29.

Carlana, M., La Ferrara, E., and Pinotti, P. (2022b). Implicit stereotypes in teachers' track recommendations. *AEA Papers and Proceedings*, 112.

Carnes, M., Devine, P. G., Baier Manwell, L., Byars-Winston, A., Fine, E., Ford, C. E., Forscher, P., Isaac, C., Kaatz, A., Magua, W., Palta, M., and Sheridan, J. (2015). The effect of an intervention to break the gender bias habit for faculty at one institution: A cluster randomized, controlled trial. *Academic Medicine*, 90(2):221–230.

Chang, E. H., Milkman, K. L., Gromet, D. M., Rebele, R. W., Massey, C., Duckworth, A. L., and Grant, A. M. (2019). The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences of the United States of America*, 116(16):7778–7783.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *The American Economic Review*, 104(9):2593–2632.

Corno, L., La Ferrara, E., and Burns, J. (2022). Interaction, stereotypes and performance. Evidence from South Africa. *American Economic Review*, 112(12):3848–75.

Coviello, D. and Persico, N. (2015). An economic analysis of black-white disparities in the new york police department's stop-and-frisk program. *The Journal of Legal Studies*, 44(2):315–360.

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., and Snowden, R. J. (2010). Faking of the implicit association test is statistically detectable and partly correctable. *Basic and applied social psychology*, 32(4):302–314.

Dasgupta, N. and Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5):800–814.

Devine, P. G., Forscher, P. S., Cox, W. T. L., Kaatz, A., Sheridan, J., and Carnes, M. (2017). A Gender Bias Habit-Breaking Intervention Led to Increased Hiring of Female Faculty in STEMM Departments. *Journal of experimental social psychology*, 73:211–215.

Dobbie, W., Goldin, J., and Yang, C. S. (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–40.

Donders, F. C. (1969). On the speed of mental processes. *Acta psychologica*, 30:412–431.

Facchini, G., Margalit, Y., and Nakata, H. (2022). Countering public opposition to immigration: The impact of information campaigns. *European Economic Review*, 141:103959.

Fiedler, K. and Bluemke, M. (2005). Faking the iat: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology*, 27(4):307–316.

Figlio, D. N. (2005). Names, expectations and the black-white test score gap. Working Paper 11195, National Bureau of Economic Research.

Fryer Jr, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261.

Gawronski, B., Morrison, M., Phills, C. E., and Galdi, S. (2017). Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis. *Personality and Social Psychology Bulletin*, 43(3):300–312.

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, 132(3):1219–1260.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., and Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of general internal medicine*, 22(9):1231–1238.

Greenwald, A. G. and Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Greenwald, A. G., Nosek, B. A., and Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of personality and social psychology*, 85(2):197.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., and Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology*, 97(1):17.

Grigorieff, A., Roth, C., and Ubfal, D. (2018). Does information change attitudes towards immigrants? representative evidence from survey experiments. *Representative Evidence from Survey Experiments (March 10, 2018)*.

Guryan, J. and Charles, K. K. (2013). Taste-based or statistical discrimination: The economics of discrimination returns to its roots. *The Economic Journal*, 123(572):F417–F432.

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

Hardin, C. D. and Banaji, M. R. (2013). The nature of implicit prejudice: Implications for personal and public policy. In *The behavioral foundations of public policy.*, pages 13–31. Princeton University Press, Princeton, NJ, US.

Heß, S. (2017). Randomization inference with Stata: A guide and software. *Stata Journal*, 17(3):630–651.

Hopkins, D. J., Sides, J., and Citrin, J. (2019). The muted consequences of correct information about immigration. *The Journal of Politics*, 81(1):315–320.

Howell, J. L., Gaither, S. E., and Ratliff, K. A. (2015). Caught in the middle: Defensive responses to IAT feedback among whites, blacks, and biracial black/whites. *Social Psychological and Personality Science*, 6(4):373–381.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jost, J. T. (2019). The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit Attitudes Are Indispensable to Social and Political Psychology. *Current Directions in Psychological Science*, 28(1):10–19.

Jussim, L. and Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and social psychology review*, 9(2):131–155.

Kane, T. J. and Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *The Journal of Economic Perspectives*, 16(4):91–114.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Karpinski, A. and Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *American Psychological Association*, 81(5):774–788.

Knowles, J., Persico, N., and Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229.

Lai, C. K., Hoffman, K. M., and Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass*, 7(5):315–330.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C. M., Sriram, N., Banaji, M. R., and Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4):1765–1785.

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 biases and effect on longer term outcomes: University admission exams and choice of field of study. *Working Paper*.

Lavy, V. and Sand, E. (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263–279.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review*, 111(6):1807–44.

Martínez, J. J. (2023). The long-term effects of teachers' gender stereotypes. *Working Paper*.

Meissner, F., Grigutsch, L. A., Koranyi, N., Müller, F., and Rothermund, K. (2019). Predicting behavior with implicit measures: Disillusioning findings, reasonable explanations, and sophisticated solutions. *Frontiers in Psychology*, 10.

Mengel, F. (2021). Gender Bias in Opinion Aggregation. *International Economic Review*, 62(3):1055–1080.

Mitchell, G. and Tetlock, P. E. (2017). Popularity as a poor proxy for utility: The case of implicit prejudice. In *Psychological science under scrutiny: Recent challenges and proposed solutions.*, pages 164–195. Wiley Blackwell.

Monteith, M. J., Voils, C. I., and Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, 19(4):395–417.

Nosek, B. A. and Hansen, J. J. (2008). Personalizing the implicit association test increases explicit evaluation of target concepts. *European Journal of Psychological Assessment*, 24(4):226–236.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B. T., Wiers, R. W., Somogyi, M., Akrami, N., Ekehammar, B., Vianello, M., Banaji, M. R., and Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10593–10597.

O'Brien, L. T., Crandall, C. S., Horstman-Reser, A., Warner, R., Alsbrooks, A., and Blodorn, A. (2010). But I'm no bigot: How prejudiced White Americans maintain unprejudiced self-images. *Journal of Applied Social Psychology*, 40(4):917–946.

OECD (2014). Are boys and girls equally prepared for life?

Olson, M. A. and Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: personalizing the IAT. *Journal of Personality and Social Psychology*, 86(5):653.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105(2):171.

Ottaway, S. A., Hayden, D. C., and Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition*, 19(2):97–144.

Paluck, E. L., Green, S. A., and Green, D. P. (2018). The contact hypothesis re-evaluated. *Behavioural Public Policy*, 3(2):1–30.

Papageorge, N. W., Gershenson, S., and Kang, K. (2020). Teacher expectations matter. *The Review of Economics and Statistics*, 102(2):234–251.

Pope, D. G., Price, J., and Wolfers, J. (2018). Awareness reduces racial bias. *Management Science*, 64(11).

Price, J. and Wolfers, J. (2010). Racial discrimination among nba referees. *The Quarterly journal of economics*, 125(4):1859–1887.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408.

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, 17(3):523–534.

Rosenthal, R. and Jacobson, L. (1968). Pygmalion in the Classroom. *The Urban Review*, 3(1):16–20.

Rothermund, K. and Wentura, D. (2004). Underlying Processes in the Implicit Association Test: Dissociating Salience From Associations. *American Psychological Association*, 133(2):139–165.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., and Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition*, 17(4):437–465.

Schimmack, U. (2021). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science*, 16(2):396–414.

Sukhera, J., Milne, A., Teunissen, P. W., Lingard, L., and Watling, C. (2018). The actual versus idealized self: Exploring responses to feedback about implicit bias in health professionals. *Academic Medicine*, 93(4):623–629.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, 77(December 2018):101981.

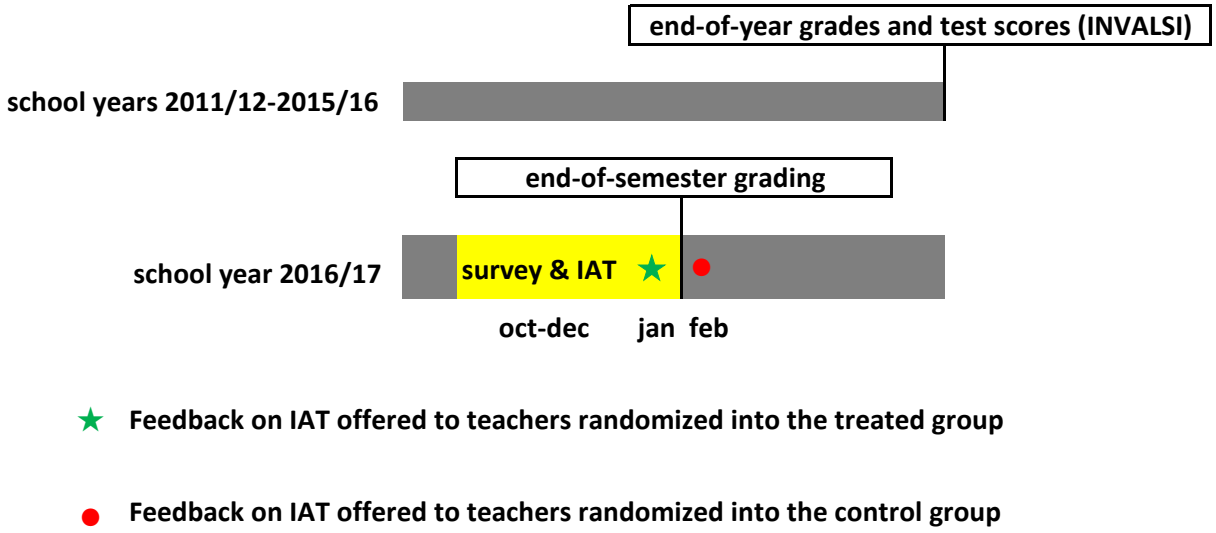
Tetlock, P. E. and Mitchell, G. (2009). Implicit Bias and Accountability Systems: What Must Organizations Do to Prevent Discrimination? *Research in Organizational Behavior*, 29:3–38.

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., and Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, 47(2):497–527.

Van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, 30(5):1045–1058.

