The Long-Term Effects of Teachers' Gender Bias *

Joan J. Martínez

November 2022

Click here for the most updated version

Abstract

This paper studies the effects of teachers' gender biases on students' long-term outcomes, including high school completion, college attendance, and formal sector employment. I measure teacher bias using differences in gender gaps between teacher-assigned and blindly-graded tests, and validate the assessment-based measure with novel data on teachers' attitudes, as captured by the Implicit Association Test (IAT). I develop a large-scale online portal available to teachers and students in Peruvian public schools to collect IAT scores nationwide. This analysis provides evidence that math teachers who strongly associate males with scientific disciplines give higher scores to male students, when compared to blindly-graded test scores, while language arts teachers who strongly associate females with humanities-based disciplines award higher grades to female students. Next, using graduation, college enrollment, and matched employer-employee data on 1.7 million public high school students who were expected to graduate between 2015 and 2019, I find that female students who are assigned to more biased teachers are less likely to complete high school and apply to college than male students. Moreover, female students assigned to more biased teachers in high school are less likely to hold a job in the formal sector after graduation and have fewer paid working hours relative to their male classmates. Exposure to gender-biased teachers also leads to monthly earnings losses for women, further exacerbating the gender pay gap.

JEL Classification: J16, J24, I24, J71

*I am indebted to Christopher Walters and Patrick Kline for their extensive support and guidance. I am thankful for comments from Sydnee Caldwell, Christopher Campos, David Card, Michela Carlana, Supreet Kaur, Lukas Leucht, Edward Miguel, Enricco Moretti, Amanda Pallais, Jesse Rothstein, and Damian Vergara. I also thank seminar participants at the UC Berkeley Labor Seminar and Development Seminar. I'm thankful for the institutional support and information access from the Ministry of Education and the Ministry of Labor of Peru. I thank Ayrton Dextre, Daniela Rossini, and Jose Miguel Valverde for assistance in the data collection. I gratefully acknowledge financial support from the UC Berkeley Economics Department, the UC Berkeley Opportunity Lab, the Weiss Fund for Research in Development Economics, the Institute for Research on Labor Employment Dissertation Fellowship, and the Center for Effective Global Action through its Psychology and Economics of Poverty Initiative and Gender-Related Research in Economics initiative. IRB approval was received from University of California, Berkeley. Corresponding author: martinezp_jj@berkeley.edu

1 Introduction

Gender disparities in labor market outcomes result from high-stakes decisions during high school. Although gender wage gaps have declined worldwide in recent decades, this pattern has been uneven between developed economies (on the one side) and middle- and low-income countries (on the other), where the gaps remain large. Moreover, labor-supply outcomes continue to differ across genders globally, as unemployment and part-time jobs are more common among female workers.[1] Recent evidence suggests that salary requests (Rousille, 2021), self-promotion (Exley and Kessler, 2022), job-attribute preferences (Le Barbanchon et al., 2020, Biasi and Sarsons, 2020, Fleche et al., 2018, Flory et al., 2015), and other behaviors that differ across genders partially drive these gaps. Such behaviors may be reinforced by preferences and attitudes acquired during school, prior to entering the labor market.

Teachers are an important potential source of gender differences in attitudes and human capital. A broad literature indicates that educators play a pivotal role in shaping students' adult outcomes (Chetty et al., 2014a,b, Rothstein, 2010, 2017). Recent studies of teacher gender bias suggest that biased educators discourage girls from pursuing science-oriented high school tracks (Alesina et al., 2018, Carlana, 2019), increase gender-score gaps (Lavy, 2008, Lavy and Sand, 2018), and influence career-choice decisions in the short run and medium run (Lavy and Megalokonomou, 2019). Two bias measures have been at the center of debate concerning their capacity to measure unobserved preferences for female or male students: assessment gaps and the Implicit Association Test (Greenwald et al., 2003, 1998, Nosek et al., 2007, 2014). The longer-run effects of gender bias have been more difficult to assess precisely because they occur long after exposure to biased teachers. As a result, the important question of how outcomes such as graduation, college attendance, and labor market opportunities are affected remains open.

This paper studies the extent to which gender bias among teachers affects students' educational careers, college applications, college attendance, and labor market outcomes. I use novel data from two sources: linked administrative information, and nationwide survey responses from public high school teachers and students in Peru. To construct and validate a measure of teachers' gender bias, I developed a government portal through which I recorded approximately 2,400 teacher responses and 4,600 student responses to Implicit Association Test and gender-attitudes surveys.[2] In addition, using comprehensive administrative records on students' education matched with the students' employment records allows me to evaluate the long-term effects of gender-bias exposure on the trajectories of five cohorts of students from high school to the formal labor market. My identification strategy compares students who are enrolled in the same cohort-grade-year-school cells but assigned to different teachers and, therefore, are exposed to different levels of gender bias.

My analysis begins with the construction of a gender-bias measure based on discrepan-

[1] According to reports by OECD (2022), ILO (2018), and ILO (2019), the gender wage gap in OECD countries today is 11.7 percent, while in middle- and low-income countries, it remains higher at 15–27 percent. ILO (2020) and World Economic Forum (2018) report labor-supply stylized facts.

[2] The educational platform was designed for this project with the approval of the Ministry of Education as part of a multiyear research partnership. The platform, Opportunities for Everyone (Oportunidades para Todos in Spanish), is available at http://www.oportunidadesparatodos.pe.

cies in gender gaps between teacher-assigned and blindly graded assessments (Botelho et al., 2015, Burgess and Greaves, 2013, Carrell et al., 2010, Lavy, 2008, Lavy and Sand, 2018). This measure can be seen as a teacher-specific difference-in-differences estimate comparing boys versus girls for teacher-assigned versus centrally assigned scores. I study the distribution of this assessment-based gender-bias parameter using empirical Bayes methods that account for sampling error in teacher-specific estimates (Aaronson et al., 2007, Chetty et al., 2014a,b). I find substantial heterogeneity in gender bias across subjects and teachers. Particularly among mathematics teachers, shifting one standard deviation upward in the distribution of teacher-level bias increases the gender gap between boys and girls by 0.06–0.09 standard deviations.

This is the first study to validate assessment-based measures of teacher bias by comparing them to newly collected data from the Implicit Association Test (IATs). The IAT is entirely unrelated to student assessments and is, therefore, less threatened by any nonrandom sorting of students to classrooms. The test has a dual interpretation: higher values indicate stronger associations between boys and science disciplines and, at the same time, between girls and humanities disciplines (Greenwald et al., 1998, Nosek et al., 2007). I document that the average Implicit Association Test score among mathematics teachers is 0.301, indicating strong boys-science and girls-humanities associations.[3] These estimates suggest that students in developing economies are likely to confront stronger stereotypes (on the part of teachers) regarding their adequacy at solving math problems. In contrast, language arts teachers' Implicit Association Test scores are, on average, 0.281, which is similar to Carlana's 2019 findings. Most importantly, the direction of the relationship between the Implicit Association Test scores and teacher-level grading bias provides compelling evidence that the latter is picking up subjective assessments of teachers. I find that teachers who believe boys are better than girls at math assign them higher math grades on average relative to girls. Similarly, teachers who believe girls are better than boys in the humanities assign them higher grades in language arts relative to boys.

After validating the assessment-based bias measure, I estimate its impact on student long-run outcomes, including high school completion, college application and enrollment, and employment and earnings in the formal sector. My research design compares students assigned to different teachers within cohort-grade-year-school cells, controlling for a rich set of observables, such as lagged test scores and students' and teachers' demographics, classroom-level controls, and school-grade-level controls. Previous studies of teacher value-added indicate that these controls are an effective means of accounting for nonrandom sorting of students to specific teachers (Chetty et al., 2014a). The research design relies on the identifying assumption that students' assignment to teachers is as good as randomly assigned when conditioning on students' demographics and lagged test scores. I focus primarily on the effects of mathematics teachers' gender bias in grading; however, I provide complementary evidence regarding language arts and science teachers.

My findings indicate that assignment to a biased mathematics teacher generates large increases in gender gaps in high school outcomes that persist into the labor market. Teachers play a significant role in students' decisions to drop out. Assigning female students during one

[3] In this setting, then, the average level of implicit stereotypes is much higher than that for Italy's middle school teachers, which is 0.09, according to Carlana (2019).

grade to a high school math teacher who exhibits one standard deviation more severe gender bias against girls, compared to the average teacher, reduces girls' likelihood of ever graduating from high school by 1.3 percentage points (1.6 percent of the group mean) relative to the case in which boys are assigned to this teacher. Furthermore, when female students switch to a classroom where the teacher's grading bias is one standard deviation higher, their likelihood of applying to college relative to boys' likelihood is 0.7 percentage points lower (2.2 percent of the group mean). In line with these results, the persistent documented effects are largely present in a smaller order if students are exposed to more biased language arts teachers. The gender bias of science teachers only affects high school completion and college application; there are no statistically significant consequences for employment outcomes.

The effects are found to be persistent and salient, as a short-lived (one-year) exposure to more biased teachers increases gender the gender gap in formal-sector employment for girls aged 18–23. This research shows that female students who are assigned to a math teacher who is one standard deviation more biased during one academic year are 1.3 percentage points less likely to hold a formal-sector job at ages 18–19 (equivalent to 18 percent of the group mean). Particularly until ages 20–21, the total effect for female students is negative, and it takes young women approximately five years after graduation for the effects to level to nearly zero. In addition, I demonstrate that female students' increased exposure to gender bias causes monthly earnings losses of USD 2.6 at ages 18–19. The magnitude of this effect is large enough that it exacerbates the gender wage gap by around 32 percent, leading female students to suffer cumulative losses equivalent to 0.91 percent of the mean monthly earnings during their first two years in the labor market. Finally, I investigate whether these effects are concentrated at the bottom of the earnings distribution, and I find marked and unequal losses for more disadvantaged females that last up to ages 18–19.

This analysis relates to the literature on evaluators' biases and stereotypes affecting human capital investment decisions and performance.[4] In this body of work, started by Rosenthal and Jacobson (1968), mainly two empirical measures have been used to detect evaluators' gender stereotypes: assessment gaps constructed with observational information (Lavy, 2008, Lavy and Sand, 2018, Lavy and Megalokonomou, 2019) and the Implicit Association Test collected through surveys or laboratory experiments (Greenwald et al., 2003, 2005, 1998, Nosek et al., 2007, 2014). My focus on linking these two measures offers critical advantages compared to previous work. First, score-based measures are susceptible to capturing unobserved heterogeneity that can be reflected in the analyzed outcomes even after adjusting for observable drivers of assessment gaps across genders. I provide direct estimates of the strong relation between observable grading that is unexpectedly favorable to boys versus girls compared to non-teacher-graded scores and the implicit gender stereotypes favoring boys following science-based fields over girls. Second, extensive evidence indicates that teachers' grading based on classroom performance is influenced by elements other than cognitive ability or students' competencies which threatens assessment-based bias measures (for example, Bertrand and Pan (2013), Figlio et al. (2019), Fortin et al. (2015), Jackson (2018)). My empirical approach for measuring teachers'

[4] I refer to *gender biases* and *gender stereotypes* interchangeably in this article although, from cognitive psychology perspective, these are not equivalent. A formal definition of stereotypes following the probability judgements approach proposed by Kahneman and Tversky (1973) is given by Bordalo et al. (2016).

gender bias partials out alternative sources of scores discrepancies related to classroom behavior and past performance. This leaves us with assessment gaps not explained by observable but confirmed to be related to teachers' implicit gender attitudes.

My results also add to this literature by measuring the effects of teacher biases on long-lasting human capital decisions and outcomes that have been previously unavailable. A growing literature focused on teacher-student interactions has documented that teachers' implicit biases and stereotypes based on gender and race disproportionately and negatively affect female students' achievement (Alan et al., 2018, Alesina et al., 2018, Burgess and Greaves, 2013, Carlana, 2019, Lavy, 2008, Lavy and Sand, 2018), choices between science- and humanities-oriented high school tracks (Carlana, 2019), or college attendance and major choice (Carrell et al., 2010, Lavy and Megalokonomou, 2019, Reuben et al., 2014). In working settings, the influence of evaluators and employers stereotypes of workers' performance has deemed to be relevant in determining productivity (Glover et al., 2017) and job assignments (Ewens and Townsend, 2020, Goldin and Rouse, 2000, Sarsons, H., 2017).[5] I contribute to this literature by providing direct evidence that stereotypes about groups not only impair performance, but also skill investments.

This article also relates to long-standing literature documenting gender gaps in labor markets and discussing its sources. An earlier group of studies focused on the effects of childbearing (Becker et al., 2019, Berniell et al., 2021, Correll et al., 2007, Hardy and Kagy, 2018, Kleven et al., 2019), occupational sorting (Francesconi and Parey, 2018) as central causes for earnings and hiring gaps affecting female workers. Nevertheless, evidence suggests that the wage gap that remains unexplained by workers' competencies or qualifications is still persistent (for instance, see Bertrand et al. (2010, 2005), Blau and Kahn (2017), Goldin (2014)). Recently in this literature, there has been a turn to evaluate behavioral sources of the persistence of gender gaps. Among these, the existence of differential preferences for job attributes such as schedule flexibility and stability (Fleche et al., 2018, Wiswall and Zafar, 2018), commuting time to the workplace (Le Barbanchon et al., 2020) or competitiveness preferences (Flory et al., 2015) have proved to be important causes. Female workers' behavioral traits such as the disposition to bargain (Biasi and Sarsons, 2020), to ask for higher wages (Rousille, 2021), confidence (Risse et al., 2018), or the subjective description of performance in tasks (Exley and Kessler, 2022), have opened new routes to explain the remaining unequal access to wages and hiring opportunities. I contribute to this literature by showing that the employment and wages of female workers respond to gender biases from exposure to them as pupils. Although I do not find a statistically significant effect on the internationalization of bias from teachers to students, this mechanism is consistent with the type of behavioral traits conveyed by educators' bias. Notably, teachers' biases are a previously undocumented source of gender disparities, allowing understanding that the sources of labor market disparities by gender were, to some extent, initiated before workers' initiation into labor markets.

Last, the paper builds on the literature on discrimination in labor markets and educational settings.[6] In particular, the methods I use relate to the most recent strand of this body of

[5] Glover et al. (2017) finds that minority workers who are supervised by managers with higher implicit stereotypes against them, as measured by the IAT, have lower performance metrics on the job they perform. Sarsons, H. (2017) documents an asymmetric penalization of surgeons' past performance measured by patient death that depends on physicians gender when determining subsequent job assignments (that is, surgical referrals).

[6] This paper draws on the framework of discriminatory behavior in which workers' (students') true productivity

work. Methods to retrieve precise individual-level estimates of discriminatory practices and the distribution of those estimates have a recent precedent in the economic literature (Kline and Walters, 2021, Kline et al., 2021). Adding to this line of research, I rely on an Empirical Bayes approach used in the statistics literature (Efron, 2012, 2016, Efron and Morris, 1972) to measure heterogeneity in discriminatory behaviors in the educational setting.

The remainder of this paper proceeds as follows. Section 2 provides a framework that conceptualizes teachers' gender bias and mechanisms, and it introduces a set of economic and psychological studies that motivate the empirical exercises. Section 3 introduces the empirical setting and describes the data. Section 4 presents the estimation strategy for measuring teachers' grading bias, characterizes the distribution of gender bias in assessment among high school teachers, and provides validation using IAT scores. Section 5 describes the methodology for measuring the long-term consequences of teachers' gender bias. Section 6 investigates the effects on academic progress and college outcomes, and Section 7 describes the effects on labor outcomes. Section 8 analyzes the consequences of teachers' biases in subjects different from mathematics. Section 9 concludes.

2 A conceptual framework for human capital investment with biased assessment

The framework in this section illustrates the theoretical mechanisms by which gender-prejudiced attitudes, stereotypes, or biases can affect human capital investment decisions. This discussion theorizes gender bias in assessment and its persistent influence on students' outcomes; therefore, it motivates the empirical analysis of the following sections. I draw from a growing body of studies in economics and educational and social psychology that assess how gender or racial biases and discriminatory behaviors affect individual decisions and how they have tangible ramifications.

The most explored among mechanisms of teachers' gender bias is the direct effect on student achievement, which is the foundation for subsequent student learning. An early literature on students' achievement gaps in mathematics is a crucial starting point.[7] Recent investigations show how gender biases affect such achievement gaps (Alesina et al., 2018, Carlana, 2019, Lavy, 2008, Lavy and Sand, 2018). In particular, Carlana's (2019) study highlights the effects of teachers' implicit stereotypes, which potentially extend to postsecondary fields of study for female students, as they prevent the students from choosing science-oriented tracks before high school. I turn to the literature that assesses the effect of racial or ethnic biases on students' performance. For instance, van den Bergh et al. (2010) find that teachers' ethnic stereotypes as measured by the IAT and self-reported measures affect the academic achievement of minority groups. Similarly, using measures of racial bias in assessment, Botelho et al. (2015) document a

(ability) is unobserved, and employers (teachers) rely on group membership and prior beliefs fueled by prejudicial beliefs. Particularly relevant is the work of Coate and Loury (1993), which proposes a setting where the employers' beliefs prior to observing their evaluations will span a prior-enforcing equilibrium where more optimistic priors on one group against the other will cause the favored group to set lower self-assessment rules for acquiring qualification.

[7] For instance, literature reviews on the gender gap in achievements can be found in Bertrand (2011), Fryer and Levitt (2010), Hyde and Mertz (2009).

grading gap between minority and nonminority students that is not accounted for by observable proficiency or students' behavior.

The cornerstone of the educational research investigating the mechanisms of teachers' biases is the associated change in expectations toward the affected student groups. The relevance of teachers' expectations has been established by a longstanding body of work in education (Babad et al., 1982, Rosenthal and Jacobson, 1968), and the work continues in recent studies that focus on the asymmetric effects of positive versus negative expectations, the accumulation of expectations of multiple teachers over time, and related questions (see a recent review in Jussim and Harber (2005)).[8] Some of this work analyzes the influence of teachers' expectations that are based on group membership (for instance, ethnicity; see Rubie-Davies et al. (2006), van den Bergh et al. (2010)) or expectations that are based not on students' individual characteristics but characteristics of aggregates such as the class (Rubie-Davies, 2010). Teachers' group-specific expectations about students' performance take shape through distinct standards based on group membership. Such standards resonate with work in economics that either theorizes (Coate and Loury, 1993) or empirically documents (Burgess and Greaves, 2013) that teachers' biased expectations based on gender or ethnicity may become self-fulfilling.

If not through expectations, through which alternative channels can teachers' biases impair students' academic progress and vitiate adult outcomes? Teachers' disproportionate allocation of educational resources has been found to affect students' academic achievement (Alan et al., 2018, Carrell et al., 2010). Recent work by Alan et al. (2021) sheds light on how teachers' ethnicity-based biases affect the distribution of seats in the classroom, leaving some students to receive less instructor attention or benefit from fewer connections with classmates. Next among teachers'-bias channels is the association of biases with measures of educators' quality as shown in the literature on value-added (Chetty et al., 2014a, Jacob and Lefgren, 2008, Kane and Staiger, 2008, Rothstein, 2010) and as explored in the context of teachers' biases by Lavy and Megalokonomou (2019).

Another set of mechanisms investigated is behavioral and psychological traits. According to studies on educational psychology, teachers' stereotypes have been found to affect students' self-assessment or the development of academic self-concept (Ertl et al., 2017). Related to this finding is Exley and Kessler's (2022) study, which finds that female workers subjectively assess their performance in so-called male-typed math and science tasks as unsatisfactory relative to the average self-assessment of male workers. Second, there is mounting evidence on the match effects of students' assignment to same-race (Dee, 2005, Gershenson et al., 2018) or same-gender teachers (Carrell et al., 2010) on achievement, attendance, and high school completion, among other short- and medium-run outcomes. It is reasonable to expect that teachers' biases are potentially lower across matched teachers and students, so this literature supports the notion that stereotypes play a role in students' outcomes. Last, recent work in social psychology investigates *internalized biases*, a psychological construct by which minorities or members of groups that are subject to negative stereotypes, beliefs, messages, and other forms of discrimination accept the negative stereotypes and beliefs about themselves and their group and, as a consequence, exhibit this behavior (David et al., 2019, Jones, C. P., 2000).[9] I follow a growing social-

[8] Other relevant reviews on the literature are in Harris and Rosenthal (1985), Hoge and Coladarci (1989).

[9] The work of Jones, C. P. (2000) started this literature. It develops a conceptual framework that postulates

and clinical-psychology literature that uncovers to what extent racial discrimination results in psychological distress (see Molina and James (2016), Roberts and Rizzo (2021), Willis et al. (2021)). In this study, I hypothesize that educators can internalize gender biases (for instance, female teachers internalize gender bias against them and act on it), and so can students. As a result, teachers that directly exhibit biases or have internalized them can perpetuate the pervasive effects of stereotypes in their students, which leads them to internalize the stereotypes. In other words, students may be led to change their self-perception of their educational adequacy or of their ability to complete gendered learning tasks. This affects subsequent decisions to drop out of high school and manifests itself in other adulthood outcomes.

Gathering the main channels introduced in this discussion, I formalize how teachers' gender biases shape students' decision to acquire qualifications (for example, a high school diploma or a college degree) in Online Appendix B. I develop a dynamic model of human capital investment as a discrete choice of whether to acquire a qualification in an environment of incomplete information about the student's academic potential—both from the student's standpoint and that of their teachers.[10] In the model, students from two groups that are fully distinguishable to teachers (for example, boys and girls) learn about their academic potential from their teachers' and a central planner's assessments. I allow teachers' assessments to manifest their stereotypes about the distribution of students' abilities in these groups. I assume that students are rational decision-makers who maximize their utility derived from becoming qualified net of the costs of acquiring qualifications. Although qualification costs vary across students, they are identically determined in both groups. The central result from the model is the implications of teachers' bias for the self-assessment of their students in equilibrium. In particular, I derive conditions for an equilibrium of self-fulfilling negative stereotypes against either of the two groups, which demonstrates that students against whom teachers hold negative gender biases have suboptimal self-assessments, which, over time, lead them to underinvest in their academic trajectories compared to students from the group regarding which teachers have a positive view of their abilities. In the following sections, I obtain empirical estimates of teachers' gender-based bias in assessment and long-term consequences that allow me to empirically assess the prevalence of a self-fulfilling equilibrium as predicted in Online Appendix B.

3 Background and data

3.1 Public school system in Peru and gender equality

The Peruvian public school system provides free education to 5.9 million students, representing 74.5 percent of the total student population in 2019. The public school sector is the largest employer of teachers in the country, employing 69.7 percent of teachers full-time or part-time. The other main type of education provider is private schools, which do not operate within a voucher or charter system. Schooling is compulsory between ages 5 and 16, and it is divided

that racial discrimination works on three levels: institutionalized (that is, unequal access to public services and opportunities due to race), personally mediated (for example, prejudices, stereotypes, racial discrimination), and internalized racism, defined as the "acceptance by members of the stigmatized races of negative messages about their own abilities and intrinsic worth."

