I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17674

The Long-Term Effects of Teachers' Gender Stereotypes on Labor Outcomes

Joan J. Martínez

FEBRUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17674

The Long-Term Effects of Teachers' Gender Stereotypes on Labor Outcomes

Joan J. Martínez
University of California Berkeley and IZA

FEBRUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Long-Term Effects of Teachers' Gender Stereotypes on Labor Outcomes[*]

Teachers' stereotypical assessments widen the gender gap in earnings and formal sector employment after high school graduation, with lasting positive effects for men and shorter-term negative effects for women. Exposure to these assessments throughout high school disproportionately affects women's graduation, employment, working hours, and earnings during late adolescence and early adulthood. Implicit Association Test scores collected through a survey indicate that students from both genders internalize stereotypes about math and language skills. Stereotyped teachers also deter females from entering male-dominated occupations. I find no evidence that these assessments affect college application or enrollment outcomes for students, irrespective of gender.

JEL Classification: J16, J24, I24, J71

Keywords: gender stereotypes, gender pay gap, value-added, math

Corresponding author:
Joan J. Martínez
Haas School of Business
University of California Berkeley
110 Sproul Hall #5800
Berkeley
USA

E-mail: martinezp_jj@berkeley.edu

[*] I am thankful to Christopher Walters, Patrick Kline and Jesse Rothstein for their extensive support and guidance. I am thankful for comments from Desmond Ang, Sydnee Caldwell, David Card, Michela Carlana, Supreet Kaur, Camille Landais, Gianmarco León-Ciliotta, Lukas Leucht, Alan Manning, Edward Miguel, Enricco Moretti, Amanda Pallais, Camille Terrier, Damian Vergara, and four anonymous referees. I also thank seminar participants at the Boston University, Bocconi University, Brookings Institution, CESifo, Chicago Harris School of Public Policy, Harvard Kennedy School, London School of Economics, Nova School of Business and Economics, Ohio State University, Stanford PacDev, and UC Berkeley. I am thankful for institutional support and information access from the Ministry of Education and the Ministry of Labor of Peru. I thank Karen Espinoza, Ayrton Dextre, and Gabriela Wong for their invaluable support in data collection. I gratefully acknowledge financial support from the UC Berkeley Economics Department, UC Berkeley Opportunity Lab, Weiss Fund, the Institute for Research on Labor & Employment Dissertation Fellowship, and the Center for Effective Global Action. IRB approval was received from the University of California, Berkeley.

1 Introduction

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).[1] With fewer opportunities for secondary and tertiary education, women in developing countries rely heavily on labor market conditions for economic mobility, making job quality and equal access to employment critical. Recent evidence indicates that gender differences in salary requests (Rousille, 2021), self-promotion in teamwork (Coffman, 2014, Exley and Kessler, 2022), job-attribute preferences (Biasi and Sarsons, 2020, Fleche, Lepinteur and Powdthavee, 2018, Flory, Leibbrandt and List, 2015, Le Barbanchon, Rathelot and Roulet, 2020), and competitiveness (Niederle and Vesterlund, 2007) contribute to shaping gender inequality in employment and pay. I extend this literature by examining how stereotyped teacher assessments in high school contribute to labor market gender gaps, a relationship that remains unstudied.

A large body of research demonstrates that educators have a lasting impact on their students' careers long after they leave the classroom (Chetty, Friedman and Rockoff, 2014*a*,*b*, Rothstein, 2010, 2017). Research on stereotypes has found that teacher prejudices and stereotypes among educators discourage girls from pursuing science-focused high school tracks (Alesina et al., 2018, Carlana, 2019), increase gender score gaps (Lavy, 2008, Lavy and Sand, 2018, Terrier, 2020), and influence college decisions (Lavy and Megalokonomou, 2024). While these studies document educational effects, assessing their labor market impact is challenging due to debates over measuring stereotypes and limited data on students' professional trajectories in labor markets long after exposure to stereotyped teachers. I contribute to the literature by bridging the gap between research on the role of stereotypes and gender inequality in labor markets.