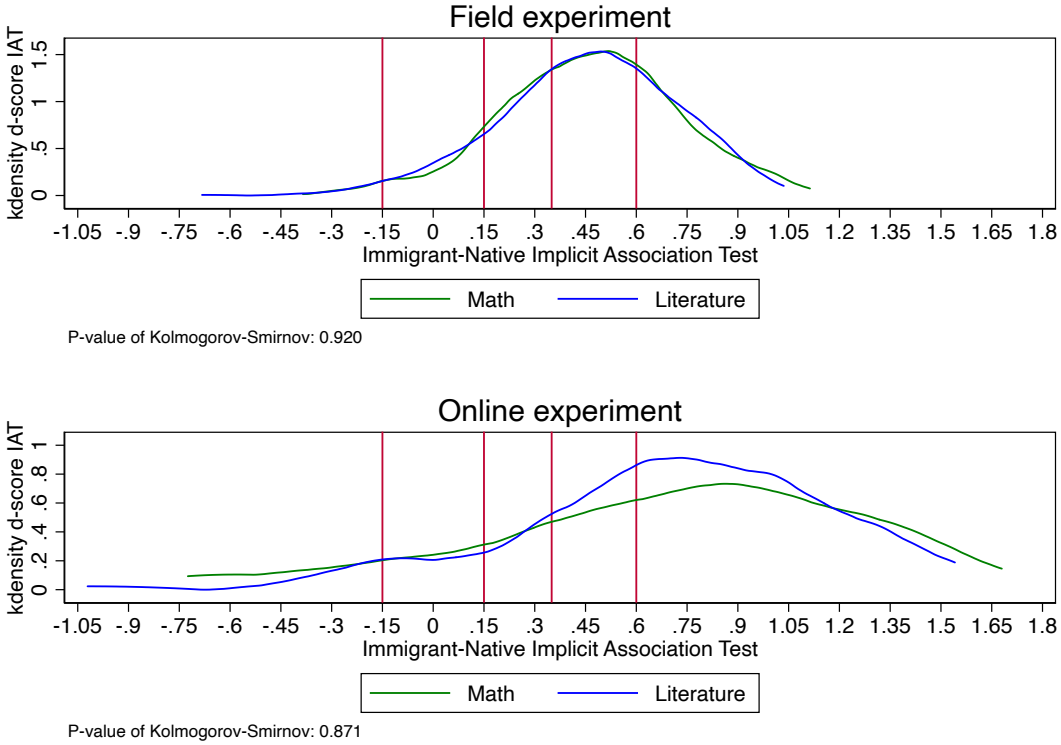
Tables and Figures

Figure 1: Timeline of data collection for the field experiment



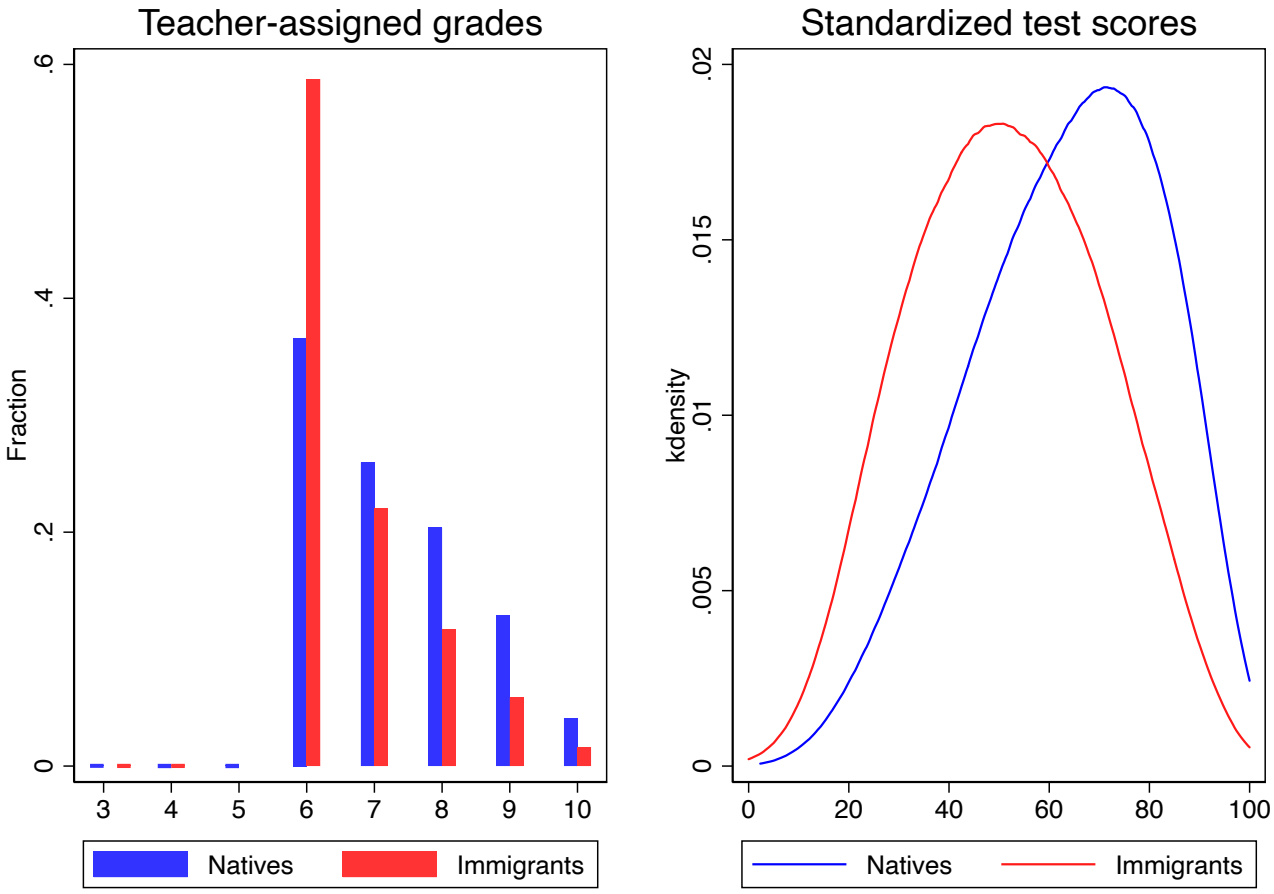
Notes: This figure shows the timeline of the data collection, survey, and field experiment. As described at length in Section 3, we obtained administrative data on end-of-year teacher-assigned grades as well as on standardized, blindly graded test scores for school years 2012/13 through 2015/16. During the first semester of the 2016/17 school year (October–January), we administered the survey and the IAT to all teachers in our sample. On January 2017, before end-of-semester grading, we sent feedback about teachers' own IAT scores to a random group of teachers. All other teachers were allowed to see their score after the end-of-semester grading (i.e., February 2018).

Figure 2: Distribution of the immigrant-native IAT score across teachers



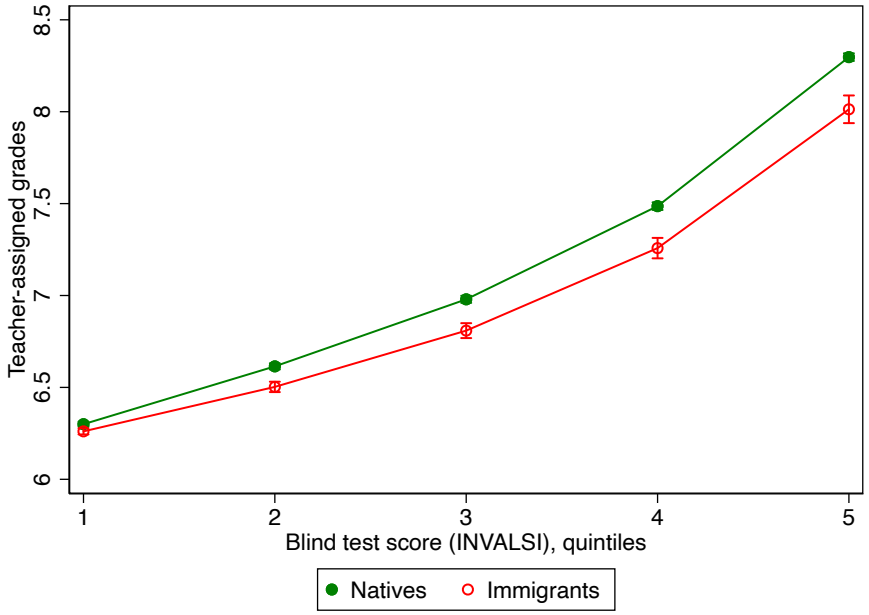
Notes: This graph shows the distribution of raw IAT scores for math and for literature teachers. A positive value indicates a stronger association between "natives" and "good" and "immigrants" and "bad." The first panel reports the IAT score for teachers participating in the first experiment (in person, 1,390 teachers), while the second panel shows the IAT score of teachers participating in the second experiment (online, 146 teachers). The vertical lines indicate the critical thresholds suggested by Greenwald et al. (2009) for defining different levels of bias. The negative association with immigrant names is absent when the IAT score is positive but below 0.15, "slight" when it lies between 0.15 and 0.35, and "moderate to severe" when it is above 0.35. Negative values of these same thresholds define the strength of positive associations.

Figure 3: Distribution of grades



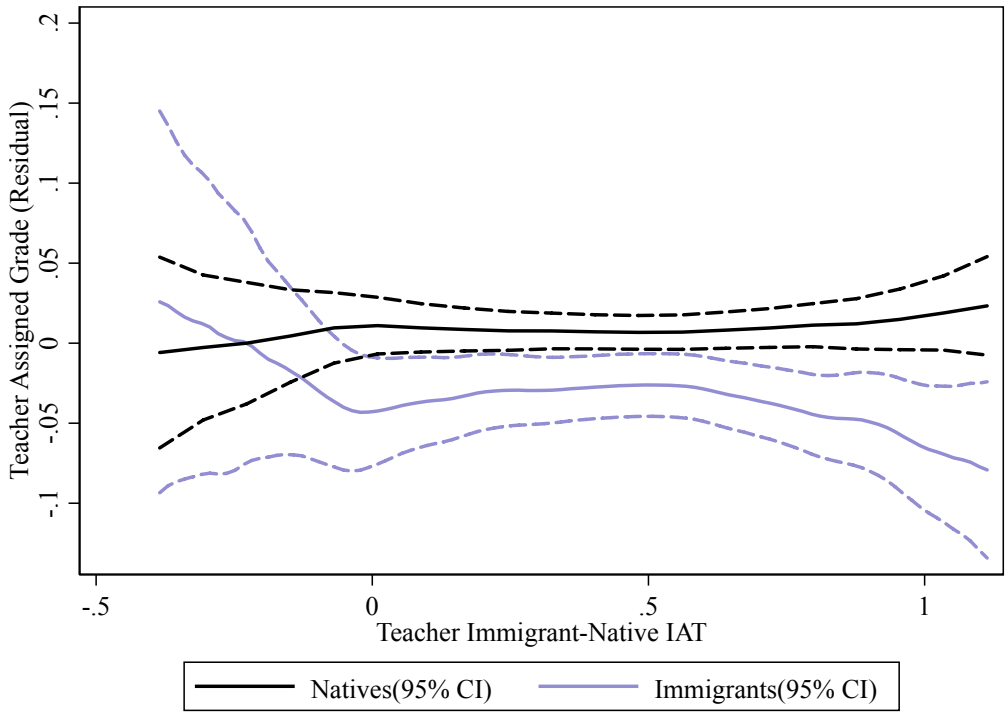
Notes: The graphs show the distribution of teacher-assigned grades (left panel) and standardized test scores INVALSI (right panel). The blue bar is for native students and the red bar is for immigrant students. For both teacher-assigned grades and standardized test scores, we report the average of math and reading scores. Students in this sample completed grade 8 between school years 2011–2012 and 2015–2016.

Figure 4: Teacher-assigned grades vs. blindly graded, standardized test scores



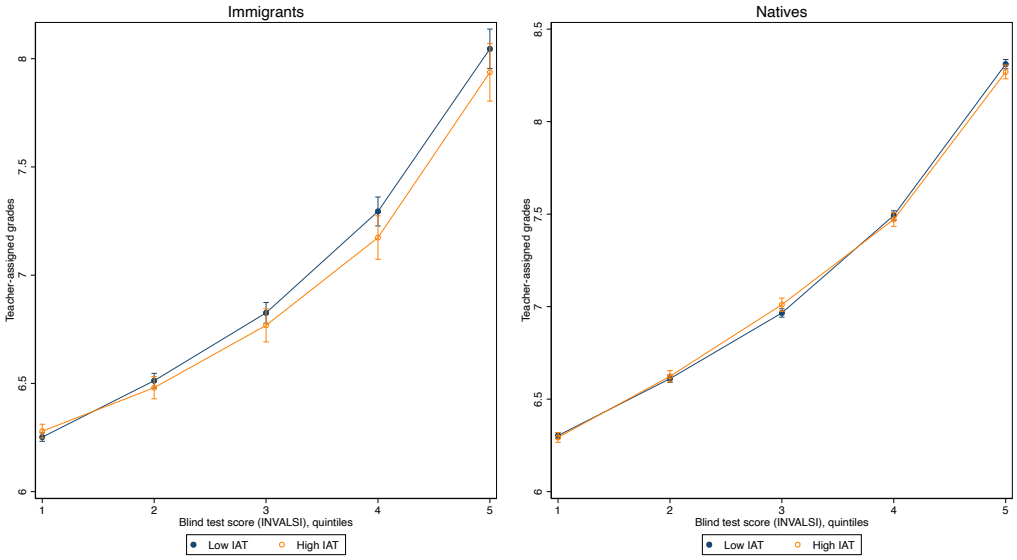
Notes: This graph shows teacher-assigned grades (non-blindly graded) on the vertical axis and quintiles of the standardized test score INVALSI (blindly graded) on the horizontal axis at the end of grade 8. Teacher-assigned grades are on a scale of 3 to 10, with 6 as the pass grade. The green squares and lines are for native students, while the red circles and lines are for immigrant students. Students in this sample completed grade 8 between school years 2011–2012 and 2015–2016.

Figure 5: The correlation between teacher-assigned grades and the IAT for immigrants and natives



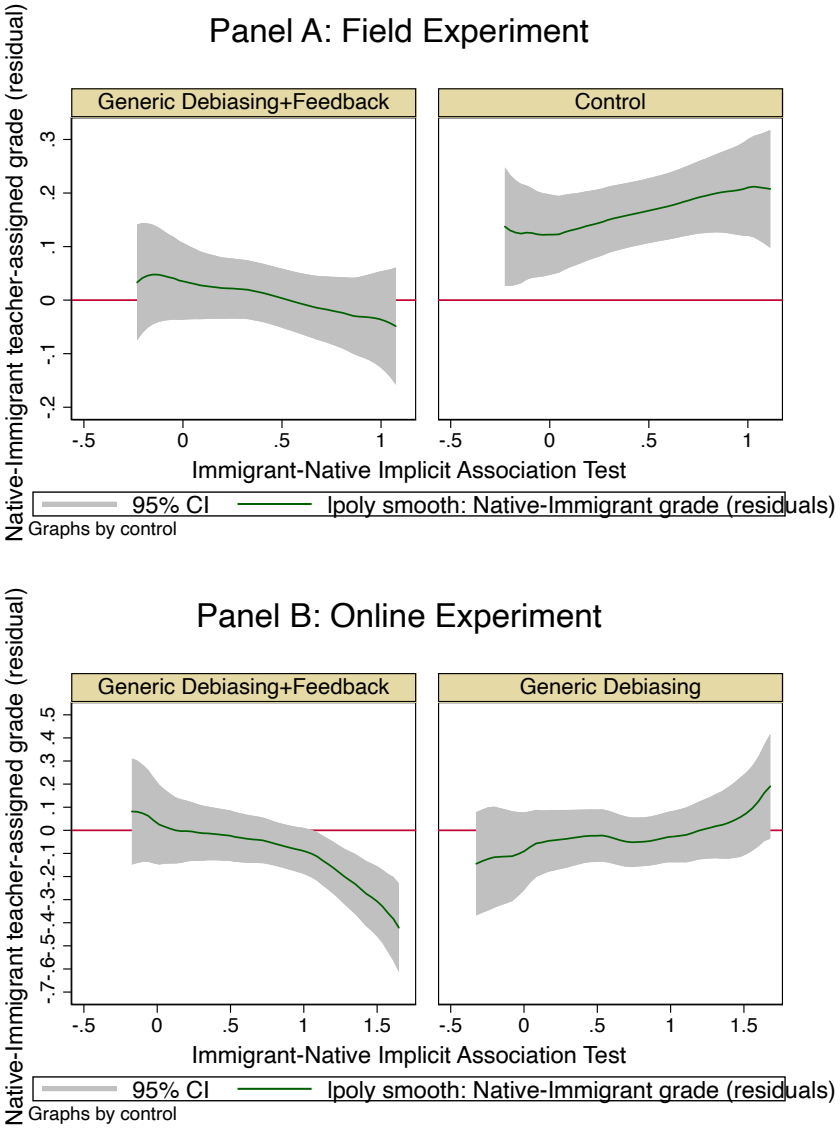
Notes: This graph shows the correlation between the residual of the teacher-assigned grade and the IAT for all teachers in the sample (math and literature). The residual is calculated absorbing the teacher fixed effects, a cubic polynomial of the INVALSI test score, and cohort fixed effects. The x-axis shows the Immigrant-Native IAT (raw d-score). The dotted lines represents the 95% confidence interval. Students in this sample completed grade 8 between school years 2011–2012 and 2015–2016.