[10] The model builds on Coate and Loury's (1993) model of self-reinforcing prior beliefs.

into three levels: preschool (ages 3 to 5), primary (ages 6 to 11), and secondary (ages 12 to 16) (UNESCO, 2016). I focus on the secondary level, in which 1.9 million students were enrolled from grades 7 to 11 in 2019.

Public school enrollment is available to all school-age children; however, those who enroll tend to come from underprivileged backgrounds. For instance, the 2018 National Census Evaluation questionnaire of eighth-grade students indicates that 43 and 27 percent of parents answered that their highest level of education was less than high school completion and high school completion, respectively (see Appendix Figure C.1). Moreover, the questionnaire indicates that the typical eighth-grade public school student in 72 percent of districts nationwide has very low socioeconomic status (see Appendix Figure C.2).

Even though access to education increased steadily between 2010 and 2019, with attendance increasing at an average annual rate of 1.1 percent, school completion still represents a challenge, especially among female students outside of metropolitan areas in the east or northwest (see Appendix Figure C.3). Recent reforms have aimed to increase the quality of education provision and provide equal opportunity for female and male students.[11] In 2016 the Ministry of Education introduced a new national curriculum that takes an inclusive approach to diversity, with a guiding principle of gender equity and inclusion (Minedu, 2016). However, it has not materialized in high school or other pedagogical content. Implementation is widely left to the discretion of regional and local units equivalent to school districts.

Gender-equality attitudes in Peru regarding, for instance, access to education or employment have not progressed to match those seen in developed countries.[12] As I show in the following sections, this trend is pervasive in the education setting.

3.2 Data and samples

In this section, I describe the two main samples I use for the empirical analysis and I describe their data sources. The first sample is used to estimate a measure of gender-biased assessment for each teacher. The second sample contains student information, which I use for evaluating the effects of gender-grading bias on a set of adulthood outcomes.

3.2.1 Student sample: Estimating teachers' assessment bias

The measure of gender bias in assessment against girls measures which teachers are more likely than others to assign examination grades that, on average, disfavor female students relative to male students, with blindly graded test scores as the benchmark. It is a double difference, with differences across two types of examination and across groups of students defined by gender.

[11] In addition to gender, geographic location and race are factors contributing to unequal access to education services. Among the reforms to improve achievement gaps and overall education quality, there have been reforms to introduce an evaluation-based selection process for teachers (Bertoni et al., 2022), establishing payment and incentive schemes for staffing schools in rural areas (Bobba et al., 2021), establishing auditing information systems preventing and reporting bullying (Minedu, 2022), and improving school equipment and infrastructure (Paxson, 2002).

[12] The WVS-7 results for Peru suggest a prominent level of gender-inequality attitudes regarding equal access to education for women. 12.9 percent of respondents agree with the statement *'University is more important for a boy than for a girl'* in Peru, compared to 9.8 percent for the US. The score for gender equality in education is on a 0-to-1 scale, where higher values indicate more favorable attitudes toward equality.

Therefore, to retrieve teacher-level measures, I construct a sample of students with end-of-year classroom scores and standardized test scores matched with their respective teachers. In this setting, high school teachers tend to specialize in a single subject but teach different grades in different school years. I exploit this turnover to calculate gender bias with scores of eight grade classes from 2015 to 2019. I then assign these measures to other classrooms that the teacher instructed in the same subject and time period (see Online Appendix A for the detailed procedure). An important assumption is that the gender bias exhibited by a teacher teaching eighth-graders is not different from the teacher's bias when they teach other grades.

The teacher in charge of instructing a subject is responsible for assigning a consolidated mark at the end of the school year. The scores assigned by teachers do not reflect students' performance on a single-format standardized test. Instead, they reflect the students' proficiency at the end of the school year in a set of competencies for a set of learning goals, evaluated in multiple examinations with a standard rubric set by the National Curricula. In contrast, standardized scores correspond to students' performance on a single evaluation. A nationwide standardized test is administered annually by the Ministry of Education to eighth-grade students in mathematics, language arts, and science in three exams at the end of the school year.[13] The format and content of the standardized and classroom evaluations are described in Online Appendix F. Table 1 shows descriptive statistics of the sample of eighth-grade pupils whose scores are used to compute a measure of teacher-level assessment bias. Students are, on average, 13 years old. More than half of the sampled students speak Spanish as their primary language, 20 percent were born in Lima, and the schools that 86 percent of them attend are in urban areas.[14] A large number of these students come from low-income families, and around 24 percent receive a conditional cash transfer.

3.2.2 Student sample: Long-term outcomes

Data sources. This analysis uses several data sources to estimate the effects of teachers' gender bias on a sample of 1.7 million public high school students graduating between 2015 and 2019. The data set from which I draw that sample tracks students' exposure to gender-biased mathematics, language arts, and science teachers between grades 7 and 11. It tracks their high school–completion, college-attendance, and labor market outcomes between ages 18 to 23, depending on the graduation cohort. Panel A of Appendix Table C.1 illustrates the grades for which I can observe gender bias and the period over which I can measure outcomes for each student cohort.

The main data set I use for constructing this sample is the universe of public school students' administrative records, including students' classroom assignments, evaluation score records, enrollment records, demographics, high school–completion records, and college records provided by the Ministry of Education. I also use employer and employee information provided by

[13]The National Standardized Evaluations of high school students administered in 2015, 2016, 2018, and 2019 were scheduled in November, and the school year finished in the second week of December at the latest. I do not use the information from 2017, as the standardized evaluation was canceled for that year because of a natural disaster in the country.

[14]The primary-language statistics for this sample of eighth-grade students differ from the primary-sample descriptive statistics in Table 2 because of missing values in the administrative data set regarding the takers of national standardized tests.

the Ministry of Labor to identify whether students get jobs in the formal labor market after graduating and to get measures of earnings and working hours. In addition, I use comprehensive information on teachers' classroom assignments, which includes their demographic information, course subject, and performance score, from a national teacher evaluation for hiring public school instructors in 2015, 2017, 2018, and 2019.[15]

Sample construction. High school students are assigned to one homeroom teacher per classroom and have different subject-specific instructors for the entirety of the school year. Therefore, students' and teachers' classroom-assignment records are linked at the classroom-subject level for each grade using classroom identifiers. The purpose is to identify the assessment bias the students were exposed to in each subject in each grade. The successful student-teacher matches constitute my base sample. Next, I join with the information on estimated teachers' gender bias using teacher identifiers.[16] Finally, after applying restrictions to preserve valid matches in the sample, I link this data set with the students' long-run outcomes using unique student-level identifiers. This constitutes the full sample that I use for my analysis below (see Online Appendix A for details).

Outcomes. The primary outcome variable is high school completion, which I measure over different time frames that allow me to keep track of the academic progress of students. Similarly to Gray-Lobe et al. (2021), I construct a projected graduation year, defined as the year the student is expected to graduate in a normal academic progression after five years of high school. To calculate this variable, I take as a reference year the school year when the student is enrolled in grade 7, and I compute the year in which the student will graduate if they advance to grade 11 without retention. For instance, a student in cohort I who is in grade 7 in 2015 will have 2019 as their projected graduation year. Using the *projected* variable, I define *on-time* high school graduation to mean that students finish in the projected graduation year and *ever* graduating from high school to mean students finish at some point, whether on time or late.[17]

The second set of outcomes consists of college application, admission, and enrollment. As before, these are defined using the on-time and ever concepts, which allows me to evaluate the impact of the gender biases of teachers not only on students' college attendance but also on whether they deviate from the standard progression. The final set of outcomes captures the experience of public high school graduates venturing into the labor market with a high school diploma. In particular, the outcomes measure whether the student is employed in the formal

[15]The National Teachers Evaluation was established in 2014 to assess the competencies of teachers aiming to attain permanent positions and access a regulated payment scale and additional supplementary wage bonifications, retirement plans, and promotion schemes.

[16]Section 3.2.1 describes the sample used to construct the bias-in-assessment measure, and Section 4 describes the empirical strategy for using teacher-assigned scores for constructing a measure of gender bias in assessment.

[17]Dropout status can change, as students may re-enter school an indefinite number of years after leaving it. My definition of high school completion considers this re-entry, but it does so over different time horizons for each cohort because data availability is limited. For example, for students who drop out in 2015, I can see whether they return during 2016–19, while if someone dropped out in 2018, I can only see whether they return during 2019. In addition, as seen in Appendix Table C.1, the data structure is an imbalanced panel so I observe each cohort of students in a different number of grades over high school. For instance, I observe cohort I only over grades 10 and 11, while cohort V is observed over grades 7 to 11. Nevertheless, the starting school year for calculating the *projected* graduation year will not be the seventh grade for all cohorts for which I use the student's enrollment records.

sector and what their annualized earnings and annualized work hours are. The time frame for tracking the evolution of these outcomes is between ages 18 to 23 after graduation. Given that some students who do not manage to graduate from high school still participate in the labor market, I consider the projected high school graduation year as the common threshold for all students.

I estimate annual employment in the formal sector, earnings, and work hours using employment records that report these outcomes monthly—a higher frequency than the quarterly or annual measures available in many administrative data sets (for instance, see Card et al. (2013), Sorkin (2018)). The granular data allows me to determine whether a worker is employed in the formal sector despite the precarious nature of work in emerging economies and the high crossover between the formal and informal sectors (OECD, 2015). I consider a worker to be employed in the formal sector if they hold a job contract with their primary employer with positive earnings in the formal sector for at least one month with their primary employer between 2015 and 2020.[18] A worker's *primary employer* is defined as the employer that pays the largest share of the worker's earnings in a quarter (Abowd et al., 2003, Lachowska et al., 2020, Sorkin, 2018). This definition is convenient for considering the scenario of multiple employers for a single worker and for retrieving annualized earnings and work hours. Online Appendix A explains the construction of the labor-outcome variables in detail.

Descriptive statistics. I study outcomes of five cohorts of students at two points in time: during their high school years and after graduation. Table 2 presents summary statistics of demographic and education characteristics during high school for the (stacked) base and full samples of students in grades 8 to 11. Information is available for these student cohorts starting in seventh grade, but the empirical strategy requires lagged test scores, so the analysis sample starts in eighth grade. The table allows us to compare the characteristics of students in the base sample, comprising successful student-teacher matches, to the full sample of students with an available grading-bias measure. Students included in the full sample are more likely to be indigenous language (Quechua) speakers, to have been retained during high school, and to be born outside Lima. Overall, over 80 percent of public school students speak Spanish, and roughly 2 out of 10 students were born in Lima and attended a school in Lima. On average, in the base and full samples, across grades 8 to 11, female students outperform male students in mathematics, language arts, and science according to standardized classroom test scores.

Once these students leave the schooling system, I observe their workers' records for 2015–20. Each cohort in the sample is observed in the employer-employee records for a different time window based on their projected graduation year. Panel B of Appendix Table C.1 displays the cohort-year cells for which I can observe labor market outcomes. Table 3 shows descriptive statistics for the sample of public school students matched to labor records in Columns (4) and (5) and a comparable benchmark group of workers between 18 and 25 years old from the entire population in Columns (1) to (3). The first panel in the table shows average earnings and hours computed using the monthly information for 2015–20, and Panel B only reports workers' characteristics for the most recent employment relation with an *dominant annual employer* as

[18] As detailed in Online Appendix A.2.1, I also consider the job spells for two and three consecutive months.

defined in Online Appendix Section A.2.1. The full regression sample includes 485,000 workers with available records and a projected graduation year between 2015 and 2019. This group represents roughly 20 percent of the base sample. Female workers in the regression sample have around 15 percent lower wages than male workers, and a similar gap is observed in the population. The rate of workers reporting that they finished high school is similar across gender groups, with a higher rate of men reporting having attained some college or technical education.

3.3 Gender-attitudes survey and the Implicit Association Test

3.3.1 Sampling and survey design

My survey aimed to quantify teachers' gender attitudes and implicit stereotypes as measured by the IAT and to evaluate the relationship between IAT scores and assessment-based bias. Figure 2 shows the data-collection timeline, which started in September 2021 and ended in September 2022. The sampling frame comprised high school teachers appearing in Ministry of Education administrative records that included contact information and school information according to the School Census 2019. During most of the data collection period, school activities were not in person because of the COVID-19 pandemic. Therefore, contact information (phone numbers) needed to be part of the sampling frame, as survey participants had to be contacted via text messages and phone calls. Approximately 230,000 teachers had available school and contact information matching the sampling criteria.[19] I conducted stratified sampling by region and selected 13,654 teachers to receive a text message inviting them to participate on a web-based platform hosting the survey. Each sampled teacher received a text message in two waves, one in October during the 2021–22 school year and another in January during the 2021–22 school year. Figure 1 illustrates the completion rate of teachers and the data coverage nationwide.

The survey was available on a web platform I designed and deployed for the Ministry of Education as part of its remote-learning program targeting high school students. The platform is part of a multi-year effort to study gender gaps and inequality sources affecting young students' trajectories.[20] During the data-collection period, 8,687 teachers registered to participate and 2,462 teachers completed the entire survey. Teachers who visited the website had to give informed consent in order to register and participate in the activities. No reference to gender bias was made in the consent form or on the web platform. The goal communicated to teachers was that the project aimed to gather information on teachers' practices in order to promote equal opportunity for all students at public schools. The survey was divided into two activities: a teachers' questionnaire and an association test.

My final survey sample includes 1,102 mathematics and 950 language arts teachers (N= 2,052) who completed both activities and were successfully matched to the teacher-student matched sample that contains a measure of gender bias in assessment. Teachers from other subjects who completed the survey and the IAT were excluded from the analysis sample because

[19]The number of teachers for which the Ministry of Education had information matching the sampling criteria was taken from personnel records by government officials with clearance to access teachers' contact information.

[20]The platform's name, Opportunities for Everyone (*Oportunidades para Todos* in Spanish), does not make explicit references to gender biases. However, it appeals to teachers and students to report information for the ministry to develop learning tools and policies that will promote learning in diversity and inclusion for all students.

information on bias in assessment was unavailable.

3.3.2 Teachers' questionnaire and Implicit Association Test

The IAT is a psychology test developed to measure implicit stereotypes (Greenwald et al., 2003, 1998, Nosek et al., 2007, 2014). Unlike self-reported gender-related attitudes, IAT responses are less influenced by cognitive processing, as they represent an indirect measure of stereotypes: the speed of categorizing science and gender terms (De Houwer and Moors, 2007, De Houwer et al., 2009). Therefore, the IAT is a more reliable assessment of implicit attitudes and overcomes social desirability bias (Egloff and Schmukle, 2002, Steffens, 2004).[21]

A growing literature finds evidence correlating IAT measures with minority job applications (Rooth, 2010), teachers' high school track recommendations (Carlana, 2019), and minority workers' job performance (Glover et al., 2017). An active area of study in psychology concerns the capacity of IAT-measured stereotypes to predict observed or self-reported behavior (Buttrick et al., 2020, De Houwer and Moors, 2007, De Houwer et al., 2009). Nevertheless, similarly to recent findings in economics, studies in psychology document a relationship between IAT-measured bias and various observed behaviors, such as health care services by practitioners (FitzGerald and Hurst, 2017) and harmful intergroup behaviors (Rudman and Ashmore, 2007).

For this study, I administer two IATs in random order to participants: the gender-science IAT and the gender-career IAT (see details on the IAT in Online Appendix Section D.1). In the former, teachers are asked to associate science and humanities with gender groups, while in the latter, they associate career and family-related terms with gender groups. More specifically, the gender-science IAT measures teachers' strength of association between science and humanities concepts, on the one hand, and group terms referring to girls and boys, on the other hand. The test instructs participants to make associations as fast as they can between words from each group (that is, sciences and humanities) and gender terms (that is, female and male), measuring the association speed in milliseconds. Quicker associations between, for example, humanities and girls or sciences and boys demonstrate the participants' stronger latent connection between these two, reflecting implicit stereotypes that exclude women from science disciplines.

The IAT score is a standardized measure with a possible range between -2 and 2, where higher values indicate a stronger implicit preference for boys in science careers and girls in humanities careers, relative to boys in humanities and girls in science. I follow the improved scoring procedure proposed by Lane et al. (2007) and Greenwald et al. (2003), which discards teachers' overly slow or fast responses to prevent contrived or fatigued random associations from being included in the analysis. Throughout the following analysis, interpretation of the IAT's gender-bias severity is based on the following discretizing IAT categories: *preference for girls*, *little to no bias*, *slight bias*, *moderate to severe bias*, and *strong bias*.[22]

[21] Socially desirable responding is a type of response bias in which participants answer questions aligning to cultural norms that let them be accepted or appraised positively by others. Such responding occurs despite their true latent attitudes (Fisher, 1993, Paulhus, 1991, Podsakoff et al., 2003). It has been documented to affect job-application behavior (Stöber, 2001) and political poll responses about race-related policies (Berinsky, 1999, Krysan, 1998).

[22] The break points for determining the categories are $-.15, .15, .35, .65$ as proposed by Greenwald et al. (2003, 2005).

Additionally, the teachers' questionnaire collects complementary information to that obtained in the IAT, particularly self-reported gender attitudes and information on how teachers exhibit implicit gender biases in the classroom. Online Appendix Section D.2 has further information on the teachers' questionnaire.

3.3.3 Descriptive statistics

Table 4 presents information on the surveyed teachers who have been matched with a measure of grading bias and therefore constitute my analysis sample. Several participant characteristics bear emphasis. Gender biases are pervasive in the sample of surveyed teachers, as 43.4 percent of mathematics teachers have a moderate to severe or strong bias against girls in science, while the same holds for 45.8 percent of language arts teachers. Regarding career stereotypes, 40 and 38.4 percent of mathematics and language arts teachers, respectively, exhibit such levels of bias against girls in professional career trajectories. More mathematics teachers are men in the analysis sample, while the remaining demographic characteristics are roughly the same across mathematics and language arts teachers and across school locations. Mathematics and language arts teachers work around 28.1 hours per week, and the class size is, on average, 25 students across both subjects. Forty percent of teachers have a tenure contract; on average, teachers in public schools have 13 years of experience. Last, consistent with the level, 16–18 percent of teachers have witnessed or experienced discrimination firsthand perpetrated by other colleagues or school principals.

4 Measuring teachers' gender bias

I consider a setting in which teachers are likely to hold gender-specific prior beliefs about their students' ability distribution.[23] Student-teacher matched data allow me to measure the prevalence of teachers' gender bias in their assessment.

4.1 Estimation strategy

I define a measure of gender bias in teacher j's assessment as the difference between male and female students' mean gaps between teacher-assigned scores and blindly graded test scores. Higher values of this measure indicate teachers are, on average, grading male students more favorably than female students, while lower values indicate the opposite. This measure allows me to compare the degree of bias across teachers.

The set $\mathcal{N} = \{1, \ldots, N\}$ denotes a student population indexed by i; teachers indexed by j are part of a population denoted $\mathcal{J} = \{1, \ldots, J\}$. Students are assigned scores by their teachers and a central planner (for example, the school-district authority or the agency in charge of standardized examinations). S_{ij}^{T} denotes the examination assessment of student i assigned by teacher j. In addition, students receive another score, denoted S_{ij}^{N}, as measured by a standardized national examination.

[23] Gender bias's influence on the assessment of competence has been studied, for instance, in Alesina et al. (2018), Botelho et al. (2015), Carlana (2019). Appendix B introduces a framework that formalizes the influence of teachers' gender bias on human capital investments.

I define a system of estimating equations, with parameters of interest $(\alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \beta_{2,j})$:

$$S_{ij}^{N} = \alpha_{1,j} + \alpha_{2,j} M_i + W_{ij}'\alpha_3 + \eta_{ij} \tag{1}$$

$$S_{ij}^{T} = \beta_{1,j} + \beta_{2,j}\, M_i + W_{ij}'\beta_3 + \epsilon_{ij} \tag{2}$$

The specification includes teacher-specific intercepts $(\alpha_{1,j}, \beta_{1,j})$ and gender effects on score assignment $(\alpha_{2,j}, \beta_{2,j})$. Under the identifying assumptions, the difference between parameters $\beta_{2,j}$ and $\alpha_{2,j}$ averaged by teacher measures the gender stereotypes that enter into their competence assessments. Covariates W_{ij} include student characteristics (for example, student age in months and place-of-birth indicators), quadratic polynomials of lagged scores in mathematics and language arts, and quadratic polynomials of lagged physical education scores as proxies for observed classroom behavior as in Botelho et al. (2015). I estimate Equations (1) and (2) separately via a fixed-effects approach using the teacher-student matched sample containing classroom grades and standardized assessments in the respective subjects. I report robust standard errors clustered at the student level. Below, I use these parameters to construct a measure of teacher-level assessment bias.

When measuring grading bias, omitted variable bias is a challenge. Boys could be assessed systematically higher or lower than girls on teacher-graded tests not because teachers are biased but because of unobservables correlated with gender. Some possible factors driving better performance by either group include the possibilities that teachers consistently prioritize content favoring one group or have a teaching style that does so; teacher scoring reflects student discipline in the classroom (Botelho et al., 2015); test-taking skills differ depending on exam format (Ben-Shakhar and Sinai, 1991, DeMars, 1998, 2000, Gallagher and De Lisi, 1994, Klein et al., 1997, Liu and Wilson, 2009); and examination conditions favor one gender over another (Gneezy et al., 2003, Niederle and Vesterlund, 2010). I address this concern by controlling for comprehensive student and teacher characteristics to adjust the female-male score gaps in the teacher-level bias measure.

4.2 Teacher-level grading bias

The teacher-level differences between teacher-assigned and centrally assigned scores attributable to the disclosure of students' gender in nonblind examinations are characterized as follows:

$$\theta_j := \beta_{2,j} - \alpha_{2,j} \tag{3}$$

I interpret θ_j as the adjusted discrepancies in gender gaps from teacher-graded versus centrally graded assessments measured in standard deviation units. When this underlying parameter of interest takes higher values for teacher j_1 relative to teacher j_2, it indicates that j_1 is more likely to assign higher scores to boys than to girls compared to the gender gap on blindly graded tests because of the teacher's stereotypes against girls. Thus, this is a relative measure of bias, allowing for comparisons of gender bias in assessment across teachers.

There are two restrictions that allow me to identify the parameter of interest in Equation (3) and to interpret it as an informative measure of gender bias for comparisons across teachers. First, I assume there are no unobserved teacher-by-gender factors in student skills that

differentially affect the gender gap on teacher assessments compared to blindly graded scores. Additionally, I assume that if there is any skill differential between boys and girls, it does not vary by teacher across their classrooms. These are essentially the assumptions needed to identify parameters in a difference-in-differences research design in which I compare boys versus girls and teacher-assigned versus blindly graded examinations. These assumptions do not rely on a cardinal interpretation of θ_j; that is, the location of θ_j's distribution need not be interpreted as a causal impact of bias. I discuss potential threats to the identification of θ_j in the following section.