This paper studies the extent to which gender-stereotyped teacher assessments regarding students' abilities affect their labor market outcomes and educational careers. I use novel data from two sources: linked administrative records on 1.6 million students' academic and professional trajectories from ages 12 to 22, including their teachers' records, and nationwide survey-based stereotype measures from teachers and students in Peru's public high schools. Two measures, systematic assessment gaps between boys and girls and the Implicit Association Test (IAT), have been at the center of debate regarding their ability to measure stereotypes influencing students' assessments in math and science relative to communication and the humanities. The IAT gauges implicit preferences through response times (Greenwald, Nosek and Banaji, 2003, Greenwald, McGhee and Schwartz, 1998, Nosek, Greenwald and Banaji, 2007, Nosek et al., 2014), while the second measure utilizes students' test scores on two types of examinations (teacher-graded and blindly graded) (see for instance, Botelho, Madeira and Rangel (2015), Burgess and Greaves (2013), Carrell, Page and West (2010), Lavy (2008), Lavy

[1] According to reports by OECD (2022) and ILO (2020), the gender wage gap in OECD countries today is 11.7%, while in middle- and low-income countries, it remains higher at 15%–27%.

and Sand (2018)). A strength of this study is that I constructed both of these measures to document the extent to which systematic gender differences in assessments reflect ability stereotypes at the individual level.

The context of this study is Peru, where significant gender disparities exist in education and labor market outcomes. Low-income workers, particularly women, face challenges, with about one-quarter of women falling below the poverty line (World Bank Group, 2017). Moreover, official statistics reveal a 30% wage gap between men and women, even among relatively well-off women working in the formal sector (INEI, 2020, Ministry of Labor of Peru, 2019). Partnering with the Ministry of Education, I co-developed and utilized a government educational portal to collect linked survey responses and IAT scores from teachers and students in public schools nationwide. This new dataset, comprising 2,541 teacher and 1,153 student responses, allows for a comprehensive analysis of the effects and mechanisms of stereotyped evaluations. Additionally, I examine the impact on long-term outcomes for five student cohorts, combining education and employment records. Using a value-added framework, I compare students in the same cohort-grade-year school cells (controlling for lagged test scores) assigned to different teachers to understand the effects of stereotypical assessment practices.[2]

Teachers' gender-stereotyped assessments are estimated through a value-added approach, isolating stereotyped from non-stereotyped exam grading variations between boys and girls. Following previous studies (for instance, Lavy (2008), Lavy and Sand (2018), Lavy and Megalokonomou (2024), Terrier (2020)), a teacher-level measure for gender disparities in student evaluations considers differences between teacher-assigned and blindly graded assessments. Controlling for student-, teacher-, and class-specific covariates interacting with students' gender, this paper identifies teachers likely to assign higher grades to boys than girls (or vice versa) in classroom exams compared to blindly graded tests in math, science, and language arts. The remaining systematic gender differences are interpreted as reflective of teachers' stereotypical views. The hypothesis is tested using newly collected IAT data on a subsample of teachers with information on gender-based math stereotypes.

Teachers' individual gender stereotypes about students' ability robustly predict their evaluating behavior, as evidenced by gender gaps in assessments compared to centrally assigned scores. Mathematics teachers, with an average IAT score of 0.32, exhibit strong associations between boys and math and science and girls and the humanities. This suggests that girls in developing countries may encounter teachers with negative stereotypes about their math abilities.[3] The relationship between IAT scores and teacher stereotyped assessment estimates provides compelling evidence that the latter reflects

[2] The educational platform was created for this study with the approval of the Ministry of Education as part of a multiyear educational program. The platform, "Opportunities for Everyone", is accessible at http://www.oportunidadesparatodos.pe.

[3] The average implicit stereotypes in this setting surpass those of middle school teachers in Italy (0.09), according to Carlana (2019). Language arts teachers have an average IAT score of 0.31, aligning with Carlana's results.

gendered preconceptions about students' abilities in math and science versus language and the humanities. Teachers with stereotypical beliefs favor boys in math and science, assigning them higher grades than girls. Conversely, teachers favoring girls in the humanities give them higher language arts grades than boys.

Next, I assess the impact of stereotyped teacher assessments on formal sector employment, worked hours, and earnings. The research design compares students assigned to different teachers within cohort-grade-year-school cells, controlling for various factors. These include lagged test scores, demographics of students and teachers, classroom-level controls, and school-grade-level controls. The empirical strategy relies on the assumption that students' assignments to teachers are as good as random when considering demographic information and lagged test scores (see, for instance, Chetty, Friedman and Rockoff (2014*a*,*b*)). I focus on the impact of stereotypical grading by mathematics teachers, as math scores are shown to be more consequential for long-term outcomes, while supplementary results for language arts are in Online Appendix C. Additional analysis discusses high school dropout rates, teachers' value-added, and gender-based industry sorting as mediating mechanisms.