Figure 6: Teacher-assigned grades vs. blindly graded, standardized test scores by teacher IAT (high vs. low)



Notes: This graph shows teacher-assigned grades (non-blindly graded) on the vertical axis and quintiles of the standardized test score INVALSI (blindly graded) on the horizontal axis at the end of grade 8. Teacher-assigned grades are on a scale of 3 to 10, with 6 as the pass grade. The blue squares and lines are for students of teachers with an IAT lower than 0.6 (high bias), while the yellow circles and lines are for students of teachers with an IAT lower than 0.6 (low bias). The left panel presents grades for immigrant students, while the right panel presents grades for native students. Students in this sample completed grade 8 between school years 2011–2012 and 2015–2016.

Figure 7: The impact of revealing stereotypes to teachers on grading



Notes: This graph shows the difference in grading of teachers in the field experiment (top panel) and online experiment (bottom panel) by their Immigrant-Native IAT score (raw d-score). First, we calculate for each grade given by teachers the residual considering the standard set of controls (for the field experiment: gender, education and occupation of parents, teacher controls such as gender, age, place of birth; for the online experiment: original grade on the question, subject, and order of questions). Then, for each teacher in our sample, we calculate the difference between the weighted mean of residuals of natives and the mean of residuals of immigrants.

Table 1: Balance table: Teacher characteristics

Panel A: Teachers in the Field Experiment					
	(1)	(2)	(3)	(4)	(5)
	Full Sample	Control	Treated	p-Value	Norm. Diff.
IAT	0.477	0.493	0.464	0.174	-0.079
	(0.261)	(0.269)	(0.253)		
Female	0.867	0.835	0.892	0.116	0.118
	(0.340)	(0.372)	(0.311)		
Teaching Math	0.494	0.506	0.485	0.397	-0.030
	(0.500)	(0.501)	(0.501)		
Advanced STEM	0.105	0.122	0.091	0.245	-0.071
	(0.307)	(0.328)	(0.288)		
Born in the North	0.665	0.679	0.653	0.577	-0.039
	(0.473)	(0.468)	(0.477)		
Age	47.455	48.114	46.929	0.406	-0.066
	(12.809)	(11.613)	(13.685)		
Full-Time Contract	0.826	0.802	0.845	0.313	0.080
	(0.380)	(0.400)	(0.362)		
Experience/10 Years	1.955	1.967	1.946	0.881	-0.012
	(1.191)	(1.191)	(1.192)		
Children	0.702	0.696	0.707	0.836	0.017
	(0.458)	(0.461)	(0.456)		
Low Edu Mother	0.448	0.468	0.431	0.434	-0.053
	(0.498)	(0.500)	(0.496)		
Middle Edu Mother	0.301	0.304	0.300	0.914	-0.006
	(0.459)	(0.461)	(0.459)		
High Edu Mother	0.150	0.148	0.152	0.926	0.008
	(0.357)	(0.356)	(0.359)		
Degree Laude	0.243	0.232	0.253	0.574	0.035
	(0.430)	(0.423)	(0.435)		
WVS Immigrants' Rights to Job	0.594	0.591	0.596	0.909	0.007
	(0.492)	(0.493)	(0.492)		
Observations	534	237	297		
Panel B: Teachers in the Online Experiment					
	(1)	(2)	(3)	(4)	(5)
	Full Sample	Control	Treated	p-Value	Norm. Diff.
IAT	0.704	0.729	0.677	0.493	-0.073
	(0.502)	(0.496)	(0.510)		
Female	0.863	0.851	0.875	0.680	0.049
	(0.345)	(0.358)	(0.333)		
Born in the North	0.603	0.568	0.639	0.451	0.102
	(0.491)	(0.499)	(0.484)		
Experience	20.308	20.770	19.833	0.606	-0.062
	(10.574)	(10.389)	(10.812)		
Teaching Math	0.349	0.338	0.361	0.746	0.034
	(0.478)	(0.476)	(0.484)		
Teaching Italian	0.479	0.486	0.472	0.856	-0.020
	(0.501)	(0.503)	(0.503)		
Underestimate Own IAT	0.801	0.811	0.792	0.783	-0.033
	(0.400)	(0.394)	(0.409)		
Observations	146	74	72		

Notes: The table shows the mean of the characteristics of the full sample of teachers for the field experiment (column 1), teachers in the control group (column 2), and teachers in the treatment group (column 3). Standard deviations are in parentheses in columns 1, 2, and 3, and the *p*-value of the difference is in column 4. The last column reports the normalized difference between group averages.

Table 2: Correlation between teacher characteristics and IAT (field experiment sample)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Dep Var: IAT Score (Stereotypes against Immigrants) in Field Experiment								
Female	-0.042**						-0.040	-0.046
	(0.020)						(0.027)	(0.033)
Born in the North		-0.021					-0.057***	-0.043**
		(0.014)					(0.017)	(0.020)
Experience/10 Years			-0.003				0.004	0.007
			(0.005)				(0.008)	(0.009)
WVS Immigrants' Rights to Job				-0.058***			-0.045*	-0.042
				(0.017)			(0.025)	(0.031)
Share of Immigrants					-0.065		-0.070	-0.212
					(0.067)		(0.067)	(0.133)
Native-Imm INVALSI(/100)						-0.040	-0.022	-0.041
						(0.083)	(0.088)	(0.119)
School FE	No	No	No	No	No	No	No	Yes
Obs.	1,384	1,384	1,384	1,384	779	779	779	779
R^2	0.063	0.061	0.061	0.066	0.093	0.092	0.117	0.203
Panel B. Dep Var: IAT Score (Stereotypes against Immigrants) in Online Experiment								
Female	-0.014						-0.007	0.110
	(0.088)						(0.090)	(0.204)
Born in the North		-0.044					-0.070	-0.221
		(0.082)					(0.081)	(0.176)
Experience/10 Years			0.057				0.050	0.013
			(0.036)				(0.039)	(0.072)
WVS Immigrants' Rights to Job				-0.139*			-0.127	-0.190
				(0.074)			(0.078)	(0.158)
School FE	No	No	No	No			No	Yes
Obs.	146	146	146	146			146	146
R^2	0.006	0.008	0.020	0.023			0.037	0.444

Notes: This table reports OLS estimates, where the dependent variable is the IAT score of teachers and the unit of observation is teacher *t* in school *s*. Panel A reports the correlations for teachers in the field experiment, while Panel B reports the correlations for teachers in the online experiment. We include controls for the order of IATs and for whether the blocks were presented in a order-compatible or order-incompatible way (which was randomized at the individual level). The variable "WVS Immigrants' Rights to Job" equals 1 for teachers believing that immigrants should have the same right to jobs as natives. The variable "Reason Gap: Prejudice" equals 1 if the teacher believes or strongly believes that the gap in high school track choices between natives and immigrants is due to prejudice. "Native-Imm INVALSI(/100)" indicates the difference in average standardized test scores of native and immigrant students assigned to the teacher in the previous four years. In columns 5–8 of Panel A, the number of observations decreases because information on past students is not available for all teachers; in these columns, we control for the number of observations with information available for at least three immigrant and native students.

Table 3: Bias in grading and teachers' IAT scores

Outcome: Teacher Grade									
	All			High Ability			Low Ability		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Immigrant	-0.097***	-0.062**	0.538	-0.179***	-0.115***	1.635**	-0.056***	-0.041	0.598
	(0.012)	(0.027)	(0.479)	(0.020)	(0.042)	(0.772)	(0.013)	(0.029)	(0.591)
IAT* Immigrant		-0.075	-0.065		-0.139*	-0.141*		-0.031	-0.031
		(0.050)	(0.049)		(0.080)	(0.079)		(0.060)	(0.058)
Obs.	42302	42302	42302	25415	25415	25415	16867	16867	16867
R^2	0.481	0.481	0.509	0.403	0.403	0.442	0.222	0.222	0.258
Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
INVALSI cubic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	No	No	Yes	No	No	Yes	No	No	Yes
Student Controls*Imm	No	No	Yes	No	No	Yes	No	No	Yes
Teacher Controls*Imm	No	No	Yes	No	No	Yes	No	No	Yes

Notes: This table reports OLS estimates, where the dependent variable is the teacher-assigned grade. The unit of observation is student i taught by teacher t in school s. "Immigrant" indicates whether the student is not Italian citizen. "IAT" indicates the Immigrant-Native IAT (d-score). Student controls include gender, generation of immigration, mother education, and province. Columns 1-3 provides the estimates for the full sample, 4-6 for high-ability students, and 7-9 for low-ability students, with a sample split based on the standardized test score INVALSI. Teacher controls include gender, place of birth, age, and age squared. Students in this sample completed grade 8 between school years 2011–2012 and 2015–2016. Standard errors are robust and clustered at the teacher level.

Table 4: Impact of revealing stereotypes to teachers in the field experiment

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var:	**Grade**			**Fail (Grade $<$ 6)**		
Panel A: Intention to Treat						
IAT Feedback*Immigrant	0.351***	0.369***	0.367***	-0.052*	-0.059**	-0.062**
	(0.111)	(0.095)	(0.096)	(0.028)	(0.025)	(0.024)
Immigrant	-0.704***	-0.683***	0.294	0.118***	0.088**	-0.223
	(0.064)	(0.154)	(0.940)	(0.018)	(0.039)	(0.266)
IAT Feedback	-0.148*	-0.166**	-0.153*	0.011	0.012	0.009
	(0.086)	(0.077)	(0.079)	(0.019)	(0.017)	(0.017)
Obs	10279	10279	10279	10279	10279	10279
R^2	0.028	0.126	0.131	0.012	0.043	0.047
Panel B: Local Average Treatment Effect						
Email*Immigrant	0.450***	0.471***	0.466***	-0.066*	-0.074**	-0.076**
	(0.138)	(0.124)	(0.124)	(0.034)	(0.030)	(0.029)
Immigrant	-0.704***	-0.632***	0.245	0.118***	0.080**	-0.214
	(0.063)	(0.163)	(0.934)	(0.018)	(0.040)	(0.264)
Email	-0.200*	-0.221**	-0.202*	0.015	0.016	0.012
	(0.114)	(0.106)	(0.107)	(0.026)	(0.022)	(0.022)
Obs	10279	10279	10279	10279	10279	10279
R^2	0.028	0.126	0.131	0.012	0.043	0.047
Mean Control Natives	7.03	7.03	7.03	0.10	0.10	0.10
Mean Control Immigrants	6.37	6.37	6.37	0.22	0.22	0.22
Students Controls	No	Yes	Yes	No	Yes	Yes
Students Controls*Imm	No	Yes	Yes	No	Yes	Yes
Teacher Controls	No	No	Yes	No	No	Yes
Teacher Controls*Imm	No	No	Yes	No	No	Yes

Notes: This table reports OLS estimates (Panel A) and IV estimates (Panel B), where the dependent variable is the grade (columns 1–3) or the probability of obtaining a grade lower than 6 (columns 4–6) at the end of the first semester of grade 8 (January). The unit of observation is student *i* in class *c* taught by teacher *t* in grade 8 of school *s*. Standard errors are robust and clustered at the school level. "IAT Feedback" is a dummy variable indicating whether the teacher was eligible for receiving the feedback before end-of-semester grading (January) or after end-of-semester grading (February). "Email" is a dummy variable indicating whether teachers eligible for receiving the feedback before end-of-semester grading actually requested it. The coefficients in Panel B are estimated by instrumental variables, using "IAT Feedback" as an instrument for "Email." Student controls include gender, generation of immigration, and education of the mother, all interacted with whether the student is an immigrant. Teacher controls include gender, place of birth, age, and age squared, interacted with whether the student is an immigrant.

Table 5: The impact of revealing stereotypes in the field and online experiment, by teacher IAT score

Dep. Var: Teacher-Assigned Grade				
	Field Experiment		**Online Experiment**	
	(1)	(2)	(3)	(4)
Immigrant	.294	0.371	0.109	0.420***
	(0.940)	(0.925)	(0.083)	(0.141)
IAT Feedback	-0.153*	-0.122	-0.216*	0.005
	(0.079)	(0.115)	(0.111)	(0.173)
IAT Feedback × Immigrant	0.367***	0.267	0.017	-0.580***
	(0.096)	(0.174)	(0.122)	(0.196)
IAT Feedback × IAT Score		-0.065		-0.329*
		(0.154)		(0.171)
IAT Feedback × Immigrant × IAT Score		0.214		0.849***
		(0.302)		(0.241)
IAT Score		0.019		-0.068
		(0.122)		(0.130)
Immigrant × IAT Score		-0.078		-0.426***
		(0.225)		(0.157)
Control Mean	6.944	6.944	7.134	7.134
Obs.	10279	10279	1460	1460
R^2	0.131	0.131	0.440	0.450
Subject, Order, Original Grade FE	No	No	Yes	Yes
Student Controls	Yes	Yes	Yes	Yes
Teacher Controls	Yes	Yes	Yes	Yes

Notes: This table reports OLS estimates, where the dependent variable is the grade assigned by teachers in the field experiment in columns 1-2 and online experiment in columns 3-4. The unit of observation is student i by teacher t. Standard errors are robust and clustered at the school level (the unit of randomization). "IAT Feedback" is a dummy variable indicating whether the teacher was eligible for receiving the IAT feedback versus the active control message. "IAT Score" is a continuous variable indicating the standard d-score of the Immigrant-Native IAT test (more details available on Appendix B.1). Student controls include gender, generation of immigration, and education of the mother, all interacted with whether the student is an immigrant for the field experiment. Teacher controls include gender, place of birth, age, and age squared, interacted with whether the student is an immigrant for the field experiment. Student controls include gender and class for the online experiment. Teacher controls include gender, place of birth, and a dummy for whether the teacher completed the IAT before the first reminder for the online experiment.