4.2.1 Heterogeneity in teacher-level bias

The teacher-level estimates $\hat{\theta}_1, \ldots, \hat{\theta}_J$ and associated standard errors denoted $s_1, \ldots, s_J$ are obtained by estimating Equations (1) and (2). The scores used for estimating $\hat{\theta}_j$ have been standardized by year and subject. Table 5 summarizes the distribution of estimates of teacher-level bias in assessment for mathematics, language arts, and science teachers. Under the identification assumptions, θ_j is a meaningful *relative* measure of bias that I use for computing the effects of one teacher's gender bias relative to another's on students' long-term outcomes. Comparing the exposure effects stemming from different teachers' bias is equivalent to a triple-difference design, as I compare boys versus girls, teacher-assigned versus centrally assigned scores, and teacher j_1 versus teacher j_2. The first row in Table 5 reports the mean gender bias by subject without any adjustment, $J^{-1}\sum_j \hat{\theta}_j = \hat{\mu}$, for teacher j of a given subject.[24] Grading bias varies considerably across teachers. Appendix Figure C.4 displays the distribution of θ_j for all three subjects; it indicates that teachers in these three subjects systematically assign higher scores to *both* girls and boys.

Analyzing the heterogeneity of bias in assessment across teachers poses another challenge. The estimates $\hat{\theta}_j$ are expected to be more variable than the underlying parameters θ_j because of sampling error. As a result, the variance of the estimated distribution of θ_j is most likely biased upward. This sampling variation can have several sources. For instance, for teachers that have limited experience or small classrooms, the estimated $\hat{\theta}_j$ are very imprecise because the number of observations is limited (see, for example, Kane and Staiger (2002)). Therefore, a debiased estimate of the underlying θ_j's variance is $\hat{\sigma}^2_U = J^{-1}\sum_j[(\hat{\theta}_j - \hat{\mu})^2 - s_j^2]$, in which I subtract the estimated standard errors s_j to depurate the observed variance.[25] Additionally, I compute a student-weighted version of the bias-corrected variance to reduce the noise estimated in $\hat{\theta}_j$ that is related to very small class size. That variance is defined as $\hat{\sigma}^2_W = J^{-1}\sum_j \left[w_j(\hat{\theta}_j - \hat{\mu})^2\right] - \sum_j w_j s_j^2$, where student weights are $w_j = \mathcal{N}(j)/\mathcal{N}$ and $\mathcal{N}(j)$ is defined as above for teacher j for a given subject and $\mathcal{N} = \sum_j \mathcal{N}(j)$ for that subject.

[24] If I interpret $\hat{\theta}_j$ as a measure of *absolute* bias instead of a relative one, the negative mean would imply that, on average, teachers are biased against boys. However, this *absolute* interpretation is not credible, because the classroom and standardized examinations could be testing slightly different skills. For instance, boys could be slightly worse on average at those skills because of differences in assessments or unobserved skill gaps across gender, as argued in the following paragraphs. This would shift the mean gender discrepancies and, therefore, shift ϑ. Nevertheless, the *relative* interpretation remains valid as long as that skill differential does not vary by teacher (as required by the second identifying assumption). Moreover, under the triple-difference interpretation of $\hat{\theta}_j$, the location of the ϑs does not play a relevant role in my analysis.

[25] Similar bias-corrected-variance procedures can be found in Aaronson et al. (2007), Abdulkadiroğlu et al. (2020), Angrist et al. (2017), Kline et al. (2021).

The second and third rows of Table 5 indicate that mathematics teachers exhibit the lowest dispersion in assessment bias. The bias-corrected weighted standard deviation of grading bias against girls is between 0.06 and 0.09 test-score standard deviations for mathematics teachers. The dispersion of mathematics teachers' gender bias is also smaller, on average, compared to language arts teachers in Lavy and Sand's (2018) findings.[26] In contrast, in my analysis sample, language arts teachers' dispersion of gender bias against girls is lower than that of their science colleagues: moving upward one standard deviation in the distribution of language arts and science teacher bias increases the gender gap of boys versus girls by 0.11 and 0.16 test-score standard deviations, respectively. The lower grading-bias dispersion in mathematics could be because instructors have limited discretion to grant partial credit on numerical exercises, or it could be that math students can more easily verify whether their answers are correct.

Discussion of identifying assumptions. The main potential threat to identification stems from the fact that teacher-assigned and centrally assigned scores could reflect the assessment of slightly different skills across boys and girls. The possibility of different assessments across examinations is largely rejected based on the analysis contained in Online Appendix F, as I compare the contents evaluated in both types of examinations and I analyze the distribution of contents across questions in the standardized tests and typical classroom examinations in mathematics and language arts.[27] My content analysis shows there are no major differences in the distribution of learning competencies evaluated across the examinations taken by students at the end of eighth grade.

A related threat to identification is that, in some classes, boys (compared to girls) could be more likely to have more of the skills that are required for the teacher-assigned tasks and not the centrally graded tasks. To understand this concern, consider a scenario in which observed and unobserved skill gaps across genders might exist. For instance, cultural norms could motivate boys in some classrooms of a given teacher to allocate their effort to improving their skill at rapid problem solving to earn them higher marks in the types of questions that are common in classroom examinations. In fact, a long-standing literature in educational psychology explores how different test formats present different challenges across genders and racial groups. For instance, Liu and Wilson (2009) investigate the existence of gender differences in performance by type of exam question. The authors conclude that male students have an advantage over female students on complex multiple-choice items and other question types in which the evaluated content departs from textbook context or requires a new use of standard methods learned in class. This aligns with the research of Gallagher and De Lisi (1994), who documents that girls outperform in questions related to textbook content. Moreover, Klein et al. (1997) finds evidence that girls have better outcomes than boys in performance assessments composed of hands-on performance tasks reported in writing. Boys perform better on multiple-choice tests.

[26] Strictly, the English and Hebrew teachers reported in Lavy (2008) are referred to here as language arts teachers.

[27] The comparison of examination contents and formats for teacher-graded and blindly graded examination corresponding to the 2019 school year are contained in Online Appendix F. I chose 2019 as a benchmark year for the comparison because it corresponds to the most recent evaluation year in my analysis period. Appendix Table F.2 contains the comparison of contents in mathematics evaluations, and Appendix Table F.3 displays the analysis of language arts examinations. Moreover, Appendix Table F.4 displays the comparison of contents assessed across different question types.

One of the plausible explanations provided by Ben-Shakhar and Sinai (1991) is that boys tend to guess more on multiple-choice exams and thus have lower omission rates than girls.

Suppose this was the case in our setting and there were unobserved skill gaps between boys and girls that drive performance gaps for each question type. Consequently, girls could be deemed to perform worse on average in these classrooms not because the educator is biased in favor of boys but because the format of the classroom examinations assesses the skills on which they have a relative advantage. Formally, according to Equation (3), this would shift the gender discrepancies in the teacher-assigned scores, reflected in $\beta_{2,j}$, but leave the blindly graded score gaps, $\alpha_{2,j}$, unaffected. I address this concern by examining the distribution of question types per competency evaluated across examinations. As shown in Online Appendix F, I find no substantial gender difference in the distribution of questions nor the distribution of questions per competency assessed in the standardized and classroom examinations.[28] Nevertheless, there might be exam-type-related skill gaps that remain unobserved and can potentially shift θ_j to the disadvantage of boys or girls.[29] Consequently, the negative means reported in the first row should not be interpreted to mean teachers are, on average, biased against boys. Moreover, my main analysis compares θ_j across teachers, so the mean of the distribution does not play a relevant role.

4.2.2 Relating grading bias to teachers' observed characteristics and implicit stereotypes

In this section, I examine the extent to which observed teacher characteristics and (mostly) unobserved implicit gender stereotypes are predictors of estimated gender-bias intensity.

Consider the following regression relating teacher j's estimated grading bias to their observed characteristics:

$$\hat{\theta}_j = \phi_1 + X_j'\phi_2 + u_j \tag{4}$$

Here, $\hat{\theta}_j$ is the estimated assessment bias for teacher j divided by the corresponding subject's bias-corrected standard deviation reported in Table 5. The covariate vector includes demographic characteristics (for example, gender, age), job-appointment characteristics (for example, experience in private and public schools, qualifications), and performance outcomes from national teachers' evaluations as proxies for teaching skills. This specification includes school fixed effects and missing-value dummies where the covariate information is unavailable; standard errors are clustered at the school level. In this exercise, I weight each teacher-level grading-bias coefficient by the inverse of the associated standard error, s_j^2, to increase precision.

Similarly, I estimate the following regression equation that relates teacher j's estimated grading bias and their implicit gender attitudes as measured by the IAT:

$$\hat{\theta}_j = \gamma_{1,s(j)} + IAT_j'\gamma_2 + X_j'\gamma_3 + \kappa_j \tag{5}$$

[28] Appendix Table F.4 reports the share of questions from each type of evaluation used to assess each learning competency for eighth grade. Only 5 out of 16 competencies were assessed on teacher-graded tests through a share of questions substantially different from the share used to assess the same competency in the standardized test (that is, share differences between five and nine percentage points).

[29] The second identifying assumption requires the skill differentials between boys and girls to not vary by teacher. This assumption ensures that, even if there are skill gaps across gender, the teacher-specific measure of gender bias has a meaningful interpretation for comparisons across exposure to different teachers.

Here, the teacher-level grading bias is divided by the bias-corrected standard deviation. IAT_j denotes the IAT score of teacher j, which has been standardized to have mean of zero and standard deviation of one. The index $s(j)$ indicates the school location s where teacher j is currently teaching according to the survey data; thus, $\gamma_{1,s(j)}$ are school-location fixed effects, and X_j is a vector of teacher characteristics that include both demographic and job-related characteristics drawn from administrative records and collected by survey. The covariates include teachers' children status (that is, whether they have daughters, sons, or both), ethnicity indicators, and the decade of birth; the order of the associations in the IAT (that is, whether order compatible or incompatible appears first); and the number of previous IATs the teacher has taken. Observations are at the teacher level, and standard errors are clustered at the school location level.

Relation with demographic and job characteristics. Teacher demographics and teaching experience are stronger predictors of assessment-based gender-bias than job experience or performance metrics in their jobs. Table 6 summarizes the estimated relations defined in Equation (4) for different configurations of teacher characteristics for the sample of mathematics instructors according to administrative records. The first two columns report coefficients for the entire sample of math teachers. The last column does so for the subsample of teachers who participated in the government's National Teacher Evaluation during 2015–19 in pursuit of tenure-track appointments. Columns (1) to (3) demonstrate that teachers with stronger gender biases toward female students tend to be women. This finding contrasts with that of Carlana (2019) and Alesina et al. (2018): female teachers are likely to exhibit less bias against girls, as measured by the IAT. I consider these results to document the extent to which gender biases have been internalized by female teachers in my setting, in accordance with the discussion in Section 2.

In addition, my results show that being older than the median age (45 years old) in this sample of teachers is associated with a weaker grading bias.[30] There are two possible arguments to explain this finding. One is in line with Altonji and Pierret's (2001) work that posits that prior prejudiced attitudes based on group membership (for example, race, gender) can be updated with continuous interactions with members of such groups. However, there is also a vast literature suggesting strong cohort effects on reported beliefs regarding discrimination.[31] In the Peruvian context, teachers' self-reported gender attitudes collected through my nationwide representative survey suggest there are no substantial cohort effects. Appendix Figure C.5 shows that, for the sample of surveyed mathematics teachers matched with a grading-bias measure, there is no marked relationship between teachers' age and gender attitudes across six self-reported measures. Nevertheless, analyzing the evolution of these gender attitudes is impossible with the available information.

Columns (2) and (3) provide evidence that more experienced teachers with private school backgrounds tend to exhibit weaker bias, while previous employment in public schools has less

[30] The median age in the sample of teachers reported in Columns (1) and (2) was 45 years in 2019, while for the sample of evaluated teachers in Column (3), the median age was 40 in 2019. However, I maintained the same definition of median age at 45 years across all columns in the table.

[31] In the US, in the last two decades racial attitudes have been declining overall, which has been attributed to time and cohort effects (Charles and Guryan, 2008).

bearing in predicting assessment bias. The pattern is consistent with Alan et al.'s (2018) finding that teachers with modern teaching styles are less likely to hold traditional beliefs about gender (including a warm approach and a growth mindset). In my empirical setting, teachers who opt to teach in, and have more experience in, private schools often adopt state-of-the-art teaching styles. Similarly essential are qualifications, as university-educated teachers tend to be less biased across all specifications. Column (3) indicates that among the evaluated teachers, those who performed better on a knowledge-based test and approved a nationwide round of evaluation usually exhibit less gender bias.[32]

Teachers' characteristics and qualifications are critical to comprehend to what extent the estimated systematic gaps in assigned scores between girls and boys actually reflect discriminatory behavior based on gender. However, it should not be understood as the driving force of gender bias. Instead, the estimated relationships provide further support regarding the direction of the discrimination: higher values of $\hat{\theta}_j$ correctly identify teachers biased against female students.

Relation with implicit gender attitudes. This study presents crucial new evidence on the extent to which implicit stereotypes, measured by IAT scores, undergird the observed bias-in-assessment estimates, $\hat{\theta}_j$. I leveraged survey data and matched a subsample of 2,052 mathematics and language arts teachers with IAT measures and the estimated grading-bias measure. The distribution of the IAT scores for teachers in this sample is displayed in Figure 3. The figure on the left indicates that male mathematics teachers have a higher IAT score across the distribution. In contrast, the figure on the right suggests that female language arts teachers have a higher IAT score across the distribution. This means that male math teachers more strongly associate male with science than female math teachers do, and the converse holds for language arts instructors. In fact, the average IAT score in math is 0.301, and among language arts teachers, the average is close, at 0.28. This finding indicates that the average high school teacher in Peru has an IAT score that almost places them in the maximum attainable category of gender bias. Moreover, gender biases are ubiquitous in the two main instruction subjects.

I use this survey-gathered information on gender biases to fit the specification in Equation (5). Each observation represents a teacher, the IAT score is the variable of interest, and robust standard errors are clustered at the school location level. Table 7 delivers evidence on the relationship between implicit stereotypes and biased grading against female students. First, I analyze the gender-science IAT's relation with bias in assessment.[33] Panel A reports coefficients in a specification with no controls other than school-location fixed effects. Panel B adds as controls a set of covariates referring to teachers' children, ethnicity, decade of birth, order of associations of IAT, and number of previously taken IATs. To my knowledge, this is the first study documenting the relation between observable bias in grading and implicit gender attitudes.

Column (1) indicates that teachers who more strongly associate male with science and female with humanities on the IAT are more likely than other teachers to exhibit strong grading bias

[32]The Ministry of Education indicates that between 2014 to 2019, about 30 nationwide evaluations were implemented, allowing more than 400,000 teachers to be evaluated Minedu (2015).

[33]The IAT scores have been standardized for this exercise.

against girls. In other words, teachers who think boys are better than girls in math- and science-related fields give boys better grades than they deserve in these subjects and give girls worse grades than they deserve. This relationship prevails whether or not I include covariates. In contrast, Column (2) suggests that teachers who strongly associate male with science and female with humanities are less likely to exhibit bias in grading girls in language arts. These coefficients imply that mathematics and language arts teachers penalize female students less when they assign grades in a subject or knowledge domain in which they believe that girls are more competent or have more potential. That the coefficients go in opposite directions and are consistent with the IAT scores in the different subjects provides compelling evidence that our measure of grading bias is a good metric for exposure to gender bias.

Finally, in light of these findings underscoring the heterogeneity of grading bias across subjects, I test the stability and consistency of grading bias exhibited by the same teacher across mathematics and language arts classes. We might expect that a biased teacher who holds stereotypes against women in science exhibits grading bias against girls in science and, to some degree, other subjects as well. I compute OLS correlations between teacher j's grading bias against female students in mathematics, language arts, and science. The sample for this exercise is limited, as high school teachers in this setting tend to specialize in a single subject. Appendix Figure C.6 shows coefficients measuring the relation between teacher-level grading bias across teacher-specific bias measures by subject, where the unit of analysis is the teacher. The relation between the grading bias across all combinations of subjects of instruction is positive and statistically significant, providing further evidence that the teacher-level measure of teacher bias is internally consistent.

4.3 Estimating the distribution of teachers' assessment bias

In this section, I describe the empirical Bayes method I use to correct sampling error in the set of estimates of $\hat{\theta}_j$. The resulting posterior means estimates serve to construct the primary variable of interest in the next step of my empirical analysis.

I leverage the set $\hat{\theta}_j$ of unbiased but noisy estimates of the true underlying teacher-level bias parameters θ_j using empirical Bayes (EB) methods. These methods are particularly useful, as they provide estimators that use the observed $\hat{\theta}_j$ for estimating aspects of the population distribution of the θ_j—beyond mean and variance—that might be of interest, even though the latter values are not themselves observed (Efron and Morris, 1972, 1975, Stein, 1964). Estimators in this framework have desired properties such as minimizing Bayes risk and minimizing quadratic error loss (Efron and Morris, 1972). Therefore, in this section I report precise estimates computed using two EB methods: parametric EB (also known as linear shrinkage estimation; see, for example, Chetty et al. (2014b), Gilraine et al. (2020), Kane and Staiger (2008)) and EB deconvolution (as proposed by Efron (2010, 2012, 2016)). I use these posterior mean estimators to construct the dependent variables in my analysis of the long-term effects of assessment bias on long-term outcomes.

4.3.1 Parametric empirical Bayes estimator

A natural way to model this problem is hierarchically by interpreting the teacher-level grading-bias estimates to be a random sample from a common *prior* population distribution of θ_j. I assume that the population parameters θ_j are drawn from a normal prior distribution with hyperparameters (μ, ϕ) so that θ_j's are conditionally independent given (μ, ϕ) :

$$\theta_j|\mu,\phi \overset{ind}{\sim} \mathcal{N}(\mu,\phi^2) \qquad \text{, for } j = 1,\dots,J \tag{6}$$

Next, consider J experiments (that is, classroom assignments) in which teacher assignment j estimates the parameter θ_j from a class size $N(j)$ that is independently distributed. This is written

$$\hat{\theta}_j|\theta_j \overset{ind}{\sim} \mathcal{N}(\theta_j, s_j^2) \tag{7}$$

for $j = 1,\dots,J$, where σ_θ^2 denotes the variance of $\hat{\theta}_j$. I am interested in the joint conditional posterior distribution $p(\theta|\mu,\phi,\hat{\theta})$ of the underlying parameters θ_j, which in this case has the form

$$\theta_j|\mu,\phi,\hat{\theta} \sim \mathcal{N}(\hat{\theta}_j^*, \Sigma_j) \tag{8}$$

and the posterior mean estimator is

$$\hat{\theta}_j^* = (\tfrac{1}{\sigma_j^2}\hat{\theta}_j + \tfrac{1}{\phi^2}\mu)/(\tfrac{1}{\sigma_j^2} + \tfrac{1}{\phi^2}), \tag{9}$$

with $\sigma_j^2 = \sigma_\theta^2/N(j)$ as the sampling variance.[34] A feasible version of this posterior mean estimator only requires the standard error of the associated $\hat{\theta}_j$. Therefore, under the previous assumptions, the parametric EB is a precision-weighted average of the prior population mean and the estimator, $\hat{\theta}_j$, that shrinks the estimated grading bias toward the respective sample mean. Figure 4 plots the posterior means' distribution under linear shrinkage. These values show lower variability as they are shrunk toward the sample mean.[35]

4.3.2 Application of empirical Bayes deconvolution methods

A key assumption of the parametric EB setup is that the prior distribution of the true parameter θ_j is Gaussian, which implies the posterior distribution specified in Equation (8) also is. The linear-shrinkage methods also carry features regarding the shrinkage procedure proposed in Equation (9). As Gilraine et al. (2020) indicates, the factor of such shrinkage only depends on the sign and not its magnitude. This property is particularly relevant in this setting, as the grading bias against girls and boys located at opposite sides of the distribution (that is, right tail and left tail, respectively) will be shrunk toward zero by the same factor as long as the class size and estimated bias are the same. To assess the importance of the normality assumption, I also estimate a more flexible model for the prior distribution of θ_j. In particular, I follow the modeling approach of the prior density proposed by Efron (2016), which uses an exponential

[34] It is easy to verify that $\Sigma_j = (\frac{1}{\sigma_j^2} + \frac{1}{\phi^2})^{-1}$.

[35] An issue not directly addressed here is the independence assumption of the estimated $\hat{\theta}_j$. Gilraine et al. (2020) proposes a test to formally test this assumption.

family of densities to model the prior, because of its efficiency compared to other deconvolution approaches.[36] In this setting, the econometrician assumes an unknown prior density, $g(\mu)$, yielding population parameters distributed as follows:

$$\theta_j \overset{ind}{\sim} G(\mu), \text{for } j = 1, \ldots, J \tag{10}$$

Here, each θ_j draws independently an observed $\hat{\theta}_j$ following a known exponential probability-density family f_j such that

$$\hat{\theta}_j \overset{ind}{\sim} f_j(\hat{\theta}_j|\theta_j). \tag{11}$$

Figure 5 displays the deconvolved density $\hat{g}(.)$ of the teacher-level bias parameters θ_j. I also report the theoretical Gaussian density of the estimated teacher-level grading bias and its observed distribution. Following Kline et al. (2021), I choose the penalization parameter such that the deconvolved density is calibrated in mean and variance using the bias-corrected variance estimates reported in Table 5. Visual inspection of deconvolved density indicates that Gaussian density is a good approximation for the prior density $g(.)$. These results suggest that a linear shrinkage estimator approximates the posterior means as if they were computed using the true underlying population.

5 Empirical strategy for estimating teachers' gender bias long-term effects

5.1 Research design of classroom exposure to teacher-level grading bias

Students' exposure to gender bias varies across school years, as teachers' assignments change across grades. In this research design, I compare students enrolled in the same cohort-grade-year-school cells who are assigned to different teachers and, therefore, are exposed to different levels of gender bias. Let $\hat{\theta}^*_{j,-i}$ denote the posterior mean of the leave-one-year-out estimator of grading bias of teacher j of student i. I standardized this estimator using its mean and the standard deviation computed per school year.[37]

The main estimating equation to retrieve the causal relationship between teachers' gender bias and student i's long-term outcomes variable, Y_i, is as follows,

$$Y_i = \delta_0 + \delta_1 \hat{\theta}^*_{j,-i} \cdot Female_i + \delta_2 Female_i + \delta_3 \hat{\theta}^*_{j,-i} + \delta_4 \mathbf{X}_i + u_i \tag{12}$$

The parameter δ_1 captures the differential effects on girls relative to boys of a one-standard-deviation increase in teacher bias against girls. The parameter δ_3 represents the effects of exposure to assessment bias among male students. $Female_i$ indicates whether student i is female.

[36] Various methods assess theoretically and empirically the deconvolution problem (for example, Fan (1991)). They can be grouped in g-modeling approaches when the prior distribution aims to be estimated Efron (2016) and in f-modeling approaches when the distribution of the observed parameters is retrieved (see, for instance, Carroll and Hall (1988), Efron (2014)).

[37] The standard deviation of teacher-level grading bias used for standardizing $\hat{\theta}^*_{j,-i}$ are calculated as the bias-corrected measures reported in Table 5 for the corresponding school year.

The covariate vector $\mathbf{X}_i$ comprises four groups of controls. First, I include student-level controls, including students' language, age in months, mother's education, and indicator variables for repetition and place of birth to account for students' social norms.[38] This set of covariates also includes teachers' characteristics such as gender, age, and indicators for the type of appointment (for example, homeroom, subject teacher, the teacher with administrative role), and, as a proxy for teacher experience, contract (for example, tenured or fixed-term contract). Teachers' behavior is potentially different across students' genders. Thus, to account for the differential effects of educational inputs on human capital decisions based on the student gender, the set of teacher characteristics is interacted with student gender. I also control for classroom-level characteristics, including class size, and student-level characteristics averaged at the classroom level to account for the composition effects of the assigned classes. The third group of controls includes school-grade averages of students' characteristics to account for confounders of exposure to biased teachers across grades in the same school. Lastly, I include second-degree polynomials of lagged individual students' scores in end-of-year examinations in mathematics and language arts and polynomials in classroom- and school-grade means of these scores. When estimating this model, I also include cohort, grade, school-year, and school fixed effects.

Leave-one-year-out estimate of bias in assessment. My preferred specification uses a leave-one-year-out version of the teacher-level grading-bias metric, denoted $\hat{\theta}^*_{j,-i}$, that is computed using the posterior-mean estimates introduced in Section 4.3. The leave-out approach avoids any correlation between the sampling variance in $\hat{\theta}_j$ and long-term outcomes that might be introduced by using the same students to study teacher bias and long-run outcomes. For example, consider the computation of the leave-one-year-out teacher's grading bias, which is the treatment variable of students expected to graduate in 2019. I exclude the data from all of the 2019 graduation cohort examinations (that is, data on eighth-grade students who took the examinations in 2015) and use the remaining students' observations to calculate all teachers' grading bias. I repeat this exercise for students in all other (projected) graduating cohorts.[39] This calculation results in a leave-one-year-out teacher-level assessment bias estimate assigned to each student cohort based on their projected graduation year.

5.2 Identification of long-term effects

A significant concern in estimating the effects of exposure to teachers' grading bias is that students might not be randomly assigned to teachers' classrooms. As a result, the estimated effects of teachers' gender bias might not capture uniquely the effects of their gender stereotypes on students' decisions (as reflected in student outcomes) but also the influence of other gender-specific unobservables that systematically drive these educational investments, such as ability differentials reinforced by gendered social norms.

[38] Following Carlana (2019), I consider student social norms to be related to the place of birth, and therefore, I include birthplace indicators as a covariate. Studies in sociology also document this relation. See for instance, Crawley et al. (2013), Uunk and Lersch (2019).

[39] Only the cohorts with projected graduations in 2018 and 2019 took the standardized examinations in 2015 and 2016, respectively. I construct a leave-one-year-out measure of teachers' grading bias, excluding scores from years 2015 and 2016, and assign those to these cohorts exclusively. The remaining cohorts are assigned the regular teacher-grading bias estimates that pool information from 2015, 2016, 2018, and 2019 scores.

To formally examine the potential confounding variables that prevent me from isolating teachers' bias in assessment, I build on the potential-outcomes framework.[40] Consider that the variable for potential long-term outcomes, $\tilde{Y}_{ij}$, for student i assigned to teacher j depends on the scores given by the teacher and central planner:

$$\tilde{Y}_{ij} = f_i(X_i, \tilde{S}^N_{ij}, \tilde{S}^T_{ij}, \rho_i) \tag{13}$$

For instance, $\tilde{Y}_{ij}$ might refer to high school completion. Here, ρ_i is a composite of students' unobserved characteristics (for instance, their educational aspirations or prior beliefs regarding their academic potential) while X_i is a set of student-demographic characteristics affecting academic-progress decisions. Moreover, student i's centrally assigned score if they were assigned to teacher j is defined as follows:

$$\tilde{S}^N_{ij} = s_i(M_i, W_{ij}, \psi_i, \eta_{ij}) \tag{14}$$

M_i is an indicator variable denoting the membership of student i in one of the two classroom groups (for example, boys). This specification allows for group membership to directly affect centrally assigned test scores through factors other than stereotypes. W_{ij} is a covariate vector of characteristics from student i and their assigned teacher j. ψ_i represents the innate ability or academic potential of the student. The score assigned by the teacher j if assigned to student i is defined as follows:

$$\tilde{S}^T_{ij} = c_i(M_i, W_{ij}, \psi_i, \vartheta_{j,g(i)}, \eta'_{ij}) \tag{15}$$

The parameter $\vartheta_{j,g(i)} \in \mathbb{R}$ captures teacher j's forecast of student i's ability based on the student's group membership g. We can think of this unobserved parameter as corresponding to teachers' gender bias against a given group if the prior beliefs of any student in that group are systematically lower.[41] Unlike in the previous equation, in which a blind examiner assigns the score, here I allow the examiner to idiosyncratically forecast student i's ability based on the student's group membership $g(i)$. Both η_{ij} and η'_{ij} are unobserved determinants of scores and are orthogonal to all other factors in their respective equations. I assume they have means of zero and are identically distributed and pairwise independent for any pair (i, j).

I now consider how the human capital investment decision depends on on potential scores assigned by the teacher and central planner:

$$\tilde{Y}_{ij} = f_i(X_i, M_i, W_{ij}, \psi_i, \vartheta_{j,g(i)}, \tilde{\eta}_{ij}) \tag{16}$$

Here, $\tilde{\eta}_{ij}$ combines any unobserved gender-driven difference in student or teacher characteristics, other than teachers' bias, that influences the outcome with any unobserved characteristic correlated with gender that influences the outcome. Among the former type of unobservable, for instance, there might be parents' requests to assign their children to a teacher with gen-

[40] See Online Appendix Section B.2 within a dynamic human capital investment model. In that setting, I allow students to have asymmetric information on their abilities or academic potential and teachers to have gender-biased assessments and provide performance feedback that guides pupils' decisions. In addition, centrally assigned scores are given as signals for students' decisions.

[41] A formal definition is introduced in Definition 1 in Online Appendix Section B.2.

der attitudes aligned with their own progressive views regarding equality in education. The latter group of unobservables might include classroom assignments for which high-performing students are more likely to be assigned to unbiased teachers that do not interfere with their self-assessment of academic potential.

In light of these concerns, the following equation expresses the selection-on-observables assumption on which I retrieve the causal estimates of being exposed to a teacher with greater grading bias against any student based on their group membership:

$$\mathbb{E}[\tilde{Y}_{ij}|\hat{\theta}^{*}_{j,-i}, M_i, \mathbf{X}_i, j(i) = j] = \lambda_0 + \lambda_1 M_i + \lambda_2 \hat{\theta}^{*}_{j,-i} + \lambda_3 \mathbf{X}_i; \qquad j = 1 \dots J \tag{17}$$

Here, $\mathbf{X}_i \equiv [X_i, W_{ij}]$ is a covariate vector of teacher characteristics, student characteristics, and lagged scores, and $\hat{\theta}^{*}_{j,-i}$ is the empirical counterpart of the parameter introduced in Equation (3). In other words, a student's assignment to teacher j is assumed to be orthogonal to unobserved determinants in levels and unobserved differences across genders that guide the decisions regarding high school completion, college outcomes, and labor market outcomes, after conditioning on observed characteristics of students, teachers, classroom-level and school-grade-level characteristics, and lagged student scores. This selection-on-observables assumption is often proposed in the value-added literature to alleviate selection concerns driven by students' sorting mainly on ability (see Chetty et al. (2014a,b), Kane and Staiger (2008), Rothstein (2010, 2017)). In this case, the conditioning variables include observable characteristics of the student, $\mathbf{X}_i$, that directly determine their investment in qualifications as well as other characteristics M_i, W_{ij}—such as educational inputs, students' group membership, and school and teacher characteristics—that influence students' decisions regarding whether to continue academically.

Another consideration for the identification of the underlying causal parameter is that the estimated teacher-level grading bias is computed using the scores of cohorts of students attending eighth grade during a specific school year. I assume that the grading bias exerted by a teacher toward previous years and specific cohort-grades is predictive of the bias exerted toward future cohorts in other grades as long as it is displayed in the same subject. This is consistent with the value-added literature that measures teachers' or schools' value added based on the scores of students in previous periods and from other students' cohorts (Abdulkadiroğlu et al., 2020, Rothstein, 2017).

6 Results on academic progress and college outcomes results

6.1 Effects on high school completion

My first main finding is that female students, who are as good as randomly assigned to more biased math teachers against girls, are less likely to graduate high school on time or ever. Table 8 shows the effects of exposure to a one-standard-deviation more biased mathematics teacher against girls, during one grade, on the likelihood of graduating high school. The outcome variable in Columns (1) to (4) is an indicator variable equal to 1 if the student graduated in the school year corresponding to their projected high school graduation date. Columns (5) to (8)

have as outcome variables an indicator of whether the student graduated a calendar year after their projected graduation date. The window of time over which graduating from high school *ever* is defined varies from cohort to cohort, ranging from one calendar year after projected graduation for the youngest cohort (class of 2019) to four for the oldest cohort (class of 2015).

Exposure to a more biased teacher increases the gender gap in high school graduation. This can be seen in Table 8, which reports estimates of Equation (12) separately by grade. In Column (1), the interaction coefficient indicates that a one-standard-deviation increase in teacher gender bias during eighth grade lowers the likelihood of graduating from high school by 3.6 percentage points for girls relative to boys. The following columns suggest that exposure to a more biased teacher in ninth grade has a differential effect on girls of 1.7 percentage points, while if exposed during tenth grade, girls are 1.5 percentage points less likely to graduate relative to boys. Overall, the point estimates follow a slightly decreasing trend over time as the students advance to senior grades.

These grade-specific estimates are aggregated into a single summary effect in Column (4). This specification stacks the data for all high school grades, with standard errors clustered by student. On average, exposure to a more biased math teacher for one grade during high school widens the high school graduation gap by 1.5 percentage points (1.8% of the mean), disfavoring girls. Columns (5) through (8) repeat this analysis with an outcome variable equal to one if the student ever graduated from high school, including late graduation. The summary of the points estimates is reported in Column (8). In this case, a one-standard-deviation increase in bias is estimated to increase the gender gap in graduation by 1.3 percentage points (1.6% of the mean). Boys' exposure to a one-standard-deviation more biased teachers increases the likelihood of graduation by 0.6 percentage points (0.7% of the mean).

6.2 Effects on college attendance

This section discusses the causal effects of teachers' gender bias on college outcomes — specifically, concerning enrollment in four-year colleges, not two-year colleges or vocational training programs. Students' average age at graduation is 16 years, and most students decide to apply to college in their junior year of high school (that is, tenth grade).[42]

The persistence of teachers' bias effects negatively impacts the college application gender gap. Table 9 presents the estimates of exposure to teachers' gender bias once the students graduate from high school and start their college trajectories. Columns (1) and (2) show a statistically significant impact on the gender gap in college applications. The estimates indicate that a one-standard-deviation increase in bias decreases girls' likelihood to apply to college on time by 0.6 percentage points relative to boys. Similarly, a one-standard-deviation increase in bias lowers the likelihood that girls ever applying to college by 0.7 percentage points relative to boys (2.2% of mean). These results differ from the findings of a recent study by Lavy and Megalokonomou (2019), in which exposure to more biased teachers against girls in the last two high school grades decreases the likelihood of postsecondary enrollment for girls and increases it for boys. In contrast, the exposure to bias among boys do not show statistically significant effects on their likelihood of applying to university. The estimates in Columns (3) and (4) show

[42] Online Appendix Section A.1.3 provides a brief description of the university system in the country.

no significant impact of teacher bias on the gender gap in college admissions. A similar result is found with respect to college enrollment reported in the last two columns. In combination with the results for high school graduation, these results suggest that the primary impact of teacher bias on educational attainment operates at the margin of high school graduation.

The evolution of the effects of grade-specific exposure to teachers' gender bias on college applications draws a detailed picture of how relevant it is in this setting junior year (eleventh grade) and freshman year (eighth grade). Figure 6 plots the main effects of exposure to teacher-level grading bias against girls on the likelihood of college application and the interaction effects for girls by high school grade. I find negative and statistically significant effects on the gender gap in college application during eighth and eleventh grade. The large effects of teacher bias found during eighth grade are plausibly driven by the performance signal sent to parents and teachers based on students' performance on the national standardized tests. Students' parents receive a detailed report comparing the student's mathematics, language arts, and science scores to the national average student performance and the performance of pupils per gender. Their performance is discretized to communicate to parents whether their children's achievement levels are "Satisfactory," "In the process," "Beginning," or "Before beginning." In addition, a detailed report is sent to teachers, in which the students are ranked within the class and their performance is compared to other eighth-grade students in the region and to the country's average performance.

6.3 Robustness checks on high school completion and college outcomes

The empirical exercise aims to identify two parameters introduced in Equation (12). First, the interaction parameter δ_1 captures the differential effects on outcomes among girls exposed to more biased teachers against women by one standard deviation above the mean grading bias. Second, the parameter δ_3 is interpreted as the effect on male students' outcomes of being exposed to a teacher with grading bias that is one standard deviation higher than the average teacher's bias. The first potential threat to identification arises if there are observed or unobserved characteristics based on which students sort to more or less biased teachers. I address this concern by showing in Table 10 that students' and teachers' baseline characteristics potentially drive such assignments are not systematically related to the teacher-level grading-bias measure.

The second threat stems from the possibility that teachers' bias toward eighth-grade students systematically differs from their bias exerted toward upper-grade students. This can potentially happen if principals assign more conservative and biased teachers to, for instance, instruct the upper grades of high school. To test whether such a scenario is of concern, I calculate the long-term effects on a restricted sample in which I only include the cohorts graduating in 2018 and 2019, which can be observed both in lower and upper grades. The results reported in Appendix Table C.2 to Appendix Table C.4 show that the set of estimates of the effects of math teachers' bias on graduation outcomes remains statistically significant in the described subsample of students. Furthermore, I also find statistically significant effects on college outcomes for the restricted sample, as displayed in Appendix Table C.4.

Finally, there is also the possibility of unobserved patterns in teacher assignments to classrooms across grades such that the bias measure for some grades is not missing at random. To

examine this potential threat, I define a sub-sample that comprises all student cohorts whose exposure to teachers' bias is observed during all grades 7 to 11–that is, for every grade they appear in the sample. The results reported in Appendix Tables C.5 to C.7 also does not indicate that this is a concern.

6.4 Teachers' bias in assessment versus value added

As discussed in Section 2, a channel that can potentially mediate the effects I find on academic progression and college outcomes is teachers' value added. Therefore, I examine whether teachers exhibit a differential value added toward boys versus girls in their classes. These gender-specific value-added estimates allow me to investigate whether higher-quality teachers are more or less gender-biased.

To formally analyze this mechanism, consider the estimating Equations (1) and (2), which form a system. I am interested in retrieving the following parameters. First, I define $\alpha_{1,j}$ as teacher j's value added toward girls and $\alpha_{1,j} + \alpha_{2,j}$ as teacher j's value added toward boys. I compute these parameters using a fixed-effects–SURE specification according to the previous equations, allowing for error terms to be correlated across equations.

With estimates of teacher value-added and gender bias in hand, I calculate the joint covariance matrix of these parameters. This intermediate step is needed for conducting a bias correction using the sampling covariances calculated from SURE estimation. Bias-corrected estimates of the correlations in value-added and bias parameters across teachers appear in Table 11 (see details on the estimation procedure in Online Appendix E). Consistent with Lavy and Megalokonomou (2019), I find that teachers with more gender bias in mathematics exhibit lower value-added toward both girls and boys. Consistently with Table 5, I find a lower dispersion of gender bias among mathematics teachers than among language arts teachers under this specification. The estimated correlations in Columns (2) and (3) reported in Panel A indicate a stronger negative association with value added against boys. The same patterns are present in Panel B among language arts teachers. This could reflect ripple effects on the classroom climate when teachers are more biased against girls (Bertrand and Pan, 2013, Figlio et al., 2019). In addition, I find an almost perfect correlation in value added for boys and girls in math and language arts, which differs from the gender-based match effects found in recent literature (Dee, 2005, Gershenson et al., 2018). This analysis suggests that low-value-added teachers tend to be more biased. Consequently, teachers' quality is an essential mechanism by which gender biases reduce high school completion and college applications.[43]

[43] Another mechanism I investigated is student tracking into biased teachers across grades in Appendix Table C.8. I found a positive correlation between bias in one grade and the next grade. However, this is explained by repeated assignments of the same teacher, as the correlation is no longer positive if I restrict the sample to students who did not get the same teacher in consecutive years.

7 Effects on labor market outcomes

7.1 Employment in the formal sector

This section reports the main findings on the set of labor market outcomes for the students' cohorts graduating between 2015 and 2019, whose trajectories I follow from age 18 to 23 (the former is the minimum legal age). The regression sample allows me to analyze the effects of teachers' gender bias on a sample of high school graduates who either have full-time jobs or work part-time in the formal sector while attending higher education institutions. Most of the students' positions are non-professional entry-level jobs, given the age of the workers. My results indicate that female students' increased exposure to teachers' gender bias affects formal sector employment significantly and persistently among girls aged 18–23.[44]

I examine the effects of exposure to a more biased teacher using the research design proposed in Section 5. In the specification used in this section, I account for local labor market dynamics related to social norms by interacting the student's place of birth with their gender. Table 12 display the estimated effects on the likelihood of formal sector employment from the empirical model in Equation (12). Each column indicates the modal ages of students throughout their first three years in the labor markets. Column (1) reports the estimated effects when students in the sample have reached the minimum legal working age of 18 years. The effects of teachers' gender bias on the gender gap in formal employment are 0.1 percentage points (18% of the mean). The bottom rows of the table report the mean outcome values for female and male students, indicating that roughly 93 percent of 18–19-year-old girls in the sample are likely working in the informal sector. Moreover, the detrimental effects on the gender gap in formal employment are 2.4 percentage points at ages 22–23 (21% of the mean), as shown in Column (3). This is a large effect equal to 27 percent of the baseline gender gap and 18 % of the mean formal sector employment level.

Figure 7 summarizes the evolution of teachers' gender bias effects by presenting analogous coefficients to those reported in Table 12. The lines display the estimated main effects of teachers' bias, the interaction effects, and the gender gap on formal sector employment; the confidence intervals are displayed in the shaded regions. The differential effects on girls remain negative and persistent until ages 22–23. Moreover, the total grading-bias effect on female students is negative and statistically significant at ages 18–19 and 20–21, ranging between -0.08 percentage points and -0.64 percentage points, respectively. In contrast, the total effect on male students is positive across all years: is 1.18 at ages 18–19 and 2.51 percentage points at ages 20–21.

Two main features of the labor market in the country support the salient total and differential effects that I find. On the one hand, the size of the informal labor sector in the country is considerable and vastly absorbs young workers. Between 2015 and 2019, the average share of informal jobs nationwide is 64 percent, above most of the countries in the region (International

[44] In Appendix Table C.13, I report the effects of teachers' bias on students' employment outcomes between ages 14 and 17 who are considered underage children with legal impediments protecting them from having any job, including those in the formal sector. The minimum legal age to work in the country is 18, so hiring companies will likely require a written permit from parents or legal wardens for these students to have legal employment contracts in the formal sector. Therefore, I focus on the labor outcomes of students starting at age 18, which reflect the decision without the parents' influence.

Monetary Fund, 2022). Informal sector jobs in Peru do not comply with labor law stipulating the right to have a contract, a retirement pension fund, health insurance, or unemployment insurance. On the other hand, another distinct characteristic of the labor markets in Peru is that the educational attainment of the labor force is underdeveloped: 68 percent of workers in the labor force have a high school diploma or less.[45] In this context, female high school graduates without job experience find it more challenging to obtain any position in the formal sector: the share of female workers under 29 years old is the lowest among workers with formal jobs (10 percent in 2018). In sum, attaining formal sector employment is highly competitive for high school graduates, given that they are unable to compensate for their lack of employment experience with qualifications that differentiate them from the rest of the workforce. My results suggest that encountering biased educators during high school will likely sidetrack students into detrimental employment trajectories in the informal sector.

7.2 Paid working hours

I also find evidence of strong large effects of gender bias on paid monthly work hours in the formal sector, a result that appears in Columns (1) to (3) in Table 13. Column (1) indicates adverse effects among female graduates ages 18–19. Increased exposure to a one-standard-deviation more biased teacher during one grade in high school decreases the paid hours worked by around one hour each month for women. This differential effect is equivalent to around 28 percent of the paid work hours gender gap; nevertheless, it seems to fade over time. In Columns (2) and (3), the unfavorable effects on women are no longer statistically significant. Similarly to the effects found on the likelihood of holding a formal sector job, the total effects on women's paid working hours (the sum of the first and second row) are no longer disfavorable to women aged 22–23. However, there is not a clear pattern of convergence to the amount of paid hours worked by men, who are working almost twice as much as women at age 22-23.

7.3 Monthly earnings

I now present estimates of the monthly earnings losses of recent high school graduates between the ages of 18 and 23. The point estimates in Columns (4) to (6) in Table 13 are expressed in 2010 USD. The first row indicates that women endure persistent effects that aggravate the earnings gender gap by ages 18–19. My estimates in Column (4) indicate that for women relative to men, the monthly earnings loss from having a teacher who is one standard deviation more biased during a single grade is USD 2.6 per month at this age –roughly three years after graduating high school. The baseline monthly earnings gap in the sample by the time the students have 18–19 years is equivalent to USD 8.3 per month; therefore, my findings indicate that exposure to gender bias exacerbates the earnings gender gap by approximately one-third. Moreover, when measured against the monthly minimum wage, the point estimate in Column (4) indicates that the monthly loss in earnings for women exposed to a one-standard-deviation more biased teacher is equivalent to 1 percent of the monthly minimum wage at ages 18–19.

[45] According to the National Institute of Statistics and Informatics of Peru, in 2019, 45 percent of the working-age population had a high school diploma as the maximum educational level attained, while 24 percent had elementary schooling as the educational level attained.

The average minimum wage during this period is approximately USD 240 per month.

Moreover, although the effects at ages 20–21 and 22–23 reported in Columns (5) and (6) are not statistically significant, they have economic importance. A potential cause for teachers' gender bias effects on women to dissipate at this stage is that women have gained significant on-the-job experience to offset the low investments endured during high school. Also, work-study employees might have passed a milestone in higher education degrees that allows detrimental effects to be less significant. In fact, by the time students are 20–21, those enrolled in a university or technical higher education have accumulated enough academic credits to start paid apprenticeships that are regulated by law and overseen by their respective higher education institutions.

Turning to the total effects on male (second row) and female graduates (sum of the first and second rows), Columns (4) to (6) in Table 13 show that the total effects on women are adverse starting at ages 18–19 and fade out only when students are aged 22–23. Therefore, any earnings catch-up by women starts five years after high school, leaving most of them with a long-lived comparative disadvantage in labor markets compared to male graduates. In contrast, exposure to a more biased teacher leads to earning-gains patterns for men aged 18 to 23 —that is, five years after graduating from high school.