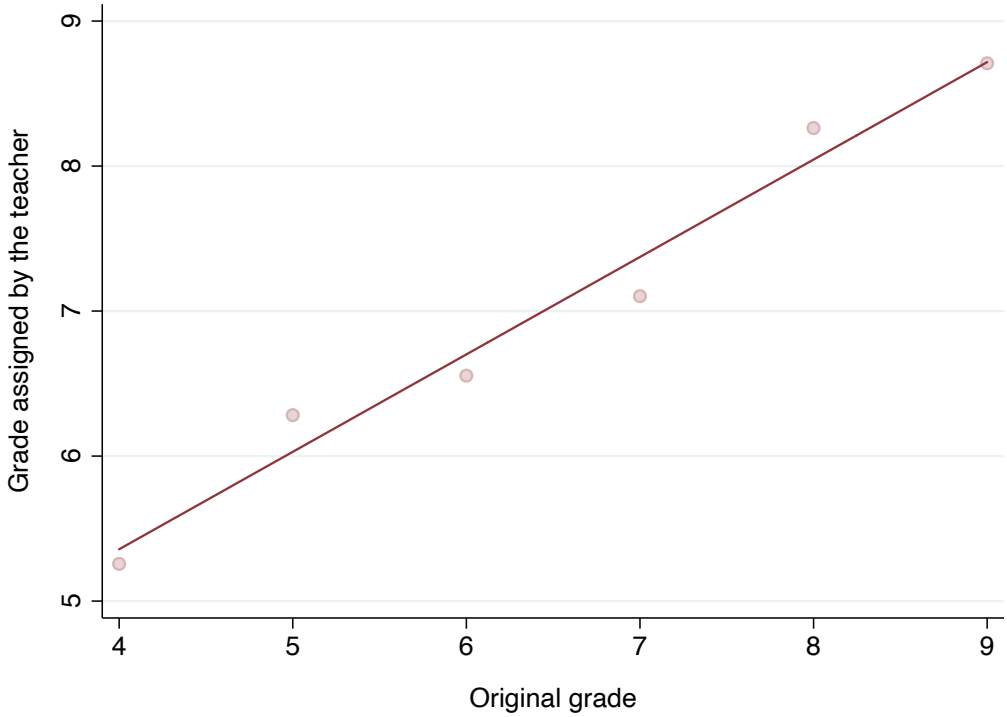
Table 6: Beliefs Updating in the Online Experiment

Dep. Var: Teacher-Assigned Grade				
	(1)	(2)	(3)	(4)
IAT Feedback	0.042	-0.056	-0.056	-0.009
	(0.205)	(0.281)	(0.274)	(0.264)
Immigrant	0.400**	0.401**	0.423**	0.463**
	(0.173)	(0.161)	(0.178)	(0.191)
IAT Feedback × Immigrant	-0.475**	-0.514**	-0.515**	-0.591**
	(0.230)	(0.237)	(0.242)	(0.249)
Moderate/Severe IAT	-0.064		-0.196	-0.087
	(0.161)		(0.172)	(0.211)
IAT Feedback × Moderate/Severe IAT	-0.317			-0.219
	(0.228)			(0.335)
Immigrant × Moderate/Severe IAT	-0.358*		-0.097	-0.270
	(0.184)		(0.222)	(0.258)
IAT Feedback × Immigrant × Moderate/Severe IAT	0.608**			0.345
	(0.262)			(0.429)
Underestimate own IAT		-0.069	0.074	-0.005
		(0.226)	(0.265)	(0.290)
IAT Feedback × Underestimate own IAT		-0.198	-0.199	-0.038
		(0.305)	(0.297)	(0.424)
Immigrant × Underestimate own IAT		-0.354**	-0.285	-0.160
		(0.172)	(0.225)	(0.231)
IAT Feedback × Immigrant × Underestimate own IAT		0.660**	0.661**	0.409
		(0.271)	(0.275)	(0.439)
Constant	5.871***	5.878***	5.948***	5.928***
	(0.323)	(0.329)	(0.330)	(0.330)
Control Mean	7.134	7.134	7.134	7.134
Obs.	1460	1460	1460	1460
R^2	0.446	0.448	0.450	0.450
Subject, Order, Original Grade FE	Yes	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes	Yes
Teacher Controls	Yes	Yes	Yes	Yes

Notes: This table reports OLS estimates, where the dependent variable is the grade assigned by teachers in the online experiment. The unit of observation is student i by teacher t. Standard errors are robust and clustered at the school level (the unit of randomization). "IAT Feedback" is a dummy variable indicating whether the teacher was eligible for receiving the IAT feedback versus the active control message. "Moderate/Severe IAT" is a dummy variable indicating whether the IAT is above 0.35 and the teachers received as feedback a moderate or severe association Immigrant-Bad Native-Good. "Underestimate own IAT" is a dummy that equals 1 if the teacher believes her IAT score is lower compared to the actual score. Student controls include gender and class. Teacher controls include gender, place of birth, and a dummy for whether the teacher completed the IAT before the first reminder.

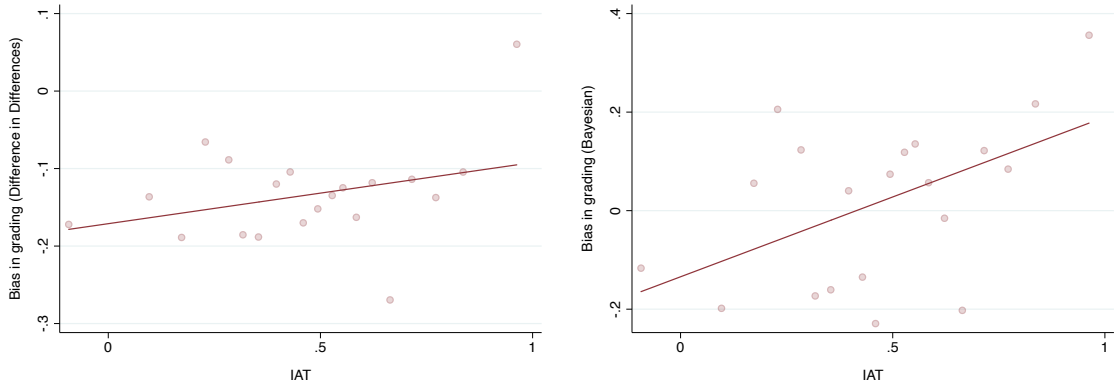
A Appendix tables and figures

Figure A.1: Teacher-assigned grades in the online experiment vs. original grades



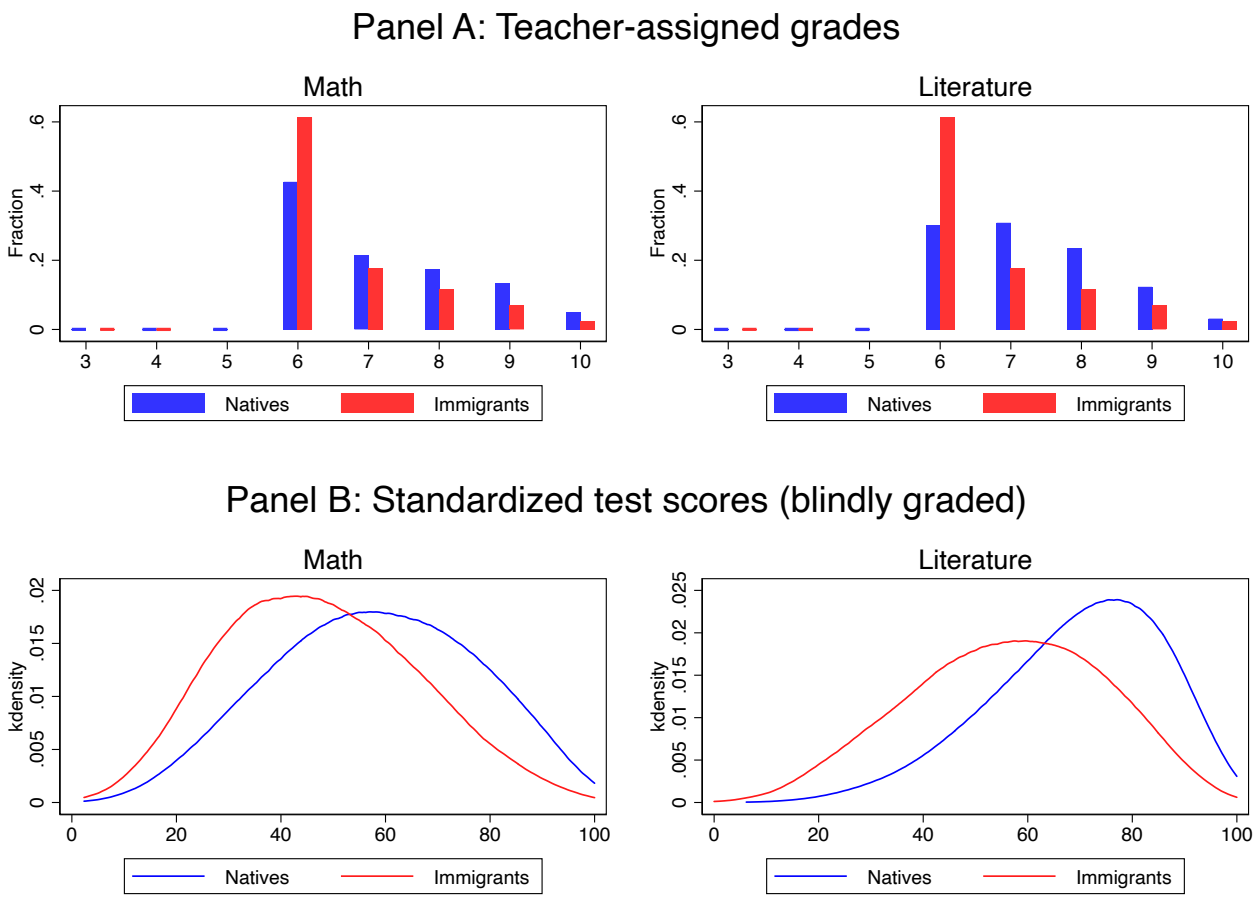
Notes: This graph shows the correlation between teacher-assigned grades in the online experiment and the original grades of the exams assigned by the teachers who prepared the answers in the online experiment.

Figure A.2: Correlation between Bias in Grading and IAT



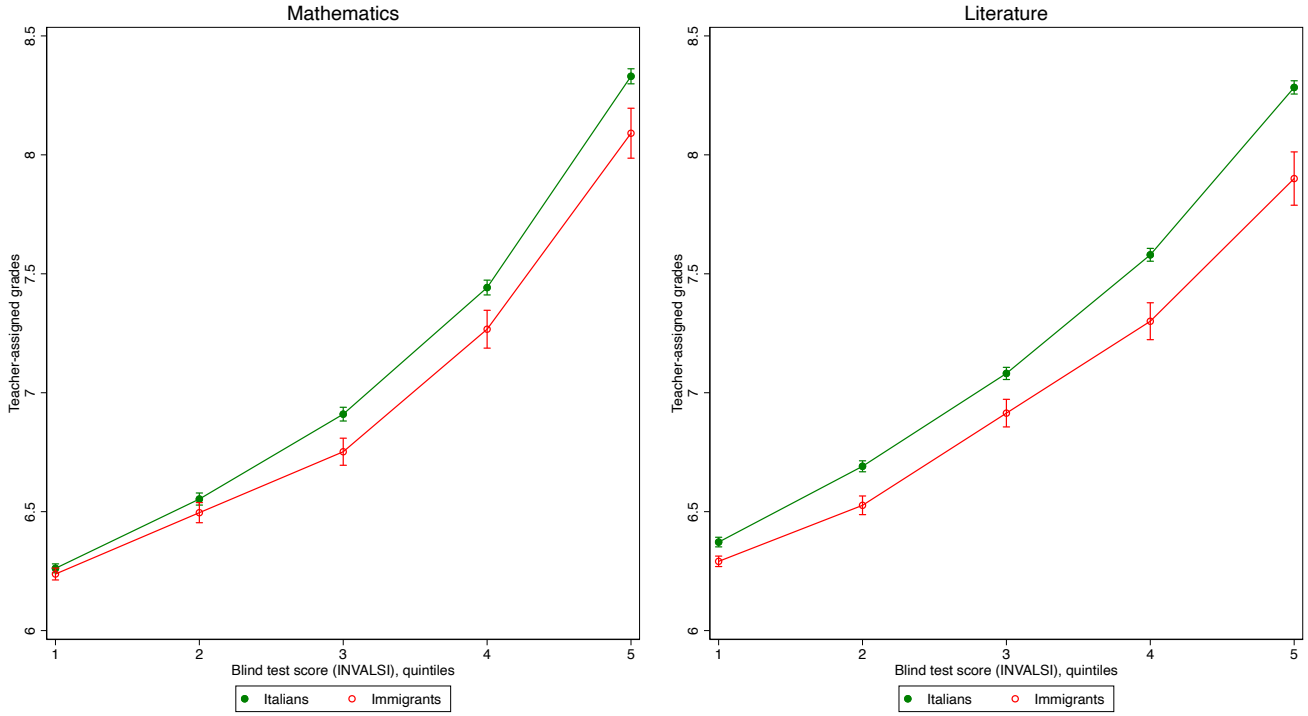
Notes: This graph shows the correlation between Immigrant-Native IAT score of the teacher and the bias in grading. In the left graph shows the IAT score of the teacher and naive estimate of bias in grading: the coefficient of the correlation is 0.08 (p-value: 0.163). The right graph shows the IAT score of the teacher and the Bayesian estimate of bias in grading: the coefficient of the correlation is 0.34 (p-value: 0.025). The description on how the measure is constructed is available in Appendix C.

Figure A.3: Distribution of grades



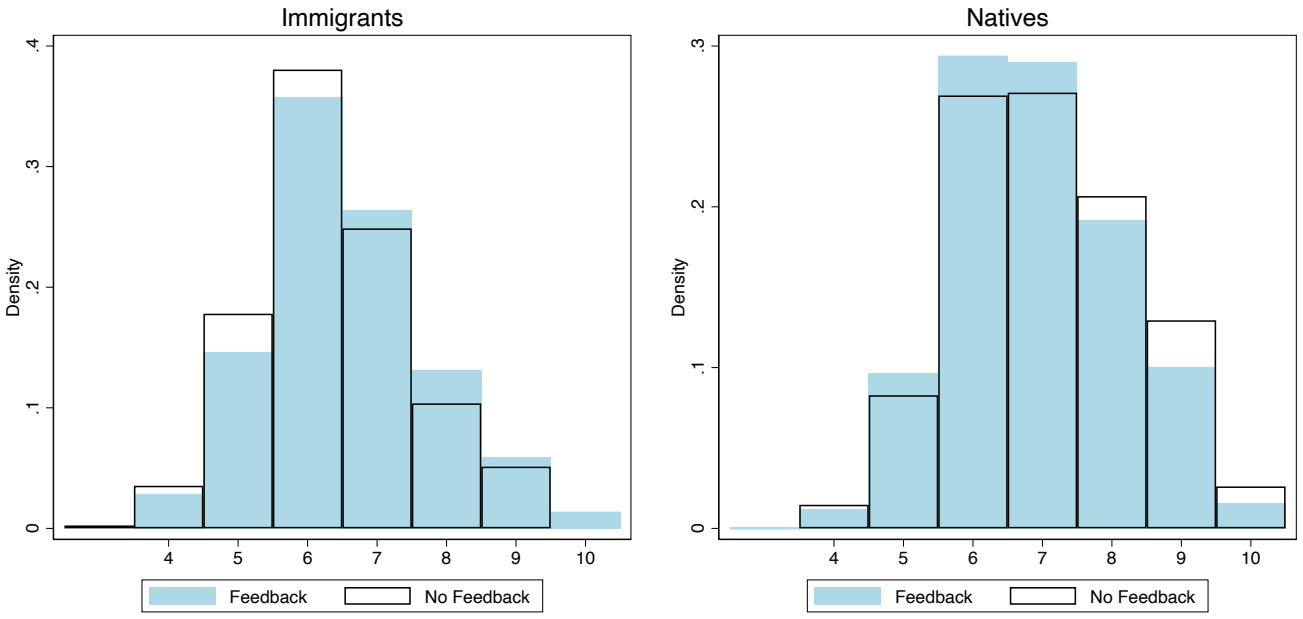
Notes: These graphs show the distribution of teacher-assigned grades (Panel A) and standardized test scores INVALSI (Panel B) in math and literature across native (blue line) and immigrant (red line) students. Students in this sample completed grade 8 between school years 2011–2012 and 2015–2016.

Figure A.4: Teacher-assigned grades vs. blindly graded, standardized test scores by subject



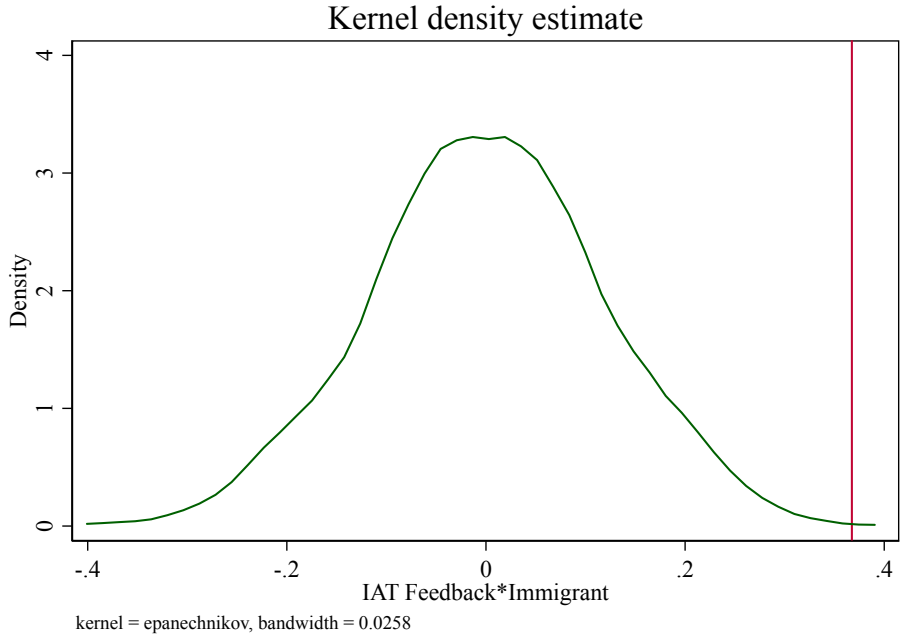
Notes: This graph shows teacher-assigned grades (non-blindly graded) on the vertical axis and quintiles of the standardized test score INVALSI (blindly graded) on the horizontal axis at the end of grade 8. Teacher-assigned grades are on a scale of 3 to 10, with 6 as the pass grade. The green squares and lines are for native students, while the red circles and lines are for immigrant students. Students in this sample completed grade 8 between school years 2011–2012 and 2015–2016.

Figure A.5: Field experiment: The impact of revealing stereotypes to teachers on grading



Notes: This graph shows the distribution of grades given to native and immigrant children by teachers eligible (light blue bars) and non-eligible (white bars) for receiving feedback about their own IAT scores before end-of-semester grading.

Figure A.6: Permutation test



Notes: This figure plots the distribution of the interaction term's coefficient "IAT Feedback*Immigrant" derived from a permutation test that runs the regression in Table 4 1,000 times, randomly assigning the treatment variable "IAT Feedback" to teachers, considering school-level clusters. The red line represents the observed coefficient from the main regression in column 1 of Table 4. In 6 out of 1,000 cases we find a coefficient higher than the one observed in Table 4. To perform the permutation test and plot the graph, we used the Stata package `ritest` (Heß, 2017), which allows us to specify permutation structures generated by clustered treatment assignments.

Table A.1: Country of birth of immigrant students from most represented nationalities (school year 2016–2017)

Place of Birth	**Number of Students**	**Share among Immigrant Students**
Romania	158,428	19.2%
Albania	112,171	13.6%
Morocco	102,121	12.4%
China	49,514	6.0%
Philippines	26,962	3.3%
India	25,851	3.1%
Moldavia	25,308	3.1%
Ukraine	19,956	2.4%
Pakistan	19,934	2.4%
Egypt	19,925	2.4%
Tunisia	18,613	2.3%
Peru	18,018	2.2%
Ecuador	16,153	2.0%
Macedonia	15,193	1.8%
Nigeria	14,853	1.8%

Source: Italian Ministry of Education. This table reports the total number of students by country of birth for the 15 most represented nationalities and their share among all immigrant students in the school year 2016–17.

Table A.2: Balance between schools in field experiment and out of the sample

	(1)	(2)	(3)	(4)	(5)
	All Students in Italy	Students in 5 Provinces in Northern Italy		p-Value (3)-(2)	Std. Diff. (3)-(2)
		Not in Sample	Exp. Sample		
Female	0.494	0.493	0.496	0.616	0.004
	(0.500)	(0.500)	(0.500)		
Immigrant	0.098	0.141	0.177	0.000	0.070
	(0.297)	(0.348)	(0.382)		
Immigrant (1st Gen)	0.051	0.075	0.066	0.011	-0.025
	(0.220)	(0.263)	(0.248)		
Immigrant (2nd Gen)	0.047	0.066	0.112	0.000	0.115
	(0.212)	(0.249)	(0.315)		
Test Score Grade 8	56.622	56.487	55.213	0.000	-0.045
	(19.046)	(19.081)	(20.534)		
Mother: Less Than Diploma	0.364	0.290	0.265	0.000	-0.039
	(0.481)	(0.454)	(0.441)		
Mother: Diploma	0.493	0.534	0.515	0.008	-0.027
	(0.500)	(0.499)	(0.500)		
Mother: More Than Diploma	0.143	0.176	0.220	0.000	0.078
	(0.350)	(0.381)	(0.414)		
Father: Less Than Diploma	0.429	0.360	0.330	0.000	-0.045
	(0.495)	(0.480)	(0.470)		
Father: Diploma	0.443	0.477	0.474	0.665	-0.004
	(0.497)	(0.499)	(0.499)		
Father: More Than Diploma	0.128	0.162	0.196	0.000	0.063
	(0.334)	(0.369)	(0.397)		
Mother: Low-Wage Occupation	0.565	0.463	0.460	0.657	-0.004
	(0.496)	(0.499)	(0.498)		
Mother: Intermediate-Wage Occupation	0.329	0.399	0.401	0.813	0.003
	(0.470)	(0.490)	(0.490)		
Mother: High-Wage Occupation	0.107	0.138	0.139	0.759	0.002
	(0.309)	(0.345)	(0.346)		
Father: Low-Wage Occupation	0.369	0.336	0.351	0.021	0.022
	(0.482)	(0.472)	(0.477)		
Father: Intermediate-Wage Occupation	0.410	0.412	0.411	0.882	-0.001
	(0.492)	(0.492)	(0.492)		
Father: High-Wage Occupation	0.222	0.252	0.237	0.019	-0.025
	(0.415)	(0.434)	(0.425)		
Class Size	22.089	22.489	22.193	0.000	-0.072
	(3.816)	(3.115)	(2.681)		
Observations	3,134,894	453,088	6,042		

The table shows the mean of the characteristics of all students in Italy (column 1) of students in schools from the five provinces of Milan, Turin, Genoa, and Padua, which were not included in the experiment (column 2) and schools included in the experiment (column 3). Column 4 shows the *p*-value of the mean difference and column 5 the normalized difference. In the experimental sample (column 3), the anonymized code for eight students do not match with the anonymized codes in the publicly available dataset. Hence, the number of observations in column 3 is 6,042 instead of 6,050. "Immigrant-Native IAT" is the d-score of the Implicit Association Test.

Table A.3: Balance table: Teacher characteristics (field experiment)

	(1) Full sample	(2) Not in the Sample	(3) Final Sample	(4) p-value	(5) Std. Diff.
Immigrant-Native IAT	0.469	0.450	0.477	0.202	0.073
	(0.262)	(0.264)	(0.261)		
Female	0.858	0.838	0.867	0.384	0.058
	(0.349)	(0.369)	(0.340)		
Teaching Math	0.484	0.459	0.494	0.154	0.050
	(0.500)	(0.499)	(0.500)		
Born in the North	0.646	0.599	0.665	0.150	0.097
	(0.479)	(0.491)	(0.473)		
Age	47.233	46.698	47.455	0.610	0.041
	(13.033)	(13.569)	(12.809)		
Full time contract	0.832	0.847	0.826	0.531	-0.040
	(0.374)	(0.361)	(0.380)		
Experience/10 years	1.942	1.911	1.955	0.702	0.026
	(1.182)	(1.164)	(1.191)		
Children	0.681	0.631	0.702	0.116	0.107
	(0.466)	(0.484)	(0.458)		
Low edu Mother	0.462	0.495	0.448	0.267	-0.067
	(0.499)	(0.501)	(0.498)		
Middle edu Mother	0.307	0.320	0.301	0.657	-0.029
	(0.462)	(0.467)	(0.459)		
High edu Mother	0.135	0.099	0.150	0.074	0.110
	(0.342)	(0.299)	(0.357)		
Degree Laude	0.230	0.198	0.243	0.132	0.077
	(0.421)	(0.400)	(0.430)		
WVS Immigrants' Rights to Job	0.585	0.563	0.594	0.477	0.044
	(0.493)	(0.497)	(0.492)		
Reason Gap: Prejudice	0.221	0.203	0.228	0.418	0.043
	(0.415)	(0.403)	(0.420)		
Reason Gap: Economic	0.640	0.595	0.659	0.042	0.094
	(0.480)	(0.492)	(0.474)		
Reason Gap: Behavior	0.192	0.171	0.200	0.293	0.053
	(0.394)	(0.378)	(0.401)		
Reason Gap: Ability	0.201	0.234	0.187	0.152	-0.082
	(0.401)	(0.424)	(0.390)		
Reason Gap: Language	0.493	0.523	0.481	0.312	-0.059
	(0.500)	(0.501)	(0.500)		
Reason Gap: Information	0.238	0.221	0.245	0.508	0.040
	(0.426)	(0.416)	(0.431)		
Observations	756	222	534		

Notes: The table shows the mean of the characteristics of the full sample of teachers for the field experiment (column 1), teachers not in the final sample (column 2), and teachers who are in the final sample of the experiment, i.e., the sample of teachers in schools that participated in the field experiment and taught 9th graders in 2017–18 (column 3). Standard deviations are in parentheses in columns 1, 2, and 3, and the *p*-value of the difference is in column 4. Standard errors are clustered at the school level. "Immigrant-Native IAT" is the d-score of the Implicit Association Test. "WVS Immigrants' Rights to Job" equals 1 for teachers believing that immigrants should have the same right to jobs as natives. "Reason Gap" represents a list of potential reasons for the immigrant-native gap in high-school track choice.