Next, I assess whether the effects on earnings arise from low- or high-paying jobs by estimating the effects of teachers' gender bias on the probability of attaining a high-paying job. In this exercise, I maintain the same research design proposed in Equation (12), using as an outcome an indicator variable that takes the value of 1 when the student i's monthly earnings are above a given earnings percentile. The percentiles are computed from the population of workers with comparable work experience between ages 16 to 25. Figure 8 plots, for each earning percentile, the effects on the probability of having a job with earnings above that percentile when the workers are ages 18–19. Here, I consider the jobs above the fifty percentile as high-earning jobs. Panel (a) shows the women's differential effects and the baseline gender gap of increased exposure to teachers' gender bias in the likelihood of having a high-earning job. Similarly, Panel (b) shows the main and differential effects. The first Panel demonstrates that girls whose monthly earnings correspond to the bottom fifty percentile are less likely to have high-earning jobs when exposed to a one-standard-deviation more biased teacher. Panel B confirms that the total effect on the likelihood of high-earning jobs among girls in the bottom fifty percentile is negative and persistent at ages 18–19 –three years out of high school.

My results indicate marked effects on the monthly earnings gap deriving from low-value-added biased teachers' that disproportionately deviate female students onto low-paying jobs when employed in the formal sector. A partial explanation of why the effects, as a share of the pay gap, are substantial lies in the weak enforcement of recent legislation stipulating equal pay for women and men.[46] As a result of the weak legal protection –which is common in middle and low-income countries– the gender earnings gap in Peru is substantial both in the formal and informal labor markets. For example, in 2015, the raw monthly earnings gap between male and female workers was 29.9 percent (INEI, 2020), while the adjusted earnings gap in hourly

[46]In 2017, laws to avoid pay discrimination between men and women were passed by the Peruvian government. However, since the law's enactment, the judiciary has started prosecuting pay discrimination in 2021.

wages was 19 percent in 2015 (Muller and Paz, 2018). Unbinding equal pay rights enforcement places little to no boundaries on earning disparities generated in schools.

Moreover, consistent with the literature on the gender pay gap, the distribution of males and females across occupations and industries is a key explanation for persistent pay differences (Cortes and Pan, 2018, Kunze, 2018). Blau et al. (2012) and Blau and Kahn (2017) emphasized that gender differences leading to occupational sorting are pronounced and play a significant role in the observed pay gaps —especially among non-college-educated women whose lack of qualifications prevents them from equalizing payment with male colleagues. Following this line of work, Goldin (2014) suggests that convergence of payment has been more favorable in occupations at the upper tail of the wage distribution. Less is known about what other factors take relevance in explaining earnings gaps at the lower tail of the wage distribution, where the high school graduates are most likely to be.

Appendix Figure C.7 indicates that industry sorting is a likely channel for educators' influence on earnings trajectories among young low-skill workers in Peru. My findings suggest that female high school graduates are deterred from participating in an increasing number of industries in the formal sector as they spend more years out of high school. The results show that biased teachers deter women and, at the same time, encourage men to seek employment in industries such as "agriculture, forestry, fishing," "transportation and warehousing," "mining, quarrying, oil and gas extraction" and "real estate and rental and leasing" between ages 18-23. Notably, the figure shows that teachers' bias does not affect differently females and males' participation in industries such as "finance and insurance,", "educational services," "health care and social assistance," and "'other services" where high-skilled professionals tend to be employed in Peru. Overall, exposure to biased teachers exacerbates the gender pay gap by inducing strong sorting effects among male students and narrower sorting mediation among females.

In sum, the prevalence of weak labor rights protection and gender-based occupation sorting shapes a setting where the role of educators in determining human capital and vocational interests can sidetrack students toward low-paying jobs where they are likely to face high wage discrepancies.

7.4 Students' internalization of gender bias

One potential mechanism through which biased teachers might affect students is changes in the student's own gender views. For example, a biased teacher might increase negative gender attitudes among girls; a phenomenon known as "internalized bias," as discussed in Section 2. I study this possibility using IAT scores collected in a sample of a younger cohort of students currently in high school and who are projected to graduate between 2022 and 2025. In particular, I use the gender-science IAT scores of these students. The information for this exercise was collected through remote and in-person sessions according to the timeline displayed in Figure 2. Focusing on the in-person sessions, approximately 4,000 students from 266 classrooms from eighth to eleventh grade participated in the data-collection sessions. The student sessions were carefully designed to promote the active participation of students and were carried out by an external educator leading these class-format workshops in which the students filled out the IAT

and student surveys.[47] The IAT is a psychological test designed initially for adults, although its validity has also been investigated in children between 7 and 17 years old (Roh et al., 2018). The version that 13–16-year-old students took was adapted from the original version to ensure they comprehended the test's language and contents and stayed engaged while taking the test. Although an external facilitator is in charge of executing the session activities, these were carried out in the presence of a teacher (neither mathematics nor a language instructor) previously appointed for this task by the school principal.

Using this information, I estimate the following regression that relates student i's gender bias, as measured by the IAT, with the teacher-level grading bias, $\hat{\theta}^*_{j,-i}$, and maintains the same research design presented in Section 5.1.[48]

$$IAT_i = \nu_0 + \nu_1 \cdot \hat{\theta}^*_{j,-i} \, Female_i + \nu_2 Female_i + \nu_3 \hat{\theta}^*_{j,-i} + \nu_4 \mathbf{X}_i + \tilde{e}_i \tag{18}$$

I do not find evidence supporting the hypothesis that exposure to biased teachers led students to internalize the gender bias of their educators in mathematics or language arts classes. Column (1) in Table 14 indicates that female students who, during one grade in high school, are in classes of math teachers who are one standard deviation more biased show a decrease in their IAT score. This effect is interpreted as reducing girls' gender biases against women attempting science disciplines. Nevertheless, the main effect reported in the second row almost entirely offsets the differential effect among girls, amounting to a total effect on girls close to zero. Regarding the exposure effects of biased language arts teachers, Column (2) indicates that biased teachers in language arts decrease the IAT score among girls reflecting a decay in their gender attitudes. The total effects on girls are overall negative but non-significant. These coefficients are not statistically significant, which prevents me from concluding the prevalence of internalization of teachers' gender biases among boys and girls.[49]

8 Consequences of exposure to bias in other subjects

This section describes the consequences of educators' gender biases when exerted in language arts and science classes. To begin, language arts teachers' gender bias has statistically significant detrimental effects on girls' academic progress across all high school grades examined. Columns (1) to (3) of Appendix Table C.9 show that female students exposed to language arts teachers who are one standard deviation more biased against girls have timely high school completion rates that are 0.04–0.01 percentage points lower. This pattern is reprised in Column (4), which indicates a differential effect of a similar order to the one reported for mathematics teachers on high school completion after students' projected graduation year. A similarity between mathematics teachers and language arts teachers is shown in the second row of Appendix Table C.9, in which the main effects of gender bias are reported. As with math, the exposure to

[47] The selection criteria for determining the sample of schools that were part of the in-person sessions are the following: (i) the school is in the Lima metropolis, (ii) the school principal provides the authorization for the in-person session, and (iii) the school has verified well-functioning computer labs with internet connection. The educators hired for the data collection either had a bachelor's degree or were senior college students majoring in education or psychology. The 90-minute data-collection sessions took place during classroom time.

[48] The variable IAT_i has been standardized to have mean of zero and standard deviation of one.

[49] This hypothesis can offer additional insights into the internalization of biases in future versions of this study.

more biased instructors in language arts brings a net benefit for male students according to the stacked grades estimate displayed in Column (4). However, these positive effects continue to be more persistent than in the case of mathematics teachers, extending beyond the eighth grade and into the ninth grade. Columns (5) to (8) of Appendix Table C.9, which pertain to the likelihood of ever graduating from high school, suggest a similar conclusion.

Another source of more negative consequences for academic progress among girls compared to boys concerns their science teachers. Appendix Table C.10 shows that only the interaction effects of teachers' gender bias on girls are statistically significant, while the main exposure effects on bias cease to be significant. Importantly, as shown in Columns (4) and (8) these differential effects are of a lower order—around half the magnitudes found among mathematics and language arts teachers.

Moving on to college-attendance outcomes, the findings I report in Appendix Table C.11 indicate that language arts instructors, unlike math instructors, do not directly affect boys' or girls' college application, admission, or enrollment outcomes. The same pattern regarding science-teacher exposure is presented in Appendix Table C.12. In addition, I consider the recurrent but transitory effects of language arts teachers' bias on girls' labor market outcomes. A conservative interpretation of the results reported in Appendix Tables C.14 to C.16 is that the differential effects of gender bias only appear in formal sector employment, paid hours worked, and monthly earnings for up to two years. These three results are largely repeated in Appendix Tables C.17 to C.19, in which it can be seen that persistence over time is not a factor in exposure to language arts teachers, whose biases affect students up to two years after graduating high school. Similarly, the effects of science teachers' bias only last till one year after graduation.

The lower intensity and lesser persistence of exposure to biases in language arts and science teachers compared to mathematics educators underscore the widely documented importance of the latter subject in academic achievement and progression.

9 Conclusions

I estimated the long-term effects of exposure to gender-biased teachers for 1.7 million students from five cohorts graduating from public schools in Peru. I tracked these students for almost a decade, starting when they were 13 years old and in eighth grade to when they turned 22 years old and were studying in higher education institutions or working. I employed a comprehensive set of administrative and survey data to characterize the distribution of their teachers' gender bias. My primary measure of bias is based on the discrepancies between teacher-assigned and centrally assigned scores to male and female students using a sample of around 60,000 teachers. This is the first study to validate that educators' implicit, and to some extent unconscious, gender attitudes are reflected in their *observed* behavior through the assessment of their students in mathematics, language arts, and science. Moreover, exploring these two measures allows me to conclude that assessment bias strongly correlates with implicit attitudes measured using survey data on a widely administered psychological test.

The central contribution of this study is its estimate that exposure to teachers' gender bias exacerbates gender gaps in earnings and formal employment between ages 18–23 for up to five

years after high school graduation. My rich set of controls allowed me to address potential sources of student-to-teacher sorting. I emphasized that educators' gender biases in students' assessment are a previously undocumented source of gender gaps in earnings, indicating educators' stereotypes' role in fostering the adverse conditions women face in labor markets.

Another key contribution is my analysis of the mechanisms that allow the effects of gender bias to persist in students' postsecondary trajectories. Teachers' distinct value-added toward female and male students is a potential mechanism, as it provides unabridged educational inputs that, according to the literature, have persistent effects on earnings. I considered the argument that teachers' signaling regarding students' performance can reflect their implicit gender attitudes and potentially distorts students' beliefs regarding boys' versus girls' adequacy in science versus humanities. In other words, I tested whether teachers' biases changed their students' implicit gender attitudes as measured by the IAT. The information I collected on students' IAT scores does not allow me to conclude that the gender biases of educators are internalized by their students.

Given the pervasiveness of teachers' gender biases and their adverse effects on students, an urgent task is to design policy interventions to reduce the prevalence of such biases in educational settings. It also remains to be studied whether students can be trained to recognize teachers' gender biases and prevent the biases from affecting their human capital decisions.

References

Aaronson, Daniel, Lisa Barrow, and William Sander, "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics*, January 2007, *25* (1), 95–135.

Abdulkadiroğlu, Atila, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters, "Do Parents Value School Effectiveness?," *American Economic Review*, May 2020, *110* (5), 1502–1539.

Abowd, John M, Paul Lengermann, and Kevin L McKinney, "The Measurement of Human Capital in the U.S. Economy," 2003, p. 101.

Alan, Sule, Enes Duysak, Elif Kubilay, and Ipek Mumcu, "Social Exclusion and Ethnic Segregation in Schools: The Role of Teacher's Ethnic Prejudice," *The Review of Economics and Statistics*, September 2021, pp. 1–45.

_ , Seda Ertac, and Ipek Mumcu, "Gender Stereotypes in the Classroom and Effects on Achievement," *The Review of Economics and Statistics*, December 2018, *100* (5), 876–890.

Alesina, Alberto, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti, "Revealing Stereotypes: Evidence from Immigrants in Schools," Technical Report w25333, National Bureau of Economic Research, Cambridge, MA December 2018.

Altonji, J. G. and C. R. Pierret, "Employer Learning and Statistical Discrimination," *The Quarterly Journal of Economics*, February 2001, *116* (1), 313–350.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters, "Leveraging Lotteries for School Value-Added: Testing and Estimation*," *The Quarterly Journal of Economics*, May 2017, *132* (2), 871–919.

Babad, Elisha Y, Jacinto Inbar, and Robert Rosenthal, "Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers.," *Journal of educational psychology*, 1982, *74* (4), 459. Publisher: American Psychological Association.

Barbanchon, Thomas Le, Roland Rathelot, and Alexandra Roulet, "Gender Differences in Job Search: Trading off commute against wage," *The Quarterly Journal of Economics*, December 2020, *136* (1), 381–426.

Becker, Sascha O., Ana Fernandes, and Doris Weichselbaumer, "Discrimination in hiring based on potential and realized fertility: Evidence from a large-scale field experiment," *Labour Economics*, August 2019, *59*, 139–152.

Ben-Shakhar, Gershon and Yakov Sinai, "Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies," *Journal of Educational Measurement*, March 1991, *28* (1), 23–35.

Berinsky, Adam J., "The Two Faces of Public Opinion," *American Journal of Political Science*, October 1999, *43* (4), 1209.

Berniell, Inés, Lucila Berniell, Dolores de la Mata, María Edo, and Mariana Marchionni, "Gender gaps in labor informality: The motherhood effect," *Journal of Development Economics*, May 2021, *150*, 102599.

Bertoni, Eleonora, Gregory Elacqua, Diana Hincapié, Carolina Méndez, and Diana Paredese, "Teachers' Preferences for Proximity and the Implications for Staffing Schools: Evidence from Peru," *Education Finance and Policy*, April 2022, pp. 1–32.

Bertrand, Marianne, "New perspectives on gender," in "Handbook of labor economics," Vol. 4, Elsevier, 2011, pp. 1543–1590.

_ and Jessica Pan, "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior," *American Economic Journal: Applied Economics*, January 2013, *5* (1), 32–64.

_ , **Claudia Goldin, and Lawrence F Katz**, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics*, July 2010, *2* (3), 228–255.

_ , **Dolly Chugh, and Sendhil Mullainathan**, "Implicit Discrimination," *American Economic Review*, April 2005, *95* (2), 94–98.

Biasi, Barbara and Heather Sarsons, "Flexible Wages, Bargaining, and the Gender Gap," Technical Report w27894, National Bureau of Economic Research, Cambridge, MA October 2020.

Blau, Francine D. and Lawrence M. Kahn, "The Gender Wage Gap: Extent, Trends, and Explanations," *Journal of Economic Literature*, September 2017, *55* (3), 789–865.

Blau, Francine, Peter Brummund, and Albert Yung-Hsu Liu, "Trends in Occupational Segregation by Gender 1970-2009: Adjusting for the Impact of Changes in the Occupational Coding System," Technical Report w17993, National Bureau of Economic Research, Cambridge, MA April 2012.

Bobba, Matteo, Tim Ederer, Gianmarco Leon-Ciliotta, Christopher Neilson, and Marco Nieddu, "Teacher Compensation and Structural Inequality: Evidence from Centralized Teacher School Choice in Peru," Technical Report w29068, National Bureau of Economic Research, Cambridge, MA July 2021.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes*," *The Quarterly Journal of Economics*, November 2016, *131* (4), 1753–1794.

Botelho, Fernando, Ricardo A. Madeira, and Marcos A. Rangel, "Racial Discrimination in Grading: Evidence from Brazil," *American Economic Journal: Applied Economics*, October 2015, *7* (4), 37–52.

Burgess, Simon and Ellen Greaves, "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities," *Journal of Labor Economics*, July 2013, *31* (3), 535–576.

Buttrick, Nicholas, Jordan Axt, Charles R. Ebersole, and Jacalyn Huband, "Reassessing the incremental predictive validity of Implicit Association Tests," *Journal of Experimental Social Psychology*, May 2020, *88*, 103941.

Card, David, Jörg Heining, and Patrick Kline, "Workplace Heterogeneity and the Rise of West German Wage Inequality*," *The Quarterly Journal of Economics*, August 2013, *128* (3), 967–1015.

Carlana, Michela, "Implicit Stereotypes: Evidence from Teachers' Gender Bias*," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1163–1224.

Carrell, Scott E., Marianne E. Page, and James E. West, "Sex and Science: How Professor Gender Perpetuates the Gender Gap *," *Quarterly Journal of Economics*, August 2010, *125* (3), 1101–1144.

Carroll, Raymond J. and Peter Hall, "Optimal Rates of Convergence for Deconvolving a Density," *Journal of the American Statistical Association*, December 1988, *83* (404), 1184–1186. Publisher: Taylor & Francis.

Charles, Kerwin Kofi and Jonathan Guryan, "Prejudice and Wages: An Empirical Assessment of Becker's *The Economics of Discrimination*," *Journal of Political Economy*, October 2008, *116* (5), 773–809.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, September 2014, *104* (9), 2593–2632.

_ , _ , **and** _ , "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, September 2014, *104* (9), 2633–2679.

Coate, Stephen and Glenn C. Loury, "Will Affirmative-Action Policies Eliminate Negative Stereotypes?," *The American Economic Review*, 1993, *83* (5), 1220–1240. Publisher: American Economic Association.

Correll, Shelley J., Stephen Benard, and In Paik, "Getting a Job: Is There a Motherhood Penalty?," *American Journal of Sociology*, March 2007, *112* (5), 1297–1339.

Cortes, Patricia and Jessica Pan, "Occupation and Gender," in Susan L. Averett, Laura M. Argys, and Saul D. Hoffman, eds., *The Oxford Handbook of Women and the Economy*, Oxford University Press, July 2018, pp. 424–452.

Crawley, Heaven, Stephen Drinkwater, and Rukhsana Kauser, "Regional Variations in Attitudes Towards Refugees: Evidence from Great Britain," *SSRN Electronic Journal*, 2013.

David, E. J. R., Tiera M. Schroeder, and Jessicaanne Fernandez, "Internalized Racism: A Systematic Review of the Psychological Literature on Racism's Most Insidious Consequence," *Journal of Social Issues*, December 2019, *75* (4), 1057–1086.

de Educación Superior Universitaria Minedu, Dirección General, "Sistema de Recolección de la información para Educación Superior (SIRIES)," February 2020.

Dee, Thomas S., "A Teacher like Me: Does Race, Ethnicity, or Gender Matter?," *The American Economic Review*, 2005, *95* (2), 158–165. Publisher: American Economic Association.

DeMars, Christine E., "Gender Differences in Mathematics and Science on a High School Proficiency Exam: The Role of Response Format," *Applied Measurement in Education*, July 1998, *11* (3), 279–299.

_ , "Test Stakes and Item Format Interactions," *Applied Measurement in Education*, January 2000, *13* (1), 55–77.

DHS, "DHS STEM Designated Degree Program List," July 2021.

Efron, Bradley, "The Future of Indirect Evidence," *Statistical Science*, May 2010, *25* (2).

_ , *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1, Cambridge University Press, 2012.

_ , "Two Modeling Strategies for Empirical Bayes Estimation," *Statistical Science*, May 2014, *29* (2).

_ , "Empirical Bayes deconvolution estimates," *Biometrika*, March 2016, *103* (1), 1–20.

_ **and Carl Morris**, "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, March 1972, *67* (337), 130–139.

_ **and** _ , "Data Analysis Using Stein's Estimator and its Generalizations," *Journal of the American Statistical Association*, June 1975, *70* (350), 311–319. Publisher: Taylor & Francis.

Egloff, Boris and Stefan C. Schmukle, "Predictive validity of an implicit association test for assessing anxiety.," *Journal of Personality and Social Psychology*, December 2002, *83* (6), 1441–1455.

Ertl, Bernhard, Silke Luttenberger, and Manuela Paechter, "The Impact of Gender Stereotypes on the Self-Concept of Female Students in STEM Subjects with an Under-Representation of Females," *Frontiers in Psychology*, May 2017, *8*, 703.

Ewens, Michael and Richard R. Townsend, "Are early stage investors biased against women?," *Journal of Financial Economics*, March 2020, *135* (3), 653–677.

Exley, Christine L and Judd B Kessler, "The Gender Gap in Self-Promotion," *The Quarterly Journal of Economics*, July 2022, *137* (3), 1345–1381.

Fan, Jianqing, "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, September 1991, *19* (3).

Figlio, David, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman, "Family disadvantage and the gender gap in behavioral and educational outcomes," *American Economic Journal: Applied Economics*, 2019, *11* (3), 338–81.

Fisher, Robert J., "Social Desirability Bias and the Validity of Indirect Questioning," *Journal of Consumer Research*, September 1993, *20* (2), 303.

FitzGerald, Chloë and Samia Hurst, "Implicit bias in healthcare professionals: a systematic review," *BMC Medical Ethics*, December 2017, *18* (1), 19.

Fleche, Sarah, Anthony Lepinteur, and Nattavudh Powdthavee, "Gender Norms and Relative Working Hours: Why Do Women Suffer More than Men from Working Longer Hours than Their Partners?," *AEA Papers and Proceedings*, May 2018, *108*, 163–168.

Flory, J. A., A. Leibbrandt, and J. A. List, "Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions," *The Review of Economic Studies*, January 2015, *82* (1), 122–155.

Fortin, Nicole M, Philip Oreopoulos, and Shelley Phipps, "Leaving boys behind gender disparities in high academic achievement," *Journal of Human Resources*, 2015, *50* (3), 549–579. Publisher: University of Wisconsin Press.

Francesconi, Marco and Matthias Parey, "Early gender gaps among university graduates," *European Economic Review*, October 2018, *109*, 63–82.

Fryer, Roland G and Steven D Levitt, "An Empirical Analysis of the Gender Gap in Mathematics," *American Economic Journal: Applied Economics*, April 2010, *2* (2), 210–240.

Gallagher, Ann M and Richard De Lisi, "Gender differences in Scholastic Aptitude Test: Mathematics problem solving among high-ability students.," *Journal of Educational Psychology*, 1994, *86* (2), 204. Publisher: American Psychological Association.

Gershenson, Seth, Cassandra M. Hart, Joshua Hyman, Constance Lindsay, and Nicholas Papageorge, "The Long-Run Impacts of Same-Race Teachers," Technical Report w25254, National Bureau of Economic Research, Cambridge, MA November 2018.

Gilraine, Michael, Jiaying Gu, and Robert McMillan, "A New Method for Estimating Teacher Value-Added," Technical Report w27094, National Bureau of Economic Research, Cambridge, MA May 2020.

Glover, Dylan, Amanda Pallais, and William Pariente, "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores*," *The Quarterly Journal of Economics*, August 2017, *132* (3), 1219–1260.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, "Performance in competitive environments: Gender differences," *The quarterly journal of economics*, 2003, *118* (3), 1049–1074. Publisher: MIT Press.

Goldin, Claudia, "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, April 2014, *104* (4), 1091–1119.

_ **and Cecilia Rouse**, "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, September 2000, *90* (4), 715–741.

Gray-Lobe, Guthrie, Parag Pathak, and Christopher Walters, "The Long-Term Effects of Universal Preschool in Boston," Technical Report w28756, National Bureau of Economic Research, Cambridge, MA May 2021.

Greenwald, Anthony G., Brian A. Nosek, and Mahzarin R. Banaji, "Understanding and using the Implicit Association Test: I. An improved scoring algorithm.," *Journal of Personality and Social Psychology*, 2003, *85* (2), 197–216.

_ **,** _ **,** _ **, and K. Christoph Klauer**, "Validity of the salience asymmetry interpretation of the Implicit Association Test: Comment on Rothermund and Wentura (2004).," *Journal of Experimental Psychology: General*, 2005, *134* (3), 420–425.

_ , **Debbie E. McGhee, and Jordan L. K. Schwartz**, "Measuring individual differences in implicit cognition: The implicit association test.," *Journal of Personality and Social Psychology*, 1998, *74* (6), 1464–1480.

Hardy, Morgan and Gisella Kagy, "Mind The (Profit) Gap: Why Are Female Enterprise Owners Earning Less than Men?," *AEA Papers and Proceedings*, May 2018, *108*, 252–255.

Harris, Monica J. and Robert Rosenthal, "Mediation of interpersonal expectancy effects: 31 meta-analyses.," *Psychological Bulletin*, May 1985, *97* (3), 363–386.

Hoge, Robert D and Theodore Coladarci, "Teacher-based judgments of academic achievement: A review of literature," *Review of educational research*, 1989, *59* (3), 297–313. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

Houwer, Jan De and Agnes Moors, "How to define and examine the implicitness of implicit measures," *Implicit measures of attitudes: Procedures and controversies*, 2007, pp. 179–194.

_ , **Sarah Teige-Mocigemba, Adriaan Spruyt, and Agnes Moors**, "Implicit measures: A normative analysis and review.," *Psychological Bulletin*, May 2009, *135* (3), 347–368.

Hyde, Janet S. and Janet E. Mertz, "Gender, culture, and mathematics performance," *Proceedings of the National Academy of Sciences*, June 2009, *106* (22), 8801–8807.

ILO, *What lies behind gender pay gaps* number 6th (2018/19). In 'Global wage report.', Geneva: ILO, 2018.

_ , *The business case for change* Women in business and management, Geneva: ILO, 2019.

_ , "ILO modelled estimates database, ILOSTAT," 2020.

INEI, "Clasificador nacional de programas e instituciones de educación superior universitaria, pedagógica, tecnológica y técnico productiva 2018," Technical Report, Instituto Nacional de Estadística e Informática, Lime 2018.

INEI, "Perú: Brechas de Género 2020. Avances hacia la igualdad de mujeres y hombres," Technical Report, Lima, Peru 2020.

International Monetary Fund, "Peru: 2022 Article IV Consultation-Press Release; Staff Report; and Statement by the Executive Director for Peru," 2022.

Jackson, C Kirabo, "What do test scores miss? The importance of teacher effects on non–test score outcomes," *Journal of Political Economy*, 2018, *126* (5), 2072–2107. Publisher: University of Chicago Press Chicago, IL.

Jacob, Brian A. and Lars Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics*, January 2008, *26* (1), 101–136.

Jones, C. P., "Levels of racism: a theoretic framework and a gardener's tale," *American Journal of Public Health*, August 2000, *90* (8), 1212–1215.

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

Kahneman, Daniel and Amos Tversky, "On the psychology of prediction.," *Psychological Review*, July 1973, *80* (4), 237–251.

Kane, Thomas and Douglas Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," Technical Report w14607, National Bureau of Economic Research, Cambridge, MA December 2008.

Kane, Thomas J and Douglas O Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, November 2002, *16* (4), 91–114.

Keane, Michael P., Petra E. Todd, and Kenneth I. Wolpin, "The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications," in "Handbook of Labor Economics," Vol. 4, Elsevier, 2011, pp. 331–461.

Klein, Stephen P., Jasna Jovanovic, Brian M. Stecher, Dan McCaffrey, Richard J. Shavelson, Edward Haertel, Guillermo Solano-Flores, and Kathy Comfort, "Gender and Racial/Ethnic Differences on Performance Assessments in Science," *Educational Evaluation and Policy Analysis*, June 1997, *19* (2), 83–97.

Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller, "Child Penalties across Countries: Evidence and Explanations," *AEA Papers and Proceedings*, May 2019, *109*, 122–126.

Kline, Patrick and Christopher Walters, "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination," *Econometrica*, 2021, *89* (2), 765–792.

_ , Evan Rose, and Christopher Walters, "Systemic Discrimination Among Large U.S. Employers," Technical Report w29053, National Bureau of Economic Research, Cambridge, MA July 2021.

Krysan, Maria, "Privacy and the Expression of White Racial Attitudes: A Comparison Across Three Contexts," *The Public Opinion Quarterly*, 1998, *62* (4), 506–544. Publisher: [Oxford University Press, American Association for Public Opinion Research].

Kunze, Astrid, "The Gender Wage Gap in Developed Countries," in Susan L. Averett, Laura M. Argys, and Saul D. Hoffman, eds., *The Oxford Handbook of Women and the Economy*, Oxford University Press, July 2018, pp. 368–394.

Lachowska, Marta, Alexandre Mas, and Stephen A. Woodbury, "Sources of Displaced Workers' Long-Term Earnings Losses," *American Economic Review*, October 2020, *110* (10), 3231–3266.

Lane, Kristin A, Mahzarin R Banaji, Brian A Nosek, and Anthony G Greenwald, "Understanding and using the Implicit Association Test: IV: What we know (so far) about the method.," 2007. Publisher: The Guilford Press.

Lavy, Victor, "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment," *Journal of Public Economics*, October 2008, *92* (10-11), 2083–2105.

_ and Edith Sand, "On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases," *Journal of Public Economics*, November 2018, *167*, 263–279.

_ and Rigissa Megalokonomou, "Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study," Technical Report w26021, National Bureau of Economic Research, Cambridge, MA June 2019.

Liu, Ou Lydia and Mark Wilson, "Gender Differences in Large-Scale Math Assessments: PISA Trend 2000 and 2003," *Applied Measurement in Education*, April 2009, *22* (2), 164–184.

Minedu, "Evaluación Nacional Docente," 2015.

_ , "Currículo Nacional de la Educación Básica," June 2016.

Minedu, "Marco de fundamentación de las pruebas de la Evaluación Censal de Estudiantes," 2017.

_ , "SíseVe," May 2022.

Molina, Kristine M. and Drexler James, "Discrimination, internalized racism, and depression: A comparative study of African American and Afro-Caribbean adults in the US," *Group Processes & Intergroup Relations*, July 2016, *19* (4), 439–461.

Muller, Miriam and Carmen de Paz, *Gender Gaps in Peru*, World Bank, Washington, DC, December 2018.

Narasimhan, Balasubramanian and Bradley Efron, "**deconvolveR** : A G-Modeling Program for Deconvolution and Empirical Bayes Estimation," *Journal of Statistical Software*, 2020, *94* (11).

NCES, "Classification of Instructional Programs," 2020.

Niederle, Muriel and Lise Vesterlund, "Explaining the Gender Gap in Math Test Scores: The Role of Competition," *Journal of Economic Perspectives*, May 2010, *24* (2), 129–144.

Nosek, Brian A., Anthony G. Greenwald, and Mahzarin R. Banaji, "The Implicit Association Test at Age 7: A Methodological and Conceptual Review.," in "Social psychology and the unconscious: The automaticity of higher mental processes." Frontiers of social psychology., New York, NY, US: Psychology Press, 2007, pp. 265–292.

_ , Yoav Bar-Anan, N. Sriram, Jordan Axt, and Anthony G. Greenwald, "Understanding and Using the Brief Implicit Association Test: Recommended Scoring Procedures," *PLoS ONE*, December 2014, *9* (12), e110938.

OECD, *OECD Employment Outlook 2015* OECD Employment Outlook, OECD, July 2015.

_ , "Gender wage gap," Technical Report, OECD 2022. Type: dataset.

Paulhus, Delroy L., "Measurement and Control of Response Bias," in "Measures of Personality and Social Psychological Attitudes," Elsevier, 1991, pp. 17–59.

Paxson, C., "The Allocation and Impact of Social Funds: Spending on School Infrastructure in Peru," *The World Bank Economic Review*, August 2002, *16* (2), 297–319.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies.," *Journal of Applied Psychology*, 2003, *88* (5), 879–903.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales, "How stereotypes impair women's careers in science," *Proceedings of the National Academy of Sciences*, March 2014, *111* (12), 4403–4408.

Risse, Leonora, Lisa Farrell, and Tim R L Fry, "Personality and pay: do gender gaps in confidence explain gender gaps in wages?," *Oxford Economic Papers*, October 2018, *70* (4), 919–949.

Roberts, Steven O. and Michael T. Rizzo, "The psychology of American racism.," *American Psychologist*, April 2021, *76* (3), 475–487.

Roh, Daeyoung, Soo-Young Bhang, Jung-Seok Choi, Yong Sil Kweon, Sang-Kyu Lee, and Marc N. Potenza, "The validation of Implicit Association Test measures for smartphone and Internet addiction in at-risk children and adolescents," *Journal of Behavioral Addictions*, March 2018, *7* (1), 79–87.

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

Rosenthal, Robert and Lenore Jacobson, "Pygmalion in the classroom," *The Urban Review*, September 1968, *3* (1), 16–20.

Rothstein, Jesse, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement *," *Quarterly Journal of Economics*, February 2010, *125* (1), 175–214.

_ , "Measuring the Impacts of Teachers: Comment," *American Economic Review*, June 2017, *107* (6), 1656–1684.

Rousille, N., "The Central Role of the Ask Gap in Gender Pay Inequality," 2021.

Rubie-Davies, Christine, John Hattie, and Richard Hamilton, "Expecting the best for students: Teacher expectations and academic outcomes," *British Journal of Educational Psychology*, September 2006, *76* (3), 429–444.

Rubie-Davies, Christine M., "Teacher expectations and perceptions of student attributes: Is there a relationship?," *British Journal of Educational Psychology*, March 2010, *80* (1), 121–135.

Rudman, Laurie A. and Richard D. Ashmore, "Discrimination and the Implicit Association Test," *Group Processes & Intergroup Relations*, July 2007, *10* (3), 359–372.

Sarsons, H., "Interpreting Signals in the Labor Market: Evidence from Medical Referrals," 2017.

Sorkin, Isaac, "Ranking Firms Using Revealed Preference*," *The Quarterly Journal of Economics*, August 2018, *133* (3), 1331–1393.

Steffens, Melanie C., "Is the Implicit Association Test Immune to Faking?," *Experimental Psychology*, June 2004, *51* (3), 165–179.

Stein, Charles, "Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean," *Annals of the Institute of Statistical Mathematics*, December 1964, *16* (1), 155–160.

Stöber, Joachim, "The Social Desirability Scale-17 (SDS-17)," *European Journal of Psychological Assessment*, September 2001, *17* (3), 222–232.

UNESCO, "Country profile: Peru," November 2016. Last Modified: 2017-04-12.

Uunk, Wilfred and Philipp M Lersch, "The Effect of Regional Gender-Role Attitudes on Female Labour Supply: A Longitudinal Test Using the BHPS, 1991–2007," *European Sociological Review*, October 2019, *35* (5), 669–683.

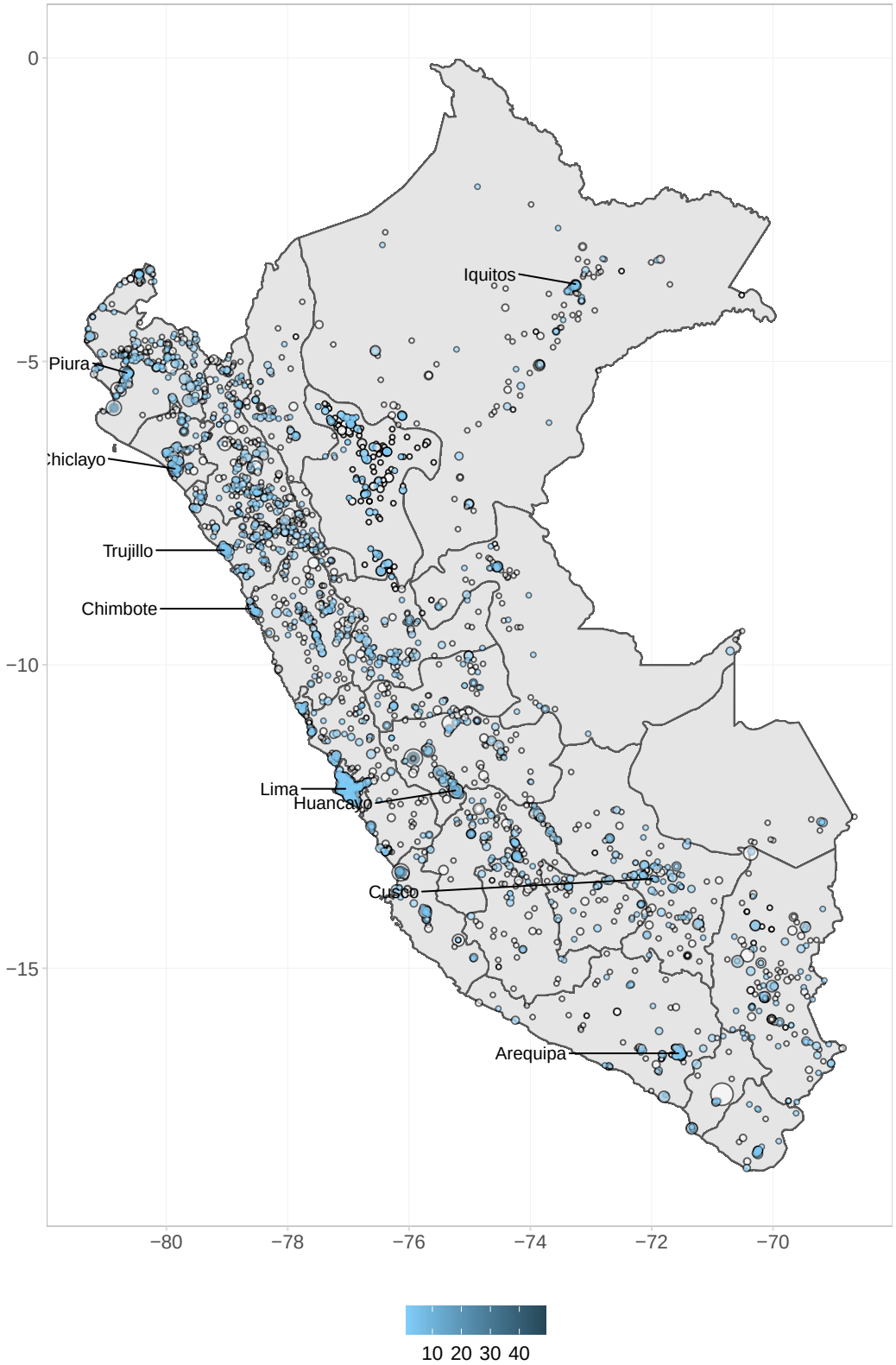
van den Bergh, Linda, Eddie Denessen, Lisette Hornstra, Marinus Voeten, and Rob W. Holland, "The Implicit Prejudiced Attitudes of Teachers: Relations to Teacher Expectations and the Ethnic Achievement Gap," *American Educational Research Journal*, June 2010, *47* (2), 497–527.

Willis, Henry A., Effua E. Sosoo, Donte L. Bernard, Aaron Neal, and Enrique W. Neblett, "The Associations Between Internalized Racism, Racial Identity, and Psychological Distress," *Emerging Adulthood*, August 2021, *9* (4), 384–400.

Wiswall, Matthew and Basit Zafar, "Preference for the Workplace, Investment in Human Capital, and Gender*," *The Quarterly Journal of Economics*, February 2018, *133* (1), 457–507.

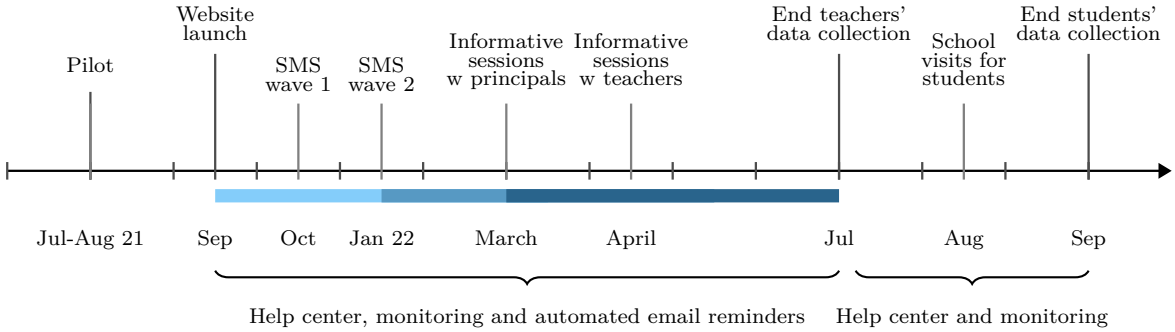
World Economic Forum, *The global gender gap report 2018*, Geneva, Switzerland: World Economic Forum, 2018. OCLC: 1126007340.

Figure 1: Map of teachers' survey completion



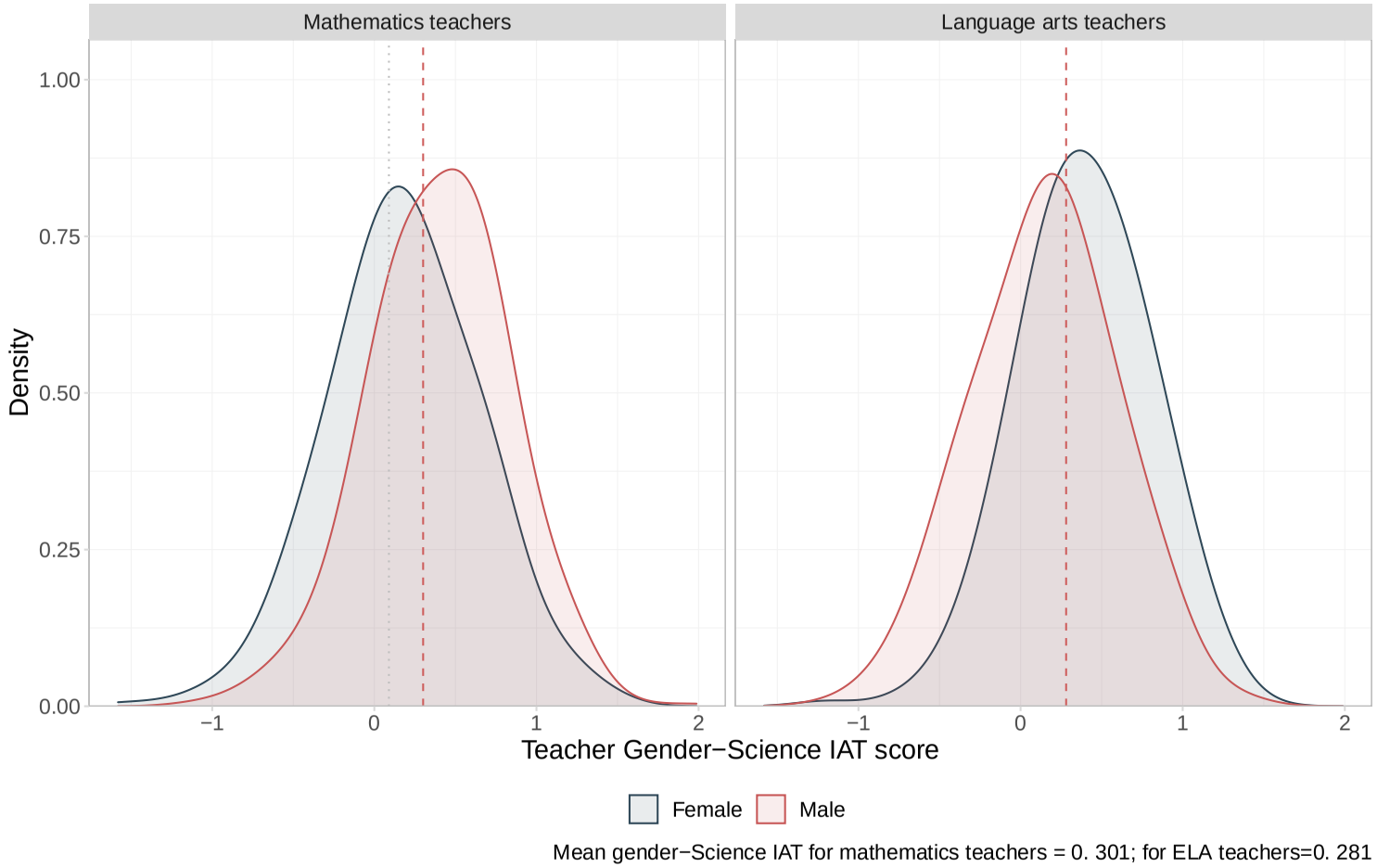
Notes: The map plots the number of teachers who registered to complete the survey and those who completed it binned at the school level. The binning uses the school names teachers are employed in as of 2019–20. White circles represent teachers that signed up for the survey but did not complete it, while the blue circles represent teachers who signed up and completed the survey. Teachers in 2,115 schools signed up, and teachers in 1,673 schools completed the survey. The teachers who completed the survey are distributed across schools as follows: 624 schools have one survey-taker teacher (46.9%), 687 (51.7%) have 2–10 survey-taker teachers and 17 schools have between 11 and 49 survey-taker teachers (1.3%).

Figure 2: Data-collection timeline



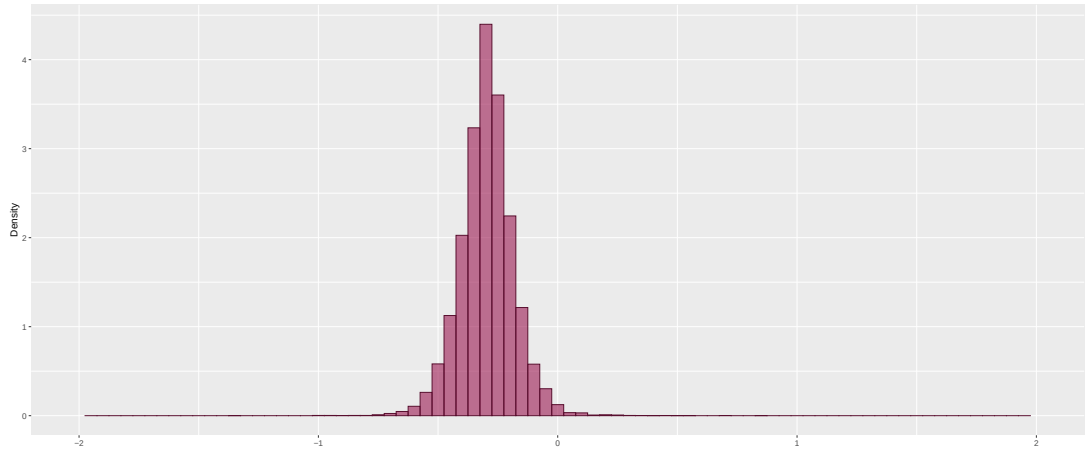
Notes: The graph summarizes the data-collection phases that started in September 2021, ended in July 2022 for teachers, and ended in August 2022 for students. The data collection encompassed the last quarter of the school year 2021 and the first three quarters of 2022. The color bar indicates the last quarter of the 2021 school year (the summer-vacation period) and the first two quarters of the 2022 school year, respectively. Before the data collection, instruments and website functionality were tested in a pilot involving around 30 teachers that participated in individual interview sessions. The website was launched in September 2022, starting the data-collection period for teachers. Sampled participants were invited to participate in the website through two waves of SMS. Once teachers registered on the website, they were sent up to four automated email reminders to complete the partially completed activities (that is, the Implicit Association Test [IAT] and the survey). At the start of the 2022 school year, informational meetings were held online for sampled teachers and their principals. Student data collection started in August 2022 and ended in September 2022. Enumerators visited schools in the metropolis of Lima, guiding students to complete the IAT and survey during the homeroom teachers' tutoring hours. Throughout the data-collection and website-functioning period, a help center was available through email, phone, and WhatsApp to aid teachers and students in the remote data collection.