Table A.4: Balance table: Students' characteristics (field experiment)

	(1) Full sample	(2) Not in the Sample	(3) Final Sample	(4) p-value	(5) Std. Diff.
Female	0.491	0.480	0.495	0.214	0.021
	(0.500)	(0.500)	(0.500)		
Immigrant	0.206	0.255	0.184	0.002	-0.122
	(0.404)	(0.436)	(0.388)		
High Education Mother	0.176	0.139	0.192	0.130	0.101
	(0.381)	(0.346)	(0.394)		
High-Wage Occupation Mother	0.115	0.103	0.120	0.563	0.038
	(0.319)	(0.304)	(0.325)		
Medium-Wage Occupation Mother	0.331	0.290	0.348	0.017	0.088
	(0.470)	(0.454)	(0.476)		
High Education Father	0.156	0.131	0.166	0.313	0.070
	(0.363)	(0.338)	(0.372)		
High-Wage Occupation Father	0.193	0.178	0.199	0.608	0.038
	(0.395)	(0.383)	(0.400)		
High-Wage Occupation Father	0.338	0.310	0.351	0.084	0.062
	(0.473)	(0.463)	(0.477)		
Grade Math June '16	7.182	7.225	7.163	0.242	-0.034
	(1.259)	(1.307)	(1.238)		
Grade Ital. June '16	7.131	7.139	7.127	0.799	-0.008
	(1.054)	(1.068)	(1.049)		
Observations	8,472	2,630	6,050		

Notes: The table shows the mean of the characteristics of the full sample of students for the field experiment (column 1), students not in the final sample (column 2), and students who are in the final sample of the experiment, i.e., students in schools that participated in the field experiment and were in the 9th grade in 2017–18 (column 3). Standard deviations are in parentheses in columns 1, 2, and 3, and the p-value of the difference is in column 4. Standard errors are clustered at the school level.

Table A.5: Correlation between teacher characteristics and willingness to receive feedback

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable: Dummy for whether the teacher wants to receive the feedback							
Immigrant-Native IAT	0.004					0.000	0.031
	(0.032)					(0.033)	(0.032)
Teaching Math		0.026				0.021	0.029
		(0.022)				(0.022)	(0.023)
Female			0.003			0.004	0.020
			(0.031)			(0.031)	(0.028)
WVS Immigrants' Rights to Job				-0.036		-0.034	-0.002
				(0.029)		(0.030)	(0.029)
Time Survey: slow					0.053*	0.053*	0.014
					(0.031)	(0.031)	(0.032)
Time Survey: fast					-0.017	-0.015	0.004
					(0.053)	(0.054)	(0.047)
Time Survey: missing					-0.096**	-0.093**	-0.032
					(0.046)	(0.046)	(0.050)
FE school	No	No	No	No	No	No	Yes
Mean dep.	0.78	0.78	0.78	0.78	0.78	0.78	0.78
Obs.	1384	1384	1384	1384	1384	1384	1384
R^2	0.000	0.001	0.000	0.001	0.004	0.006	0.247

Notes: The table shows the correlations between whether the teacher decided to receive the feedback on their own IAT score and teacher characteristics. Robust standard errors clustered at the school level are in parentheses. All columns include dummy variables for missing characteristics (if any). "Immigrant-Native IAT" is the d-score of the Implicit Association Test. "Time Survey: Fast" equals 1 for teachers who took fewer than 11 minutes to complete the survey. "Time Survey: Slow" equals 1 for teachers who took more than 20 minutes to complete the survey. The average completion time is around 15.5 minutes. "Time Survey: Missing" indicates that a teacher did not complete the survey with the tablet and only did the IAT. "WVS Immigrants' Rights to Job" equals 1 for teachers believing that immigrants should have the same right to jobs as natives.

Table A.6: Correlation between teacher characteristics and IAT

	(1)	(2)	(3)	(4)	(5)
Dep. Var.: IAT Score (Stereotypes against Immigrants) in Field Experiment					
Children	-0.006			0.004	0.004
	(0.014)			(0.016)	(0.017)
Middle Edu Mother		0.027		0.027	0.030
		(0.017)		(0.018)	(0.019)
High Edu Mother		-0.022		-0.025	-0.032
		(0.021)		(0.022)	(0.023)
Reason Gap: Economic			-0.008	-0.000	0.003
			(0.015)	(0.016)	(0.017)
Reason Gap: Behavior			-0.002	-0.003	-0.006
			(0.018)	(0.019)	(0.020)
Reason Gap: Ability			0.023	0.019	0.035
			(0.020)	(0.020)	(0.022)
Reason Gap: Language			0.017	0.022	0.014
			(0.015)	(0.015)	(0.017)
Reason Gap: Information			-0.009	-0.009	-0.013
			(0.016)	(0.017)	(0.018)
Reason Gap: Prejudice			0.032*	0.033*	0.031
			(0.018)	(0.018)	(0.020)
Experience/10 Years				0.000	0.002
				(0.007)	(0.007)
Female				-0.040**	-0.045**
				(0.020)	(0.021)
Born in the North				-0.024*	-0.020
				(0.015)	(0.016)
WVS Immigrants' Rights to Job				-0.054***	-0.047**
				(0.016)	(0.019)
IAT Order Controls	Yes	Yes	Yes	Yes	Yes
Obs.	1,384	1,384	1,384	1,384	1,384
R^2	0.060	0.065	0.065	0.085	0.152

Notes: This table reports OLS estimates, where the dependent variable is the Immigrant-Native IAT score of teachers and the unit of observation is teacher *t* in school *s*. We include controls for the order of IATs and for whether the blocks were presented in an order-compatible or order-incompatible way (which was randomized at the individual level). The variable "WVS Immigrants' Rights to Job" equals 1 for teachers believing that immigrants should have the same right to jobs as natives.

Table A.7: Balance table: Student characteristics (field experiment)

	(1) Full Sample	(2) Control	(3) Treated	(4) p-Value	(5) Norm. Diff.
Female	0.495	0.502	0.490	0.408	-0.017
	(0.500)	(0.500)	(0.500)		
Immigrant	0.184	0.174	0.193	0.497	0.035
	(0.388)	(0.379)	(0.395)		
First-Gen Imm	0.084	0.079	0.088	0.568	0.023
	(0.277)	(0.270)	(0.283)		
Grade Ital. June '16	7.127	7.141	7.116	0.724	-0.017
	(1.049)	(1.052)	(1.046)		
Grade Math June '16	7.163	7.198	7.134	0.393	-0.037
	(1.238)	(1.248)	(1.228)		
Grade ItaL. June '15	7.203	7.231	7.180	0.427	-0.034
	(1.053)	(1.052)	(1.054)		
Grade Math June '15	7.337	7.369	7.309	0.380	-0.033
	(1.287)	(1.287)	(1.287)		
Low Education Mother	0.231	0.205	0.254	0.207	0.083
	(0.422)	(0.404)	(0.435)		
High Education Mother	0.192	0.166	0.213	0.271	0.085
	(0.394)	(0.372)	(0.410)		
Mother Low Skill	0.160	0.143	0.174	0.161	0.060
	(0.366)	(0.350)	(0.379)		
Mother Mid-Skill	0.348	0.342	0.353	0.754	0.016
	(0.476)	(0.475)	(0.478)		
Mother High-Skill	0.120	0.100	0.137	0.257	0.081
	(0.325)	(0.300)	(0.344)		
Low Education Father	0.281	0.255	0.302	0.288	0.074
	(0.449)	(0.436)	(0.459)		
High Education Father	0.166	0.152	0.178	0.556	0.049
	(0.372)	(0.360)	(0.383)		
Low-Wage Occupation Father	0.258	0.244	0.271	0.467	0.044
	(0.438)	(0.429)	(0.444)		
Medium-Wage Occupation Father	0.351	0.341	0.359	0.615	0.027
	(0.477)	(0.474)	(0.480)		
High-Wage Occupation Father	0.199	0.178	0.217	0.442	0.069
	(0.400)	(0.383)	(0.412)		
Observations	6,050	2,775	3,275		

Notes: The table shows the mean of the characteristics of the full sample of students for the field experiment (column 1), students in the control group (column 2), and students in the treatment group (column 3). Standard deviations are in parentheses in columns 1, 2, and 3, and the *p*-value of the difference is in column 4. The last column reports the normalized difference between group averages. If both the math and literature teacher participate in the experiment, there is only one student-level observation used for this table. Standard errors are clustered at the school level.

Table A.8: Bias in grading and teachers' IAT scores

Outcome: First Difference, Std Grade–Std Test Score									
	All			High Ability			Low Ability		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Immigrant	-0.079***	-0.050**	0.405	-0.154***	-0.100***	1.395**	-0.048***	-0.035	0.493
	(0.010)	(0.023)	(0.402)	(0.017)	(0.036)	(0.644)	(0.011)	(0.025)	(0.498)
IAT* Immigrant		-0.063	-0.054		-0.116*	-0.119*		-0.026	-0.026
		(0.043)	(0.043)		(0.068)	(0.067)		(0.051)	(0.049)
Obs.	42,302	42,302	42,302	25,415	25,415	25,415	16,867	16,867	16,867
R^2	0.357	0.357	0.391	0.213	0.213	0.264	0.447	0.447	0.473
Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
INVALSI Cubic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	No	No	Yes	No	No	Yes	No	No	Yes
Student Controls*Imm	No	No	Yes	No	No	Yes	No	No	Yes
Teacher Controls*Imm	No	No	Yes	No	No	Yes	No	No	Yes

Notes: This table reports OLS estimates, where the dependent variable is the standardized difference between teacher-assigned grades and test scores (INVALSI). The unit of observation is student *i* taught by teacher *t* in school *s*. "Immigrant" indicates whether the student is not Italian citizen. "IAT" indicated the Immigrant-Native IAT (d-score). Student controls include gender, generation of immigration, mother education, and province. Columns 1-3 provides the estimates for the full sample, 4-6 for high-ability students, and 7-9 for low-ability students, with a sample split based on the standardized test score INVALSI. Teacher controls include gender, place of birth, age, and age squared. Students in this sample completed grade 8 between school years 2011–2012 and 2015–2016. Standard errors are robust and clustered at the teacher level.

Table A.9: Estimation of the impact of revealing stereotypes to teachers on student grades

Dep. Var.: Teacher-Assigned Grade (Transformed)			
	(1)	(2)	(3)
IAT Feedback*Immigrant	0.226***	0.236***	0.232***
	(0.069)	(0.059)	(0.060)
Immigrant	-0.629***	-0.640***	-0.177
	(0.040)	(0.088)	(0.631)
IAT Feedback	-0.112*	-0.126**	-0.118**
	(0.057)	(0.051)	(0.053)
Student Controls	No	Yes	Yes
Teacher Controls	No	No	Yes
Obs.	10279	10279	10279
R^2	0.053	0.151	0.155

Notes: This table reports OLS estimates for teacher-assigned grades, transformed to map the grades for the end of the first semester to the grades of the end of the second semester. Robust standard errors clustered at the school level are in parentheses. "Immigrant" is a dummy variable that assumes value 1 if the student is from an immigrant background. "IAT Feedback" is a dummy variable indicating whether the teacher was eligible for receiving the feedback before end-of-semester grading (January) or after end-of-semester grading (February). Student controls (also interacted with immigrant controls) include gender, generation of immigration, year birth, mother education, and province. Teacher controls (also interacted with immigrant controls) include gender, born north, age, and age squared.

Table A.10: Estimation of the impact of revealing stereotypes to teachers on student grades in the field experiment

Dependent Variable: Teacher Assigned Grades	(1)	(2)	(3)
IAT Feedback*Immigrant	0.367***	-0.051	0.289***
	(0.096)	(0.158)	(0.096)
Immigrant	0.294	0.178	0.247
	(0.940)	(0.880)	(0.964)
IAT Feedback	-0.153*	-0.061	-0.122
	(0.079)	(0.098)	(0.084)
IAT Feedback*WVS*Immigrant		0.581***	
		(0.177)	
IAT Feedback*WVS		-0.155*	
		(0.086)	
IAT Feedback*Reason Gap Prejudice*Immigrant			0.325*
			(0.179)
IAT Feedback*Reason Gap Prejudice			-0.116
			(0.099)
Obs.	10279	10279	10279
R^2	0.131	0.133	0.134
Mean Control Natives	6.57	6.57	6.57
Mean Control Immigrants	5.86	5.86	
Student Controls	Yes	Yes	Yes
Student Controls*Imm	Yes	Yes	Yes
Teacher Controls	Yes	Yes	Yes
Teacher Controls*Imm	Yes	Yes	Yes

Notes: This table reports OLS estimates, where the dependent variable is the grade at the end of the first semester of grade 8 (January). The unit of observation is student i in class c taught by teacher t in grade 8 of school s. Standard errors are robust and clustered at the school level. "Immigrant" is a dummy variable that assumes value 1 if the student is from an immigrant background. "IAT Feedback" is a dummy variable indicating whether the teacher was eligible for receiving the feedback before end-of-semester grading (January) or after end-of-semester grading (February). "WVS" equals 1 for teachers who agree with the statement that "immigrants and natives should have equal opportunities to access available jobs. "Reason Gap Prejudice" equals 1 for teachers who agree that prejudice is one of the factors explaning the differences in high-school track choice of natives and immigrants. Student controls include gender, generation of immigration, and education of the mother, all interacted with whether the student is an immigrant. Teacher controls include gender, place of birth, age, and age squared, interacted with whether the student is an immigrant.