Figure 3: Distribution of teachers' gender-science IAT scores

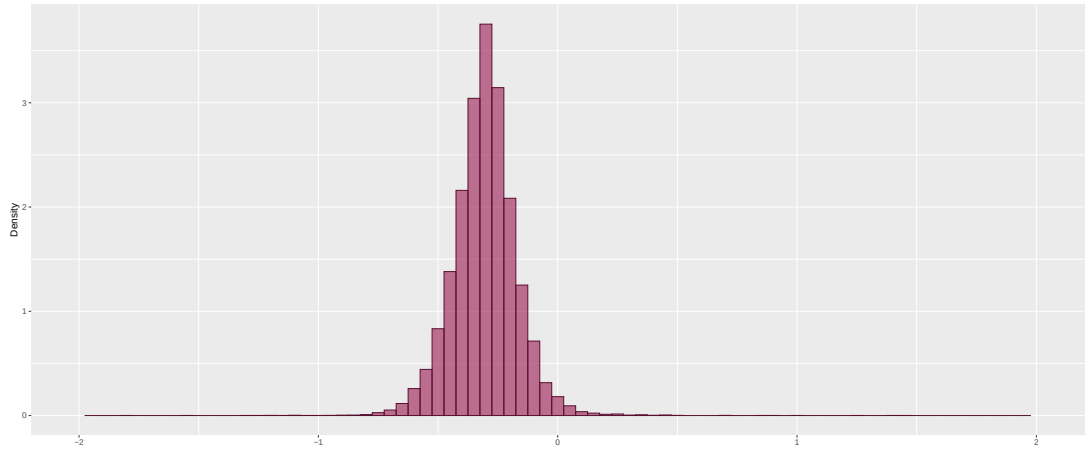


Notes: This figure presents the density of IAT scores for 2,052 (Math= 1,102; ELA=950) teachers successfully matched with administrative records, with an available measure of teacher-level estimated grading bias. According to Nosek et al. (2007), IAT scores below -.15 are *in favor of girls*, between -.15 and .15 indicate *little to no bias against girls*, .15 to .35 indicate *slight bias against girls*, and scores above .35 indicate *moderate to severe bias against girls*.

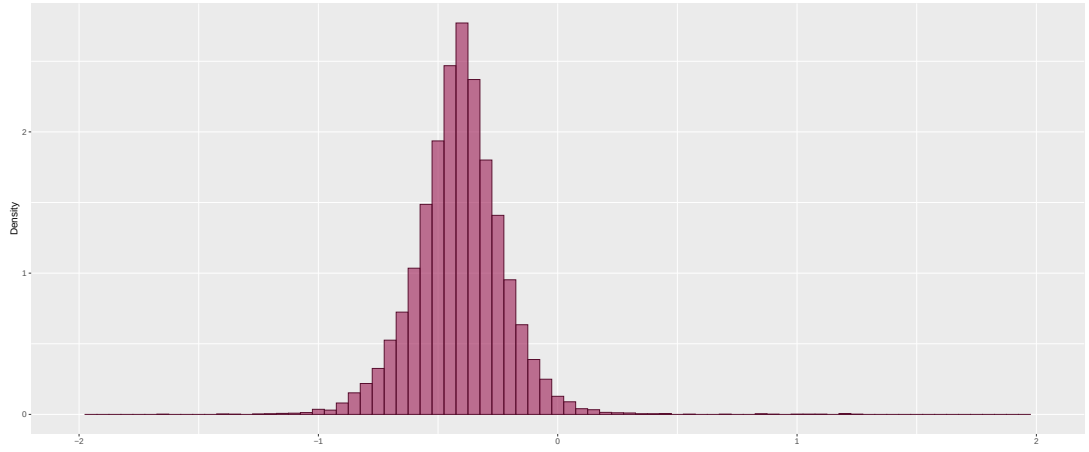
Figure 4: Distribution of posterior mean of teacher-level grading-bias estimates under linear shrinkage



(a) Mathematics



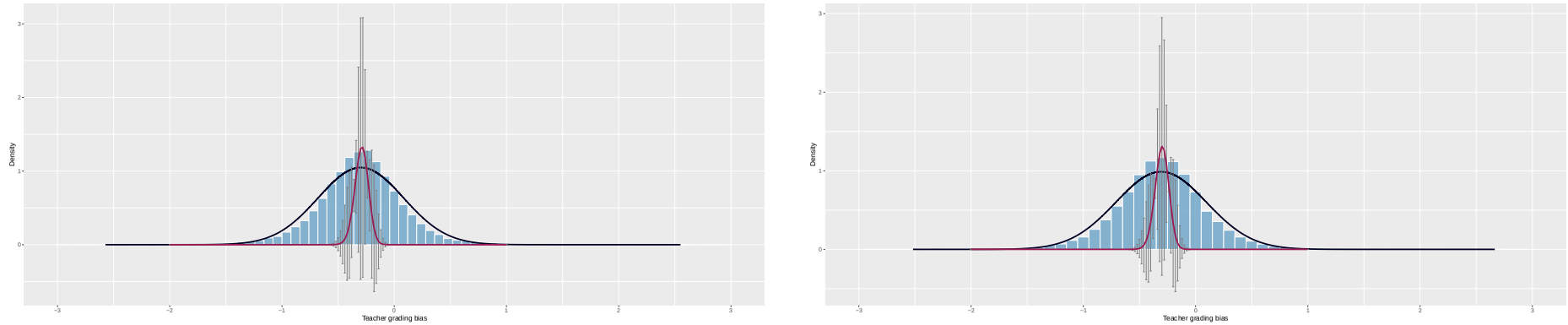
(b) Language arts



(c) Science

Notes: The graph shows the distributions of posterior means and $\hat{\theta}_j^*$ of teacher-level bias estimates $\hat{\theta}_j$ as defined in Equation (9). The teacher-level bias estimates are shrunk toward the group mean. For visualization, the estimates are limited to -4 to +4.

Figure 5: Distribution of posterior means of teacher-level grading bias under EB deconvolution methods

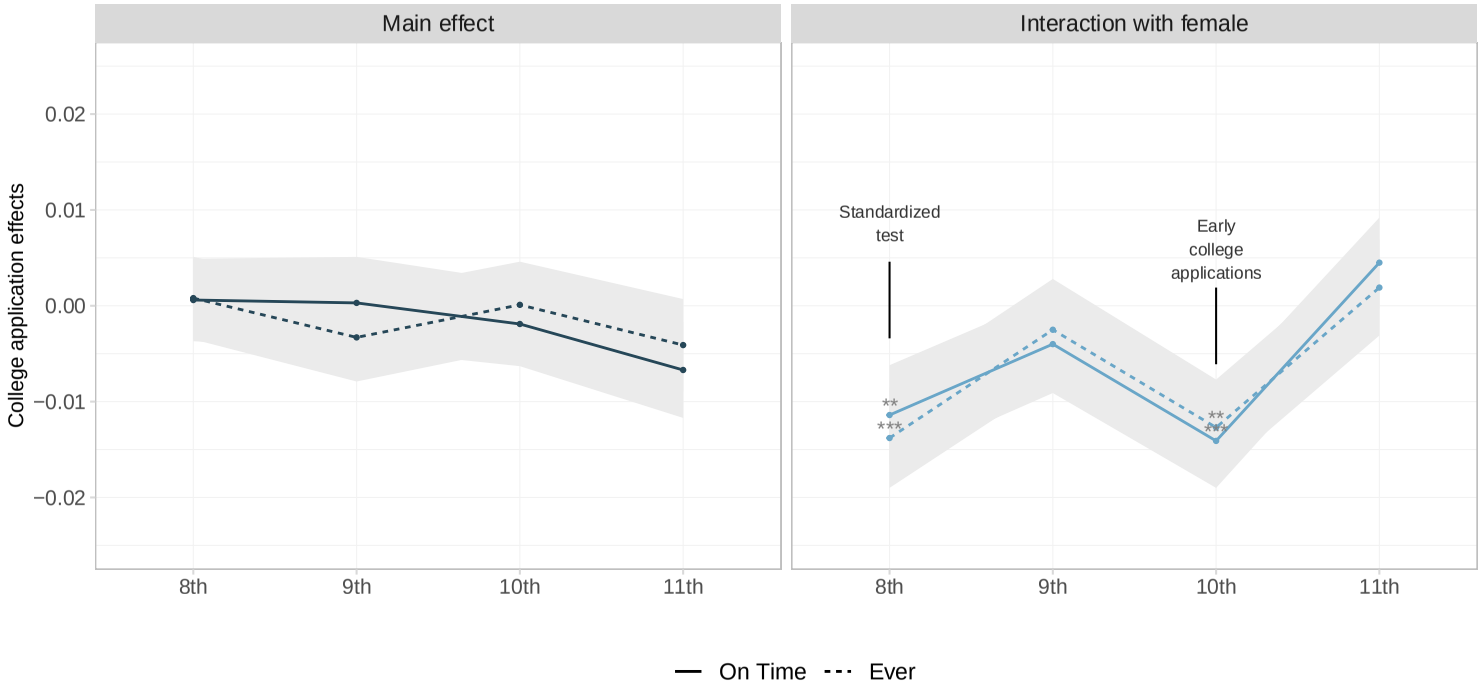


(a) Mathematics

(b) Language arts

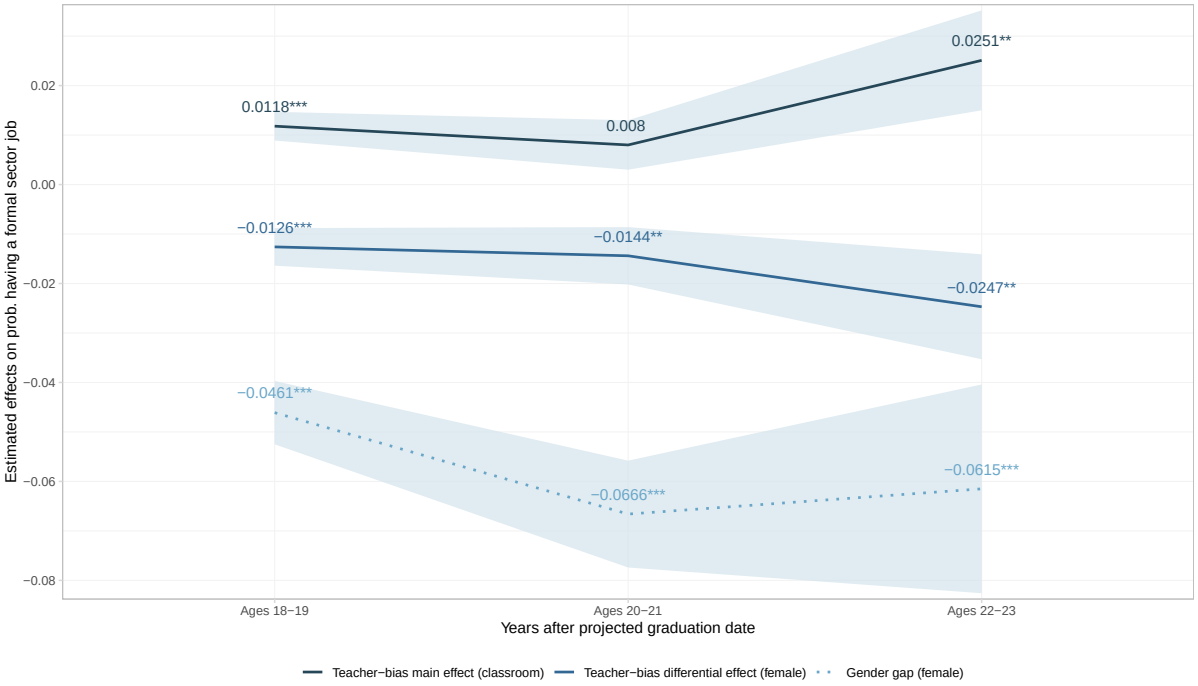
Notes: The figure presents the EB deconvolution estimates of the prior density of teacher-level grading bias and a benchmark Gaussian density. The histogram shows the distribution of estimated teacher-level grading bias. The red line density is the prior density of underlying population parameters estimated by the EB deconvolution method proposed in Efron (2016). This density was computed numerically, maximizing the respective (penalized) log-likelihood function using the deconvolveR package developed by Narasimhan and Efron (2020). The complexity penalty parameter was chosen for the deconvolved density to match the mean and bias-corrected variance of the estimated teacher-level grading bias. The parameter support was set to be a finite discrete of points between -2 and 1 for mathematics and language arts. The grey bars correspond to ± one estimated standard error. The black line density corresponds to the Gaussian density with the mean and variance of the posterior deconvolved distribution.

Figure 6: Effects on college application per grade



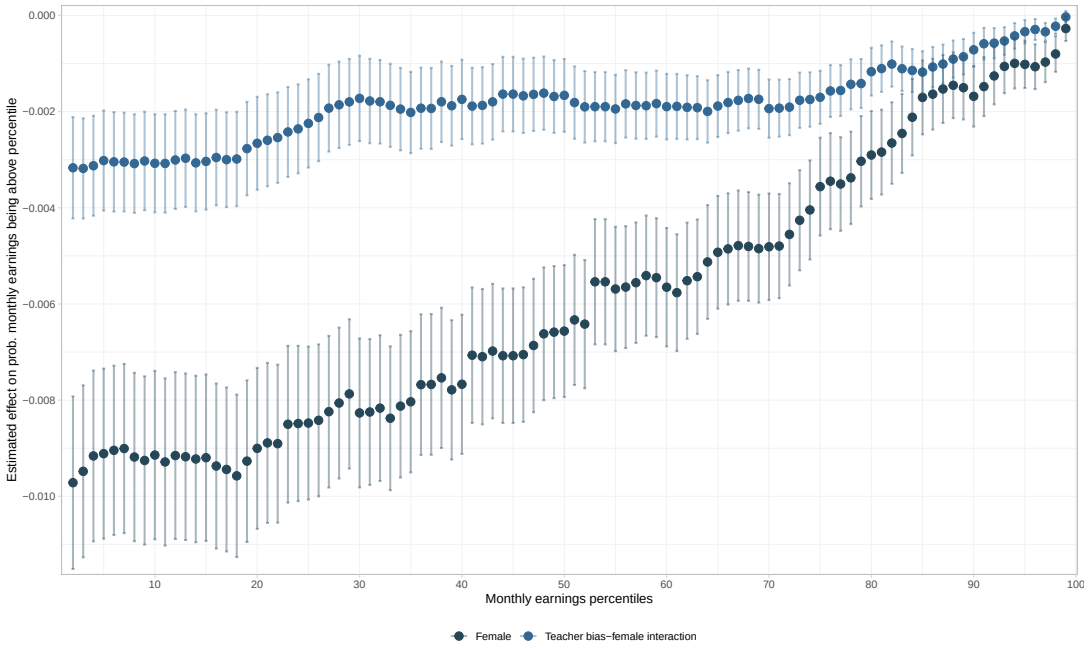
Notes: This figure shows the estimated effects of teacher' gender bias in assessment on the likelihood of applying to college on time and ever according to the estimating Equation (12). Panel A reports point estimates of the main effects of exposure to a one-standard-deviation more biased teacher against girls. Panel B reports the interaction effects of exposure with students being female. Clustered standard errors at the school level interacted are shown in shaded regions.

Figure 7: Effects on formal sector employment, ages 18–23

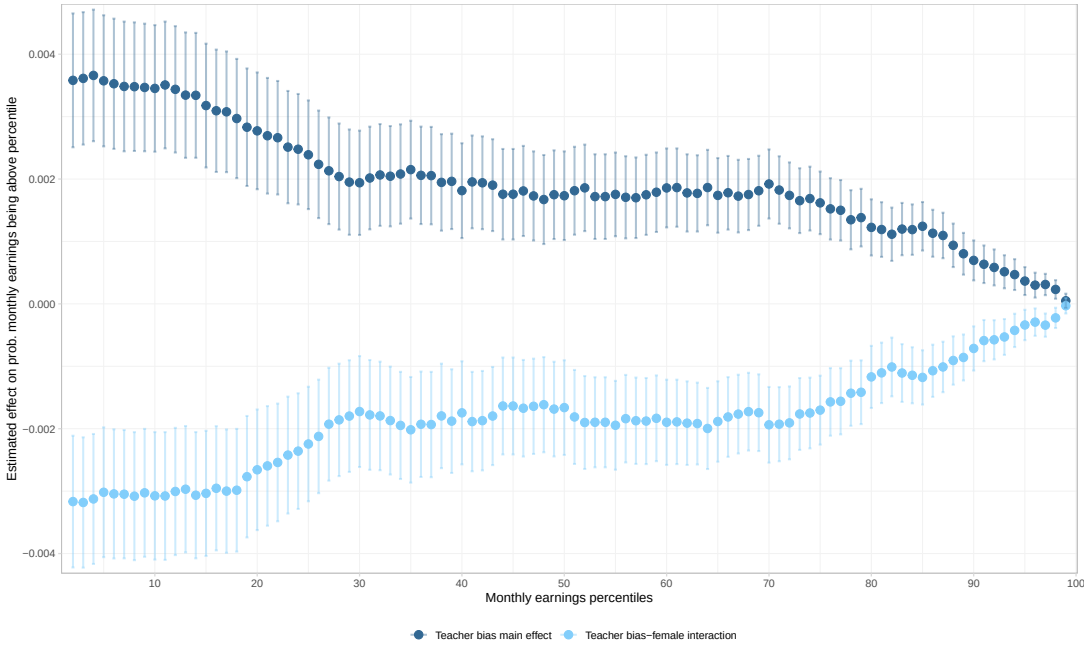


Notes: This figure shows the estimated effects of teacher' gender bias in assessment on the likelihood of having a formal sector job at ages 18–23 after graduating from high school. The point estimates of the main effect of exposure to more biased teachers are shown in the top solid line. The second solid line shows point estimates of the differential effects on women for being exposed to a more biased teacher. The dotted line shows the effects on women on the likelihood of employment (that is, the gender gap). I include controls according to estimating Equation (12). Clustered standard errors at the school location level, interacting with student gender, are shown in shaded regions.

Figure 8: Effects of teachers' bias in assessment on probability of students' earnings being above a percentile after graduating high school, ages 18–19



(a) Interaction effects and effects on gender-gap of teacher-level grading bias



(b) Interaction and main teacher-level grading bias effects

Notes: This graph shows the estimated effects on the probability of monthly earnings of student i being above a given percentile p where $p \in (0, 100)$. The estimates are according to Equation (12), and the percentiles have been computed using the population of workers from the base sample who have between 16 and 25 years old.

Table 1: Descriptive statistics for eighth-grade students with teacher-assessed and blindly graded test scores

	Female (1)	Male (2)	Total (3)
A. Demographic characteristics			
Age, years	12.750	12.873	12.812
Spanish	0.589	0.586	0.587
Indigenous language (Quechua)	0.076	0.079	0.078
Other	0.017	0.018	0.017
Missing language	0.318	0.317	0.318
Family receives CCT	0.231	0.246	0.239
Mother main caregiver	0.688	0.672	0.680
Father main caregiver	0.291	0.308	0.300
Other main caregiver	0.021	0.020	0.020
B. School attributes			
School in Lima	0.223	0.211	0.217
School in urban area	0.869	0.857	0.863
C. Students' educational aspirations			
HS	0.058	0.075	0.067
Technical	0.059	0.109	0.084
College	0.287	0.292	0.290
Graduate	0.259	0.182	0.220
Missing	0.337	0.342	0.339
Num. of students	690,449	696,800	1,387,249
Num. of schools			9,358
Num. of teachers			61,419
Observations			3,513,562

Notes: This table reports summary statistics on eighth-grade students who attended schools during 2015, 2016, 2018, and 2019. Students in the sample have available information on the standardized examination and teacher-assigned test scores in mathematics, language arts, and science. In addition, pupils in the sample have been successfully matched with their respective subject teachers. The bottom rows report the number of schools the students in the sample attend and the number of teachers matched to these students. Observations are at the student-subject level.

Table 2: Descriptive statistics for 8th- to 11th-grade of high school teacher-student matched sample

	Base sample			Full regression sample		
	Male (1)	Female (2)	Total (3)	Male (4)	Female (5)	Total (6)
A. Demographic characteristics						
Spanish	0.871	0.875	0.873	0.841	0.842	0.842
Indigenous language (Quechua)	0.104	0.100	0.102	0.130	0.128	0.129
Other language	0.024	0.025	0.025	0.028	0.030	0.029
Born in Lima	0.239	0.246	0.243	0.195	0.201	0.198
Mother is main caregiver	0.702	0.720	0.711	0.682	0.698	0.689
B. Students' senior year						
2015	0.062	0.066	0.064	0.046	0.049	0.047
2016	0.133	0.140	0.136	0.105	0.110	0.108
2017	0.204	0.212	0.208	0.170	0.175	0.173
2018	0.224	0.227	0.226	0.254	0.261	0.258
2019	0.239	0.240	0.239	0.278	0.283	0.280
C. Student achievement						
Mathematics	-0.116	0.042	-0.039	-0.163	-0.009	-0.089
Language arts	-0.224	0.148	-0.042	-0.278	0.086	-0.105
Science	-0.171	0.101	-0.038	-0.210	0.061	-0.080
Observations	7,876,733	7,508,521	15,385,254	4,169,532	3,816,044	7,985,576
Num. of students	1,083,084	1,048,483	2,131,560	890,078	846,049	1,736,120
Num. of teachers			122,591			53,975

Notes: This table reports summary statistics on 8th- to 11th-grade students enrolled in school in 2015-2019 with available end-of-grade scores for a previous school year. Columns 1 to 3 describe the *base* sample, which includes all students successfully matched with a teacher who instructs mathematics, language arts, or science. Columns 4 to 6 report statistics on the *full* regression sample, which excludes students in the base sample whose teachers' have no information on their bias measure. Report card scores measure student achievement; they have been standardized per year and subject, using the respective samples of students. The observations are at the student-grade-subject level.

Table 3: Descriptive statistics for students after high school graduation: labor market characteristics

	Base sample			Workers in regression sample	
	Female (1)	Male (2)	Non reported (3)	Female (4)	Male (5)
A. Average earnings and hours, 2015-2020					
Monthly earnings (2010 USD)	218.113	254.862	321.517	218.702	255.303
Paid work hours per month	40.940	46.902	65.518	41.036	46.791
Hourly wage (2010 USD)	2.759	2.979	3.421	2.765	2.986
B. Worker characteristics					
Worker age, years	19.202	19.244	19.437	19.209	19.246
Special education	0.000	0.000	0.000	0.001	0.001
Less than HS	0.046	0.064	0.163	0.045	0.063
HS graduate	0.621	0.691	0.779	0.616	0.687
Some college or technical	0.195	0.118	0.039	0.194	0.118
Technical degree	0.051	0.046	0.019	0.056	0.049
No information	0.088	0.082	0.000	0.085	0.079
C. NAICS industry					
Firm size (num. workers per firm)	3,850.377	3,185.798	2,334.833	3,809.744	3,173.817
11 agriculture, forestry, fishing	0.218	0.224	0.270	0.216	0.223
21-23 mining, utilities, construction	0.017	0.082	0.073	0.017	0.082
31-33 manufacturing	0.180	0.179	0.253	0.178	0.178
42-49 trade, transportation	0.222	0.213	0.150	0.222	0.213
51-59 information, finance, prof. serv.	0.183	0.167	0.124	0.184	0.167
61-62 educational and health care serv.	0.019	0.008	0.006	0.020	0.009
71-72 arts, recreation, hospitality serv.	0.099	0.074	0.058	0.100	0.074
81 other services	0.057	0.046	0.060	0.057	0.046
92-99 public admin., unclas.	0.005	0.007	0.004	0.005	0.007
Number of employees	176,963	281,511	2,014,575	187,919	297,087
Number of firms	30,214	45,798	190,055	31,990	47,667
Observations	176,963	281,511	2,014,575	187,919	297,087

Notes: This table shows descriptive statistics for comparing characteristics of recently graduated workers and students constituting the high school estimation sample with available matched employer-employee records. Students appearing in these records were reported by an employer to have a formal sector job corresponding to January 2015 and December 2020. Columns 1 to 3 show means for workers in the base sample, comprising workers between 16 and 25 years old. This age range retrieves comparable work experience to cohorts of recently graduated high school students in the regression sample. Workers in the base sample also comply with all exclusion filters described in the Online Appendix Section A.2.1. Columns 4 and 5 show means for students with projected graduation years 2015-20 who (i) are part of the estimation sample and thus have been matched with public high school student records, and (ii) appear as being employed for at least one month over the period 2015-2020 in administrative employment records. Panel A reports the means of monthly earnings, paid working hours, and hourly wages that have been previously processed to reflect the compensations from the work contact with the *dominant annual employer* as defined in Online Appendix Section A.2.1. Panels B and C report workers' characteristics and occupation categories as recorded by the most recent work contact with a *dominant annual employer*.

Table 4: Descriptive statistics for surveyed teachers with implicit-stereotypes measure

	Mathematics (1)	Language arts (2)	Total (3)
A. Demographic characteristics			
Female	0.454	0.543	0.495
Age, years	42.977	41.540	42.312
Mixed	0.650	0.660	0.654
Indigenous language (Quechua)	0.227	0.204	0.216
White	0.022	0.019	0.020
Asian	0.003	0.002	0.002
Afroperuvian	0.031	0.025	0.028
Has at least one daughter	0.532	0.482	0.509
Has no daughters	0.257	0.245	0.251
Has neither daughter nor sons	0.211	0.273	0.239
B. Teaching appointment			
School in Lima	0.240	0.253	0.246
Teaching hours per week	28.438	27.811	28.148
Number of students per class	24.222	24.983	24.574
Classoom teacher	0.606	0.581	0.595
First grade	0.532	0.563	0.546
Second grade	0.562	0.590	0.575
Third grade	0.531	0.649	0.585
Fourth grade	0.554	0.627	0.587
Fifth grade	0.552	0.586	0.568
C. Job characteristics			
Tenure	0.430	0.401	0.417
Contract	0.518	0.554	0.535
No contract	0.018	0.019	0.019
Other contract	0.034	0.026	0.030
Public sch experience, years	12.635	11.382	12.055
Private sch experience, years	2.647	3.030	2.824
College major, education	0.851	0.914	0.880
College major, other	0.149	0.086	0.120
College	0.644	0.655	0.649
Technical institution	0.447	0.419	0.434
Other institution	0.015	0.013	0.014
D. Discrimination and IAT			
Experienced discrimination	0.164	0.185	0.173
Witnessed discrimination	0.175	0.190	0.182
Gender-science stereotype (IAT score)	0.301	0.281	0.292
Gender-career stereotype (IAT score)	0.265	0.270	0.267
Observations	1,102	950	2,052

Notes: This table reports descriptive statistics for the analysis sample of surveyed high school teachers (grades 7 to 11) matched with assessment bias. Teachers' age and teaching experiences were reported between September 2021 and July 2020. Teaching appointment and job characteristics reported correspond to the years 2021 and 2022. Experience and witnessing discrimination is a self-reported measure of discriminatory behaviors by colleagues or principals in the current school year 2021-2022 due to race, immigrant status, socioeconomic level, gender, sexual identity, and religion. Raw (non-standardized) IAT scores are reported. Teachers have a *strong bias* if raw IAT score is above .65, *moderate to severe bias* when the score is between .65 and .35, *slight bias* if the score is between .15 and .35, *little to no bias* if the score is between -.15 and .15, and *preference for girls* if the score is below -.15.

Table 5: Variation in teacher-level bias in assessment estimates

	Teachers' subjects		
	Mathematics (1)	Language arts (2)	Science (3)
Unadjusted mean	-0.2978	-0.3069	-0.4119
Bias corrected standard deviation			
Unweighted	0.0612	0.0706	0.0880
Student-weighted	0.0973	0.1106	0.1576
Num. of teachers	21,094	21,078	13,482

Notes: This table presents the estimated means and standard deviations of the estimated teacher-level bias in assessment $\hat{\theta}_j$, calculated from estimates of Equations (1) and (2). The first row reports the mean, $\hat{\mu}$, of the unadjusted $\hat{\theta}_j$. The second and third rows display bias-corrected variance estimates that use the associated standard errors s_j to correct for bias due to sampling error in $\hat{\theta}_j$. The second row reports an *unweighted* bias-corrected variance, $\hat{\sigma}_U$, and the third row presents the bias-corrected *weighted* variance, $\hat{\sigma}_W$, where the sampling error was computed with student weights are $w_j = \mathcal{N}_j/\mathcal{N}$ with $\mathcal{N}(j)$ equivalent to the number of students assigned to teacher j and $\mathcal{N} = \sum_j \mathcal{N}(j)$. Observations are at the teacher level. Standard errors were computed using clusters at the student level.

Table 6: Relationship between assessment bias and mathematics teachers' characteristics

	(1)	(2)	(3)
A. Demographic characteristics			
Female	0.3269***	0.3286***	0.4455***
	(0.043)	(0.043)	(0.0789)
Age, older than median	-0.2312***	-0.2523***	-0.086
	(0.0497)	(0.0518)	(0.093)
Higher ed., university	-0.1503**	-0.0835	-0.0808
	(0.0652)	(0.0669)	(0.0813)
B. Teaching experience in private schools, years			
Less than 2 yrs		-0.3688***	-0.4137***
		(0.0928)	(0.1099)
2-5 yrs		-0.1749**	-0.2173**
		(0.0816)	(0.0967)
6-10 yrs		-0.2806**	-0.3658***
		(0.1153)	(0.1415)
More than 10 yrs		-0.4032**	-0.5239**
		(0.1681)	(0.2034)
C. Teaching experience in public schools, years			
Less than 2 yrs		-0.2994	-0.1336
		(0.2584)	(0.5585)
2-5 yrs		-0.4028*	-0.2201
		(0.2326)	(0.5278)
6-10 yrs		-0.191	-0.0855
		(0.2326)	(0.5279)
More than 10 yrs		-0.2262	-0.1952
		(0.2377)	(0.531)
D. Teacher evaluation performance			
Skill and knowledge test, z score			-0.1473**
			(0.0586)
Passed test			-0.0616
			(0.1455)
R-squared	0.0057	0.0058	0.0124
Observations	20,411	20,411	7,161

Notes: This table reports the estimated relationship between mathematics teachers' grading bias and a set of covariates. The reported coefficients were estimated according to Equation 4, and the outcome variable has been divided by the bias-corrected student-weighted standard deviation. The median age of teachers is 45 years old in the sample as of 2019. Higher education indicators have as a base category *Technical Institute*. *Teaching experience in public schools* and *Teaching experience in private schools* are measured as of 2019 based on personnel records from the Ministry of Education. The base category for the experience variables is *No experience*. Teacher-evaluation performance covariates are drawn from the National Teacher Evaluations 2015, 2017, 2018, and 2019. *Skills and knowledge test z score* is the standardized score of the national-stage evaluation of knowledge and teaching skills designed to test teachers applying to permanent public-school teaching positions. *Passing status* indicates when the teacher has approved the national-stage exam in any of the years mentioned above. Observations are at the teacher level. I include school-fixed effects and missing-value dummies where the information is unavailable. Standard errors are clustered at the school level.

Table 7: Relationship between assessment bias and implicit gender stereotypes

	Mathematics teachers	Language arts teachers
	Gender-science	Gender-science
	(1)	(2)
	A. No controls	
IAT score	0.1682***	-0.4943***
	(0.0489)	(0.045)
R-squared	0.035	0.0882
Observations	1,102	949
	B. With covariates	
IAT score	0.1279**	-0.5138***
	(0.0501)	(0.045)
R-squared	0.1127	0.1272
Observations	1,102	949

Notes: This table reports regression estimates of implicit stereotypes measured with Implicit Association Test (IAT) scores on mathematics and language arts teachers' assessment bias estimated according to Equation (5). The IAT score has been standardized to have a mean of 0 and a standard deviation of 1. Column 1 shows the effect of increasing math teachers' implicit gender bias against girls (as measured by the IAT score) on their grading bias in mathematics by one standard deviation. Column 2 reports similar regression effects of increasing by one standard deviation language arts teachers' implicit gender bias against girls on their grading bias in language arts. In Column 1, the dependent variable was divided by the bias-corrected standard deviation of $\hat{\theta}_j$ for math teachers, while Column 2 employs a similar procedure for language arts teachers. Therefore, the reported coefficients are expressed in standard-deviation units. The first panel reports estimates including only school-location fixed effects. The second panel adds control for teachers' children categories (i.e., daughters, sons, or both), indicators of the decades of teachers' date of birth, teachers' race, the order of the associations in the IAT (i.e., order compatible or order incompatible appearing first), and the number of previous IATs the teacher has taken. Observations are at the teacher level, and clustered robust standard errors by school location are reported in parentheses.

Table 8: High school graduation effects of exposure to mathematics-teacher assessment bias

	Graduated on time				Graduated ever			
	8th grade (1)	9th grade (2)	10th grade (3)	Stacked grades (4)	8th grade (5)	9th grade (6)	10th grade (7)	Stacked grades (8)
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.0364***	-0.0173***	-0.0147***	-0.0151***	-0.0387***	-0.0135***	-0.0136***	-0.0133***
	(0.0054)	(0.0045)	(0.0037)	(0.0023)	(0.0053)	(0.0043)	(0.0034)	(0.0022)
$\hat{\theta}^*_{j,-i}$	0.0208***	0.0029	0.002	0.0051*	0.0227***	0.0028	0.0021	0.006**
	(0.0059)	(0.0061)	(0.0051)	(0.0027)	(0.0057)	(0.006)	(0.005)	(0.0028)
Female	-0.0054	-0.0226**	-0.0097	-0.0089**	-0.0197**	-0.0376***	-0.0204***	-0.0191***
	(0.0103)	(0.0088)	(0.0069)	(0.0041)	(0.01)	(0.0086)	(0.0065)	(0.0041)
Baseline controls	×	×	×	×	×	×	×	×
Lagged scores	×	×	×	×	×	×	×	×
$\bar{Y}$ female	0.7513	0.7883	0.8572	0.8239	0.782	0.8249	0.8933	0.8544
$\bar{Y}$ male	0.6833	0.7327	0.8105	0.7741	0.7434	0.8019	0.8784	0.8318
R-squared	0.3271	0.364	0.4275	0.4189	0.2674	0.2871	0.2897	0.2831
Observations	629,489	667,250	765,620	2,905,775	629,489	667,250	765,620	2,905,775

Notes: This table reports estimates of the effects of one-grade exposure to a mathematics teacher's gender assessment-bias on high school graduation on time and ever. The estimations use the full regression sample of students between grades 8 and 10 with projected graduation years between 2015 and 2019. The teacher-level bias in assessment is a leave-one-year-out estimate assigned to each student cohort based on their projected graduation year. The outcome variable in Columns 1 to 4 is an indicator variable taking value one when the student graduated in the school year corresponding to their projected high school graduation date. According to enrollment records, the projected graduation year was computed based on the earliest year the student was enrolled in high school. Columns 5 to 8 have as outcome variables an indicator for graduation at any time. The window of time over which graduating from high school *ever* is defined varies cohort-to-cohort, ranging from one for the youngest cohort (class of 2019) to four calendar years after projected graduation for the oldest cohort (class of 2015). Covariates include baseline student-level characteristics and teacher attributes interacted with student-gender, classroom and school-level controls, and quadratic polynomials of lagged mathematics and language arts scores. The observation unit in grade-specific estimations (Columns 1 to 3 and Columns 5 to 7) is the student; this specification includes cohort, year, and school fixed effects and clustered standard errors at the school level. The observation unit in all-grades estimation (Columns 4 and 8) is the student grade; this specification uses a sample of students stacked across grades and includes grade, cohort, year, and school fixed effects, as well as two-way clustered standard errors by student and school.

Table 9: College-attendance effects of exposure to mathematics-teacher assessment bias

	College application		College admission		College enrollment	
	On time	Ever	On time	Ever	On time	Ever
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.0066**	-0.0073**	-0.0011	-0.001	-0.0006	-0.0011
	(0.0032)	(0.0032)	(0.0026)	(0.0027)	(0.0026)	(0.0026)
$\hat{\theta}^*_{j,-i}$	0.0007	0.0015	-0.0028	-0.0028	-0.0031	-0.0032
	(0.0029)	(0.0027)	(0.0022)	(0.0022)	(0.0022)	(0.0022)
Female	0.0161***	0.0076	0.0063	0.0042	0.0056	0.0051
	(0.0054)	(0.0056)	(0.0044)	(0.0045)	(0.0042)	(0.0042)
Baseline controls	×	×	×	×	×	×
Lagged scores	×	×	×	×	×	×
$\bar{Y}$ female	0.2438	0.3334	0.1987	0.2077	0.1806	0.1843
$\bar{Y}$ male	0.1944	0.281	0.1644	0.1735	0.151	0.1538
R-squared	0.1048	0.1323	0.0869	0.0955	0.0901	0.0928
Observations	2,905,775	2,905,775	2,905,775	2,905,775	2,905,775	2,905,775

Notes: This table reports estimates of the effects of one-grade exposure to a mathematics teacher's gender bias in assessment on college-attendance outcomes. The estimation uses a stacked sample of students between grades 8 and 11 (13 to 16 years old) with projected graduation years between 2015 and 2019 that are part of the full regression sample described in Section 3.2.2. The estimation controls for the set of covariates are as described in the notes of Table 8 and include grade, cohort, year, and school fixed effects. Two-way clustered standard errors by student and school are reported in parentheses.

Table 10: Test of student sorting to teachers based on observed students' characteristics

	Spanish-speaking student	Mother college-education degree	Mother graduate degree	Student birthplace Lima	Teacher contract, fixed term
	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}^*_{j,-i} \cdot$ Female	0.0021	-0.0003	0.0001	0.0007	-0.0027
	(0.0014)	(0.0012)	(0.0002)	(0.0018)	0.0056
$\hat{\theta}^*_{j,-i}$	0.0002	0.0005	0	-0.0007	0.0115
	(0.0008)	(0.0009)	(0.0002)	(0.0014)	0.0123
Baseline controls	×	×	×	×	×
Lagged scores	×	×	×	×	×
R-squared	0.7422	0.1484	0.0172	0.5487	0.4951
Observations	2,547,760	2,547,760	2,547,760	2,547,760	2,547,760

Notes: This table reports estimates of the effects of teacher-level grading bias on student and teacher characteristics, which are potential sources of student sorting. Baseline controls include all covariates detailed in the notes of Table 8, except for the variable included as the dependent variable in the corresponding column. Lagged scores are also included as controls and fixed effects on grade, cohort, year, and school. Two-way standard errors are clustered at the student and teacher levels.

Table 11: Bias-corrected correlation matrix of parameters of teacher value-added and gender bias in assessment

		Teacher VA parameters	
	Assessment bias (1)	VA towards females (2)	VA towards males (3)
	A. Mathematics		
Assessment bias	1	-0.1157	-0.2143
VA toward females		1	0.9884
VA toward males			1
Standard deviation	0.1160	0.5758	0.5986
	B. Language arts		
Assessment bias	1	-0.3642	-0.6577
VA toward females		1	0.8673
VA toward males			1
Standard deviation	0.1240	0.1259	0.1793

Notes: This table reports the correlation matrix between teacher-level assessment bias and value-added parameters. Numbers in brackets are standard deviations of the parameters reported in each column. I define the assessment bias $\theta_j = \beta_{2,j} - \alpha_{2,j}$, teachers' value-added for males $\alpha_{1,j} + \alpha_{2,j}$, and value-added for females $\alpha_{1,j}$. The parameters were estimated using a fixed-effects-SURE specification of according to Equation (1) and Equation (2). The variance-covariance matrix of joint parameters was adjusted for sampling error using SURE-calculated sampling covariances. The estimating equations control for quadratic polynomial lagged scores in language and math, student-level, and teacher-level baseline controls, classroom-grade and school-grade means of baseline covariates, and cohort, grade, school, and year fixed effects.

Table 12: Employment in the formal sector: effects of exposure to mathematics teachers' assessment bias

	Employed in formal sector		
	Ages 18–19 (1)	Ages 20–21 (2)	Ages 22–23 (3)
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.0126***	-0.0144**	-0.0247**
	(0.0038)	(0.0058)	(0.0106)
$\hat{\theta}^*_{j,-i}$	0.0118***	0.008	0.0251**
	(0.0029)	(0.005)	(0.0101)
Female	-0.0461***	-0.0666***	-0.0615***
	(0.0064)	(0.0108)	(0.0211)
Baseline controls	×	×	×
Lagged scores	×	×	×
$\bar{Y}$ female	0.0694	0.0963	0.1153
$\bar{Y}$ male	0.1154	0.151	0.1743
Observations	884,231	395,940	118,550

Notes: This table reports estimated coefficients of the effects of exposure to more biased teachers during one grade in high school on the probability of having a formal sector job at ages 18 to 23 after high school graduation. The sample includes students in the *full regression sample* with projected graduation between 2015 and 2019. The outcome variable is the likelihood of having a formal sector job with a *dominant annual employer* for at least one month in the year t after graduating from high school when students' modal ages are 18–19, 20–21, and 22–23. See Appendix A for details on this variable calculation. The treatment variable was computed as a leave-one-year-out version of teacher-level grading bias, divided by its respective standard deviation computed at the year-subject level. Covariates include student-level characteristics, teacher characteristics, lagged scores, classroom- and school-grade-level controls, and cohort, grade, and year fixed effects. Additionally, I controlled for school-location fixed effects are interacted with gender. The observation unit is the student grade. Clustered standard errors by the school are reported in parentheses.

Table 13: Paid hours work and monthly earnings: effects of exposure to mathematics teachers' assessment bias

	Paid monthly work hours			Monthly earnings (2010 dollars)		
	Ages 18–19 (1)	Ages 20–21 (2)	Ages 22–23 (3)	Ages 18–19 (4)	Ages 20–21 (5)	Ages 22–23 (6)
$\hat{\theta}^*_{j,-i} \cdot$ Female	-1.0624***	-0.8031	-1.6004	-2.6219***	-1.304	-3.0888
	(0.3679)	(0.5758)	(1.0464)	(0.7671)	(1.1849)	(2.2144)
$\hat{\theta}^*_{j,-i}$	1.0016***	0.2338	2.7815***	2.4911***	0.2876	5.281***
	(0.3041)	(0.5005)	(0.9513)	(0.6362)	(1.045)	(2.0422)
Female	-3.8758***	-6.1869***	-5.0676**	-8.2754***	-11.181***	-11.0696**
	(0.6472)	(1.1269)	(2.2476)	(1.3323)	(2.3933)	(4.6696)
Baseline controls	×	×	×	×	×	×
Lagged scores	×	×	×	×	×	×
$\bar{Y}$ female cond.	96.6	98.3	98.5	212.7	220.7	223.7
$\bar{Y}$ male cond.	103.4	103.4	102.2	247.3	251.3	251.5
$\bar{Y}$ female uncond.	5.5	7.7	8.6	9.8	14.2	16.6
$\bar{Y}$ male uncond.	9.8	12.8	13.8	18.6	24.8	27.8
Observations	884,231	395,940	118,550	884,231	395,940	118,550

Notes: This table reports estimated coefficients of the effects of exposure to more biased teachers during one grade in high school on paid monthly work hours and monthly earnings at ages 18 to 23 after high school graduation. The sample includes students in the *full regression sample* with projected graduation between 2015 and 2019. In Columns 1 to 3 the outcome variable is the number of work hours paid per month in the contract with the dominant annual employer, and in Columns 4 to 6 the outcome variable is monthly earnings measured in 2010 USD obtained in the contract with the dominant annual employer. See Appendix A for details on this variable calculation. The treatment variable was computed as a leave-one-year-out version of teacher-level grading bias, divided by its respective standard deviation computed at the year-subject level. Covariates include student-level characteristics, teacher characteristics, lagged scores, classroom- and school-grade-level controls, and cohort, grade, and year fixed effects. Additionally, I controlled for school-location fixed effects are interacted with gender. The observation unit is the student grade. Clustered standard errors by the school are reported in parentheses.

Table 14: Mechanisms analysis of internalized gender bias among students as mediator of labor outcomes

	Students gender-science IAT Score	
	Mathematics	Language arts
	(1)	(2)
$\hat{\theta}^{*}_{j,-i}\cdot$ Female	-0.1496	-0.0262
	(0.1898)	(0.1377)
$\hat{\theta}^{*}_{j,-i}\cdot$	0.1515	-0.0146
	(0.1737)	(0.1167)
Female	0.0331	-0.7631**
	(0.3567)	(0.3634)
Baseline controls	×	×
Lagged scores	×	×
Mean IAT score female	0.2967	0.2919
Mean IAT score male	0.3212	0.3176
R-squared	0.1187	0.114
Observations	4,640	4,666

Notes: This table reports the relationship between a one-standard-deviation increase in teachers' grading bias against girls and their students' Implicit Association Test (IAT). Students in the estimation sample have a projected graduation year from high school between 2020 and 2025. The students' IAT was standardized with a mean of 0 and a standard deviation of 1. The controls include all controls from the main research design for identifying the impacts of grading bias on long-term outcomes. Observations are at the student-grade level, and standard errors are clustered at the school level.