Table A.11: Beliefs Updating in the Online Experiment

Dep. Var: Teacher-Assigned Grade				
	(1)	(2)	(3)	(4)
IAT Feedback	0.005	-0.044	-0.073	-0.018
	(0.173)	(0.166)	(0.166)	(0.178)
Immigrant	0.420***	0.352**	0.389**	0.426***
	(0.141)	(0.144)	(0.149)	(0.159)
IAT Feedback × Immigrant	-0.580***	-0.458**	-0.471**	-0.546***
	(0.196)	(0.188)	(0.185)	(0.202)
IAT Score	-0.068		-0.332	-0.232
	(0.130)		(0.207)	(0.236)
IAT Feedback × IAT Score	-0.329*			-0.261
	(0.171)			(0.465)
Immigrant × IAT Score	-0.426***		-0.137	-0.272
	(0.157)		(0.287)	(0.414)
IAT Feedback × Immigrant × IAT Score	0.849***			0.350
	(0.241)			(0.559)
(IAT-Expected IAT)		-0.034	0.201	0.132
		(0.123)	(0.206)	(0.229)
IAT Feedback × (IAT-Expected IAT)		-0.270*	-0.243	-0.043
		(0.161)	(0.160)	(0.420)
Immigrant × (IAT-Expected IAT)		-0.356**	-0.261	-0.167
		(0.156)	(0.280)	(0.361)
IAT Feedback × Immigrant × (IAT-Expected IAT)		0.738***	0.753***	0.486
		(0.226)	(0.223)	(0.514)
Constant	5.851***	5.814***	5.979***	5.954***
	(0.324)	(0.303)	(0.325)	(0.325)
Control Mean	7.134	7.134	7.134	7.134
Obs.	1460	1460	1460	1460
R^2	0.450	0.453	0.455	0.455
Subject, Order, Original Grade FE	Yes	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes	Yes
Teacher Controls	Yes	Yes	Yes	Yes

Notes: This table reports OLS estimates, where the dependent variable is the grade assigned by teachers in the online experiment. The unit of observation is student i by teacher t. Standard errors are robust and clustered at the school level (the unit of randomization). "Immigrant" is a dummy variable that assumes value 1 if the student is from an immigrant background. "IAT Feedback" is a dummy variable indicating whether the teacher was eligible for receiving the IAT feedback versus the active control message. "IAT Score" is a continuous variable indicating the standard d-score of the IAT test (more details available on Apendix B.1). "IAT-Expected IAT" is a continuous variable calculated as the difference between IAT score and the expected score. The expected score is the average of the score in each IAT category. For the "expected severely biased category" we imputed the average IAT score of the teachers with $IAT > 0.6$. Student controls include gender and class. Teacher controls include gender, place of birth, and a dummy for whether the teacher completed the IAT before the first reminder.

B Online Appendix

B.1 Description of the IAT

The IAT that we developed for this study associates immigrant/native names with positive/negative adjectives in the specific schooling context. As usual in the IATs, it presents two sets of stimuli. The first set includes typical Italian names (e.g., Francesca or Luca) and common names among immigrant children in Italy (e.g., Fatima or Mohamed), respectively. The second set consists of positive adjectives (e.g., smart) and negative ones (e.g., lazy).

One word at a time (either a name or an adjective) appeared at the center of the screen, and individuals were instructed to categorize it to the left or to the right according to different labels displayed on the top of the screen. For instance, the right label might have said "Immigrant," and the left one might have said "Italian." Names and adjectives randomly appeared at the center of the screen, and subjects were asked to categorize the words as quickly as possible. In one type of round, subjects were asked to categorize native-sounding names and negative adjectives to the same side of the screen, whereas in another, they were asked to categorize immigrant-sounding names and negative adjectives to the same side. The order of the two types of rounds was randomly selected at the individual level. Each teacher in our survey completed two immigrant-native IATs, one using male names and one using female names, and the order of the IAT with male and female names was randomized at individual level.

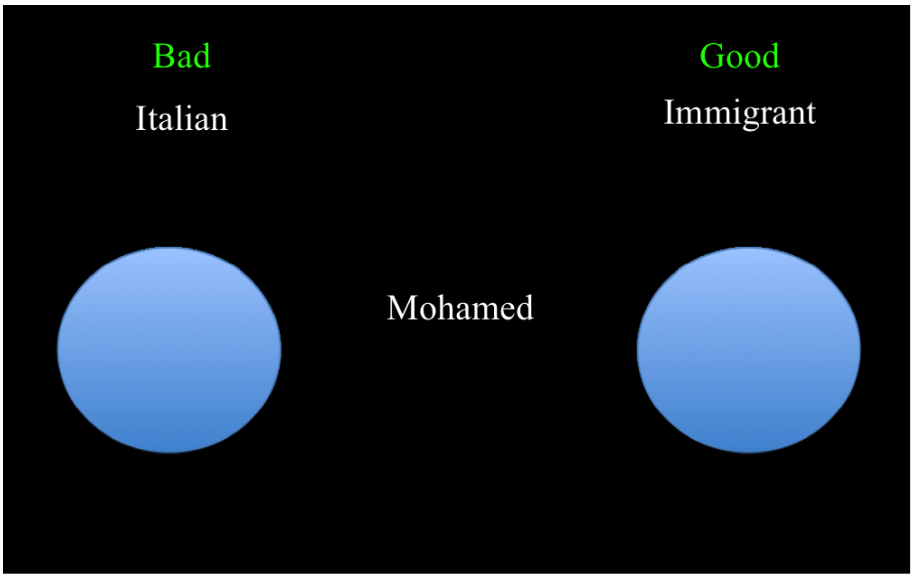
The IAT comprises seven blocks. Half of the teachers randomly selected at the individual level and completed the IAT in the order as presented in Table B.1 ("order-compatible" task first), while the other half completed the IAT with the blocks in the following order: 1, 5, 6, 7, 2, 3, and 4 ("order-incompatible" task first). Figure B.1 presents a sample screenshot of the latter task, while all the words presented to teachers are shown in the box below (with the original in Italian in parentheses). On average, there is a small difference in the IAT score between individuals who performed the order-compatible task first versus the order-incompatible task first. Hence, in all regressions where there are no teacher fixed effects, we control for whether the first task was order compatible.

The blocks used to calculate the IAT score are blocks 3, 4, 6, and 7. The number of words that need to be categorized is 20 in blocks 3 and 6 and 40 in blocks 4 and 7, as in the standard IAT with 7 blocks. The scoring procedure follows the guidelines of the improved scoring algorithm defined by Greenwald et al. (2003).

Table B.1: Schematic overview of the immigrant IAT

Blocks	Left Categories	Right Categories
1	Italian	Immigrant
2	Good	Bad
3	Italian Good	Immigrant Bad
4	Italian Good	Immigrant Bad
5	Bad	Good
6	Italian Bad	Immigrant Good
7	Italian Bad	Immigrant Good

Figure B.1: Example of the screenshot of the tablet in the "order-incompatible" task



- **IAT with male names of immigrants and natives**

  1. Immigrant *(Immigrato)*: Youssef, Mohamed, Gheorghe, Alejandro, Li Yi, Pascual
  2. Italian *(Italiano)*: Marco, Simone, Daniele, Francesco, Lorenzo, Mattia
  3. Good *(Bravo)*: Prepared *(Preparato)*, Intelligent *(Intelligente)*, Capable *(Capace)*, Studious *(Studioso)*, Able *(Abile)*, Precise *(Attento)*, Willing *(Volenteroso)*, Respectful *(Rispettoso)*
  4. Bad *(Impreparato)*: Disrespectful *(Irrispettoso)*, Slow *(Tardo)*, Incapable *(Incapace)*, Boisterous *(Irrequieto)*, Lazy *(Pigro)*, Distracted *(Distratto)*, Demotivated *(Demotivato)*, Insufficient *(Scarso)*

- **IAT with female names of immigrants and natives**

  1. Immigrant *(Immigrata)*: Fatima, Naila, Adina, Iryna, Jiaxin, Beatriz
  2. Italian *(Italiana)*: Valentina, Sara, Giorgia, Francesca, Elisa, Alice
  3. Good *(Brava)*: Prepared *(Preparata)*, Intelligent *(Intelligente)*, Capable *(Capace)*, Studious *(Studiosa)*, Able *(Abile)*, Precise *(Attenta)*, Willing *(Volenterosa)*, Respectful *(Rispettosa)*
  4. Bad *(Impreparata)*: Disrespectful *(Irrispettosa)*, Slow *(Tarda)*, Incapable *(Incapace)*, Boisterous *(Irrequieta)*, Lazy *(Pigra)*, Distracted *(Distratta)*, Demotivated *(Demotivata)*, Insufficient *(Scarsa)*

- **Online experiment: IAT immigrant-native (both male and female names)**

  1. Immigrant *(Immigrato)*: Fatima, Mohamed, Adina, Alejandro, Jiaxin, Pascual
  2. Italian *(Italiano)*: Valentina, Simone, Giorgia, Francesco, Elisa, Mattia
  3. Good *(Bravo)*: Prepared *(Preparato)*, Intelligent *(Intelligente)*, Capable *(Capace)*, Studious *(Studioso)*, Able *(Abile)*, Precise *(Attento)*, Willing *(Volenteroso)*, Respectful *(Rispettoso)*
  4. Bad *(Impreparato)*: Disrespectful *(Irrispettoso)*, Slow *(Tardo)*, Incapable *(Incapace)*, Boisterous *(Irrequieto)*, Lazy *(Pigro)*, Distracted *(Distratto)*, Demotivated *(Demotivato)*, Insufficient *(Scarso)*

B.2 Teacher questionnaire

B.2.1 Field experiment

1) *Immigrant children, with the same grades of natives, are more likely to choose a vocational track. According to your experience, how much do you think these factors affect the choice of immigrants? Answers on a scale of 1 to 5.*

1. *Economic reasons*
2. *Bad behavior at school*
3. *Insufficient abilities for more demanding schools*
4. *Knowledge of the language*
5. *No information about educational and occupational careers*
6. *Perception of prejudices in school or at work*

2) Do you agree or disagree with the following statements? When jobs are scarce, employers should give priority to Italian people over immigrants. Possible answers: Agree, Neither agree nor disagree, Disagree, Don't know

B.2.2 Online experiment: Baseline

SECTION 0: Introduction **Note: The survey is sent as a unique link to the contact information on teachers. We do not need to ask the school name.**		
Dear Teacher, Thank you so much for agreeing to participate in this research study. We ask you to complete this first survey by (DATE1). It will take less than 15 minutes. Later, we will ask you to help us grade some questions in the subject you teach between (DATE2) and (DATE3). This will take no longer than 45 minutes. To thank you for your time, you will receive an **Amazon gift card of 40 euros** after you complete both parts of the research study. Thank you in advance for your collaboration. Best regards, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti		
0.0 Consent form to teachers		
0.1 GDPR		
0.2	You are:	• Male • Female
0.3	Where were you born?	• Province: • Abroad (country):
0.4 How many years have you been teaching? Dropdown menu from 0 to 40, "More than 40 years"		

Table B.2 – *Continued on next page*

Table B.2 – *Continued from previous page*

0.5	In which subject have you obtained a university degree?	• I did not obtain a university degree • Math • Biology/natural sciences • Physics/chemistry/ astronomy • Languages • Literature • Psychology • Engineering • Education • Philosophy • History • Geography/geology • Other degree: _____

Table B.2 – *Continued on next page*

Table B.2 – *Continued from previous page*

0.6	Do you have special responsibilities within the school?	• Vice principal • Math area chair • Literature area chair • English area chair • Math games • Responsible for career counseling
0.7 In which classes have you taught during the school year 2020–21? Add list of classes (1A, 2A)		
SECTION 1: IAT (immigrant-native, bad-good IAT)		
SECTION 2: Self-perception: Now we would like to ask you some questions about your general opinions and about your perceptions of the task you just performed.		
0.4 How many years have you been teaching? Dropdown menu from 0 to 40, "More than 40 years"		
2.1	When jobs are scarce, employers should give priority to people of this country over immigrants.	• Strongly agree • Agree • Disagree • Strongly disagree

Table B.2 – *Continued on next page*

Table B.2 – *Continued from previous page*

2.2	There are innate difference in the math skills of men and women.	• Strongly agree • Agree • Disagree • Strongly disagree
2.3 Sorting names of immigrants with good (and natives with bad) has been		**Sorting names of immigrants with bad (and natives with good) has been**
• A lot easier • Moderately easier • Slightly easier • The same • Slightly easier • Moderately easier • A lot easier		
2.4 Sorting names of females with scientific subjects (and males with humanistic subjects) has been		**Sorting names of females with humanistic (and males with scientific) has been**
• A lot easier • Moderately easier • Slightly easier • The same • Slightly easier • Moderately easier • A lot easier		
SECTION 3: Grading questions		

Table B.2 – *Continued on next page*

Table B.2 – *Continued from previous page*

3.1	Immigrant students are more likely to choose a vocational track in high school compared to natives even when they do equally well in middle school. Based on your experience, how much can these factors influence the choice of immigrants? 1. Economic reasons 2. Problems related to behavior at school 3. Ability not sufficient for more difficult high schools 4. Knowledge of Italian language 5. Absence of information on education or occupation opportunities 6. Perception of prejudices in school/work	• Very much • Much • Sufficiently • A bit • Not at all
3.2	When you grade your students at the end of the semester, how much weight do you assign to the following aspects? (Choose the weights to sum to 100. There are no right or wrong answers; it depends on your teaching style.) 1. Grades in written exams in class _____ 2. Grades in oral exams in class _____ 3. Attention and behavior in class _____ 4. Diligence in doing the homework _____	
Thank you very much for your participation!		

B.2.3 Online experiment: Endline

SECTION 0: Introduction **Note: The survey is sent as a unique link to the contact information for teachers.**
Dear Teacher, Thank you so much for agreeing to participate in this research study. We will ask you to help us grading some questions in the subject you teach. Please complete the task by February 28. To thank you for your time, after the grading, you will receive an Amazon gift card of 40 euros. Thank you in advance for your collaboration. Best regards, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti
SECTION 1. Each teacher will see the answer on one question from 10 students (4 with immigrant names, 6 with native names). **They will need to grade each question on a scale from 3 to 10 (as usual in the Italian schooling system).**
SECTION 2: Explicit bias questions

Table B.3 – *Continued on next page*

Table B.3 – *Continued from previous page*

2.1	Immigrant students are more likely to choose a vocational track in high school compared to natives even when they do equally well in middle school. Based on your experience, how much can these factors influence the choice of immigrants? 1. Economic reasons 2. Problems related to behavior at school 3. Ability not sufficient for more difficult high schools 4. Knowledge of Italian language 5. Absence of information on education or occupation opportunities 6. Perception of prejudices in school/work	• Very much • Much • Sufficiently • A bit • Not at all
2.2	When jobs are scarce, employers should prioritize people from their own country over immigrants.	• Totally agree • Agree • Disagree • Totally disagree

B.3 Email with the feedback

B.3.1 Field experiment

The exact wording of the email with the feedback about one's own implicit bias is reported in this appendix translated in English. Instead of the XXX, teachers saw the precise score (e.g., 0.25). We followed the standard categorization of IAT scores (Greenwald et al., 2009): no association if the score is between –0.15 and 0.15, slight association for values between |0.15| and |0.35|, moderate association between |0.35| and |0.60|, and strong association for scores higher than |0.60|.

Subject: Result of the Implicit Association Test – Research Project of Bocconi University

Dear teacher,

As per your request, we are writing you to let you know your result of the Implicit Association Test that you completed during the questionnaire administered by Bocconi University and related to the research titled "The role of teachers in high school track choice." You did this test using a tablet in the school building where you work. The Implicit Association Test was administered to teachers in middle school to measure and increase the awareness of potential unconscious preferences or associations.

Implicit Association Test: this test investigates the automatic associations between immigrant and Italian names with positive associations (e.g., good) and negative associations (e.g., bad). You completed this test separately with male and female names.
Your immigrant-native Implicit Association Test score using male names of natives and immigrants is XXX, which suggests a (slight/moderate/strong) association between positive attributes and Italian/immigrant names, and between negative attributes and immigrant/Italian names (or no automatic associations between positive attributes and Italian or immigrant names).
Your immigrant-native Implicit Association Test score using female names of natives and immigrants is XXX, which suggests a (slight/moderate/strong) association between positive attributes and Italian/immigrant names, and between negative attributes and immigrant/Italian names (or no automatic associations between positive attributes and Italian or immigrant names).

We want to underscore that this test reveals implicit attitudes and not behaviors. Our attitudes may derive from the cultural and social context where we live, and it is not obvious that explicit and implicit behaviors coincide. All of your responses will be held in confidence: only the researchers involved in this study will have access to the information you provide. Your responses will not be shared with other people. Data collected will be published in aggregate form, and it will not be possible to link them with the teacher or the school. We hope that you found this test useful. Thank you for the time you dedicated to our research.

The Research Team

B.3.2 Online experiment

TREATMENT 1: Active control group
Subject: Research Project of Bocconi and Harvard University

Dear teacher,

A few weeks ago, you completed an online questionnaire administered by researchers at Bocconi and Harvard University. We are writing you to confirm that we received the first part of the questionnaire to share some additional information.

An enormous body of literature confirms that we all have biases—some explicit, many implicit. However, it is important to avoid our implicit biases or stereotypes related to a specific group from systematically influencing our behavior toward students, thus influencing a child's self-image or burdening him/her with low expectations that will make the child feel lacking or inadequate. Acknowledging and understanding our biases and those of our colleagues can help minimize the influence they have on our daily interaction with students, including our encouragements and disciplinary procedures, teachers' track recommendations, and grades.

Thank you for the time you dedicated to our research. In about a month we will send you the last part of the questionnaire. To thank you for your time, you will receive a 40 euro Amazon gift card after completing the last part of the research study as well.

Many thanks,
The Research Team

TREATMENT 2: Reveal own bias treatment

Subject: Research Project of Bocconi and Harvard University

Dear teacher,

A few weeks ago, you completed an online questionnaire administered by researchers at Bocconi and Harvard University. We are writing you to confirm that we received the first part of the questionnaire and to share some additional information.

The survey included an Implicit Association Test, a tool used in social psychology to measure and increase the awareness of potential preferences or unconscious associations.

We are reporting below the result of the Implicit Association Test that you completed.

This test was aimed at investigating the automatic associations between immigrant and Italian names with positive associations (e.g., good) and negative associations (e.g., bad).

Your immigrant-native Implicit Association Test score using names of Italians and immigrants is XXX, which suggests a (slight/moderate/strong) automatic association between positive attributes and Italian/immigrant and negative attributes and immigrant/Italian (or no automatic associations between positive attributes and Italian or immigrant).

We want to iterate that this test reveals implicit attitudes and not behaviors. Our attitudes may derive from the cultural and social context where we live, and it is not obvious that explicit and implicit attitudes coincide. We remind you that all of your responses will be held in confidence: only the researchers involved in this study will have access to the information you provide. Your responses will not be shared with other people. Data collected will be published in aggregate form, and it will not be possible to link them with the teacher or the school. We hope that you found this test to be useful.

An enormous body of literature confirms that we all have biases—some explicit, many implicit. However, it is important to avoid our implicit biases or stereotypes related to a specific group from systematically influencing our behavior toward students, thus influencing a child's self-image or burdening him with low expectations that will make the child feel lacking or inadequate. Acknowledging and understanding our biases and those of our colleagues can help minimize the influence they have on our daily interaction with students, including our encouragements and disciplinary procedures, teachers' track recommendations, and grades.

Thank you for the time you dedicated to our research. In about a month we will send you the last part of the questionnaire. To thank you for your time, you will receive a 40 euro Amazon gift card after completing the last part of the research study as well.

The Research Team

B.4 Examples of grading task in math, Italian, and English

Figure B.2: Grading task in math

Una fabbrica di cioccolato produce cioccolatini a forma di piramide con le seguenti dimensioni:

base quadrata di lato 2,7 cm;
altezza di 3 cm;
peso specifico di 0,48 g/cm3.

Ogni kilogrammo di cioccolato, quanti cioccolatini produrrà?

Risposta 3
Nome:
Classe:

Dati

$l = 2,7$ cm
$h = 3$ cm
$P_s = 0,48\ g/cm^3$

Richiesta
numero di cioccolatini

Svolgimento

Calcolo il volume della piramide

$$V = \frac{A_b \times h}{3} = \frac{3 \times 3 \times 2,7}{3}\ cm^3 = 8,1\ cm^3$$

Calcolo il peso della piramide :

$P = P_s \times V = 0,48 \times 8,1 = 3,888\ g = 4\ g$
$1\ kg = 1000\ g$

Calcolo il numero dei cioccolatini:

$1000 : 4 = 250$

Risposta

Si possono produrre 250 cioccolatini

Figure B.3: Grading task in Italian

Scrivi in un testo di una quindicina di righe un episodio della tua infanzia che ti sembra avere un significato particolarmente importante e spiega il motivo della tua scelta. Il destinatario è un adulto con cui hai rapporti familiari.

Risposta 4:
Nome:
Classe:

Questo episodio della mia infanzia credo sia importante in quanto quando è avvenuto aveva come unico scopo il divertimento, ma credo che in realtà abbia trovato il modo di contribuire ai comportamenti che assumo crescendo, insegnandomi alcune cose che solo ora saprei di aver imparato quel giorno.
Era un weekend estivo ed io e la mia famiglia ci eravamo incontrati con il nostro solito gruppo per goderci la giornata soleggiata. Eravamo sei bambine, di tre diverse fascie di età, io e le mie sorelle e le nostre amiche, anche loro tre sorelle, come noi. Nel pomeriggio ci stavamo annoiando e non sapevamo cosa fare. Eravamo circondate da un bosco conosciuto dalla nascita e così ci venne un'idea; avremmo usato il pomeriggio per un'escursione. Entusiaste della pensata, ci preparammo, e decidemmo di legarci in vita una funicella, per rendere l'avventura più realistica. Fatto ciò, ci incamminammo lungo il sentiero, che presto però abbandonammo, camminando tra gli alberi in fila indiana, una dopo l'altra. Una tra le cose bella fu che in alcuni pezzi ci aiutammo a vicenda in base a quello che riuscivamo a fare, chi più, chi meno. Di per sé non fu molto faticoso, ma si sa, i bambini tendono ad accrescere tutte le emozioni.

Figure B.4: Grading task in English

Write a short text of about 100-150 words that describes one or more past days using past simple, affirmative or negative form, regular and irregular verbs.

Risposta 1
Nome:
Classe:

My classmates and I went to a chocolate factory last year. It was a 2-hour ride, so we all fell asleep on the bus. In the factory, we made chocolate. First, we poured the coconut milk in a bowl. The coconut milk was without taste, so we chose the flavour we liked. For example, I liked strawberry, so I poured strawberry milk into the bowl. After that, we put the mixed milk into a special freezer, which can freeze the milk into chocolate in three minutes. Magic! Finally, we used the models to make different shapes of the chocolate. Luckily, we could eat the scrumps. It was so much fun. I can't wait to go there again!

C Bayesian estimate of bias in grading

To avoid estimation error arising from sample variation, we calculated empirical Bayes estimates of teacher bias.[32] This method has been suggested by Kane and Staiger (2002) and is followed by several studies to estimate teacher value added (Chetty et al., 2014; Kane and Staiger, 2008) and teacher bias (Terrier, 2020). We follow the method of Terrier (2020) to make sure that less reliable estimates are shrunk to the mean:

1. First, we calculate the teachers' bias in grading by subtracting the standardized score in the blind test to the standardized grade assigned by the teacher.

2. Second, for each teacher, we measure the bias toward immigrant students in a regression by regressing a dummy equal to 1 if the student is an immigrant student on the bias in teachers' previous grades for that student. We then save the coefficient and standard error for each teacher.

3. Third, we calculate the mean error variance (MEV) by taking the mean of the squared standard errors (noise) and storing the variance of the observed bias (variance of the regression coefficient).

4. We then obtain the true variance by subtracting from the variance of the observed bias the mean error variance (MEV).

5. The reliability ratio is then calculated by dividing the true variance by the total variance (true variance plus noise).

6. Finally, we obtain the empirical Bayes estimator by multiplying the coefficient of the bias by the reliability ratio.

[32] We restrict the sample to teachers that have at least 3 immigrants students in their classes and overall at least 10 students in our dataset. We lose less than 1% of the observation due to this selection and the results are not substantially changed when we include them in the analysis.

D Additional Results using Gender Specific IAT

Table D.1: Bias in grading and teachers' IAT scores

Panel A– Outcome: Teacher Grade									
	All			High Ability			Low Ability		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Immigrant	-0.097***	-0.072***	0.497	-0.179***	-0.127***	1.559**	-0.056***	-0.049**	0.568
	(0.012)	(0.022)	(0.479)	(0.020)	(0.036)	(0.773)	(0.013)	(0.022)	(0.583)
IAT Gender Specific * Immigrant		-0.052	-0.038		-0.113*	-0.098*		-0.014	-0.007
		(0.037)	(0.035)		(0.065)	(0.059)		(0.040)	(0.038)
Obs.	42302	42302	42302	25415	25415	25415	16867	16867	16867
R^2	0.481	0.481	0.509	0.403	0.403	0.442	0.222	0.222	0.258
Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
INVALSI cubic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	No	No	Yes	No	No	Yes	No	No	Yes
Student Controls*Imm	No	No	Yes	No	No	Yes	No	No	Yes
Teacher Controls*Imm	No	No	Yes	No	No	Yes	No	No	Yes

Panel B– Outcome: First Difference, Std Grade–Std Test Score									
	All			High Ability			Low Ability		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Immigrant	-0.079***	-0.058***	0.373	-0.154***	-0.109***	1.333**	-0.048***	-0.042**	0.468
	(0.010)	(0.018)	(0.401)	(0.017)	(0.030)	(0.645)	(0.011)	(0.019)	(0.491)
IAT Gender Specific * Immigrant		-0.045	-0.032		-0.096*	-0.083*		-0.012	-0.006
		(0.032)	(0.030)		(0.055)	(0.050)		(0.034)	(0.032)
Obs.	42302	42302	42302	25415	25415	25415	16867	16867	16867
R^2	0.357	0.357	0.391	0.213	0.213	0.264	0.447	0.447	0.473
Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
INVALSI cubic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	No	No	Yes	No	No	Yes	No	No	Yes
Student Controls*Imm	No	No	Yes	No	No	Yes	No	No	Yes
Teacher Controls*Imm	No	No	Yes	No	No	Yes	No	No	Yes

Notes: This table reports OLS estimates, where the dependent variable is the teacher-assigned grade in Panel A and the standardized difference between teacher-assigned grades and test scores (INVALSI) in Panel B. The unit of observation is student i taught by teacher t in school s. "IAT Gender Specific" is a continuous variable indicating the standard d-score of the IAT test, using the Native-Immigrant IAT with female names for female students and the Native-Immigrant IAT with male names for male students (more details available on Appendix B.1). Student controls include gender, generation of immigration, mother education, and province. Teacher controls include gender, place of birth, age, and age squared. Standard errors are clustered at the teacher level.

Table D.2: The impact of revealing stereotypes in the field and online experiment, by teacher IAT score

Dep. Var: Teacher-Assigned Grade		
	Field Experiment	
	(1)	(2)
Feedback × Immigrant	0.367***	0.242*
	(0.096)	(0.141)
Immigrant	0.294	0.486
	(0.940)	(0.941)
Feedback	-0.153*	-0.148
	(0.079)	(0.101)
Feedback × Immigrant × IAT Gender Specific		0.268
		(0.212)
IAT Gender Specific		0.008
		(0.085)
Immigrant × IAT Gender Specific		-0.086
		(0.162)
Feedback × IAT Gender Specific		-0.011
		(0.114)
Constant	6.997***	7.019***
	(0.759)	(0.750)
Control Mean	6.944	6.944
Obs.	10279	10230
R^2	0.131	0.132
Student Controls	Yes	Yes
Teacher Controls	Yes	Yes

Notes: This table reports OLS estimates, where the dependent variable is the grade assigned by teachers in the field experiment. The unit of observation is student i by teacher t. Standard errors are robust and clustered at the school level (the unit of randomization). "IAT Feedback" is a dummy variable indicating whether the teacher was eligible for receiving the IAT feedback versus the active control message. "IAT Gender Specific" is a continuous variable indicating the standard d-score of the IAT test, using the Native-Immigrant IAT with female names for female students and the Native-Immigrant IAT with male names for male students (more details available on Appendix B.1). Student controls include gender, generation of immigration, and education of the mother, all interacted with whether the student is an immigrant. Teacher controls include gender, place of birth, age, and age squared, interacted with whether the student is an immigrant.