Boys benefit from long-lasting advantages in labor markets, while girls face detrimental, smaller, and short-lived effects. A female student's likelihood of full-time formal employment at age 18–19 decreases by 0.2 percentage points for every standard deviation that her math teacher's stereotyped assessments exceed the group mean. No detectable effects on girls' formal sector employment occur after this age. Boys experience positive effects over five years, with a 0.4–2.2 percentage point increase in their likelihood of formal sector employment (equivalent to between 4% and 12% of the mean). One grade level of increased exposure to stereotyped assessments for girls is associated with a USD 3.6–8.9 annual loss in earnings between ages 17 and 19, leading to an 8% increase in the gender pay gap in the first two years post-graduation. Women's earnings losses level off between the third and fifth year, while men see annual earnings gains ranging from USD 8.3 to USD 52.4 in the first five years after graduation, equivalent to 3.5% to 20% of the minimum wage per month.

Analyzing the impact of exposure to average stereotypical practices throughout high school, women with teachers exhibiting one-standard deviation higher stereotypical practices experienced an annual earnings loss of USD 8.4 at ages 18–19, two years post-graduation. Comparable effects for one-year and high school-average exposures suggest potential offsetting by different teachers in certain grades. The impact disproportionately affects the lower end of the earnings distribution, resulting in significant and unequal losses for disadvantaged females up to ages 18–19, with no clear pattern thereafter. The primary mechanism for widened gender gaps in earnings and employment is the hindrance of high school graduation. A one-year assignment to a high school math teacher with gender-stereotyped assessments reduce girls' graduation probability by 0.6–0.5 percentage points (0.7% and 0.6% of the mean), while boys' graduation likelihood increases by 0.9–0.8 percentage points (1.2% and 1% of the mean).

To the best of my knowledge, this study is the first to show that boys and girls in math classes are more likely to internalize gender stereotypes from their teachers, especially when taught by a math teacher with a one-standard-deviation higher IAT score. Analyzing the collected IAT scores, the study finds a negative impact on students' self-perceptions of math and science abilities for both genders. While survey evidence indicates there are no significant effects on reported beliefs about working in STEM fields, a separate analysis using administrative data of recent graduates indicates that teachers with gender-stereotyped views influence students' career choices, resulting in persistent industry sorting three years after high school graduation at age 19–20. The findings indicate that exposure to stronger stereotypes influences students' perceptions and career choices to some extent.

This analysis contributes to the literature on evaluators' biases and stereotypes impacting human capital decisions, productivity, and job performance. [4] Previous research utilized two main measures: assessment gaps from observational data (Lavy, 2008, Lavy and Sand, 2018, Lavy and Megalokonomou, 2024, Terrier, 2020) and IAT data (Greenwald, Nosek and Banaji, 2003, Greenwald et al., 2005, Greenwald, McGhee and Schwartz, 1998, Nosek, Greenwald and Banaji, 2007, Nosek et al., 2014, Rooth, 2010). Linking these two measures offers critical advantages compared to previous work. First, it addresses the limitations of score-based approaches, which may capture unobserved heterogeneity. This analysis provides direct estimates of the robust relationship between stereotyped grading (a behavioral outcome), favoring boys over girls, and implicit gender stereotypes in science-based fields. Second, I also address concerns about factors influencing teachers' grading beyond cognitive abilities, aiming to rule out alternative explanations for gender differences in test scores, such as students' behavior and ability proxied by past performance (Bertrand and Pan, 2013, Figlio et al., 2019, Fortin, Oreopoulos and Phipps, 2015, Jackson, 2018).

I also contribute to the literature on education and labor stereotypes by examining the impact of teacher stereotype assessments on long-term human capital decisions and previously unavailable outcomes. Previous research on teacher-student interactions highlights the negative effects of teachers' implicit biases on female students' achievement (Alan, Ertac and Mumcu, 2018, Alesina et al., 2018, Burgess and Greaves, 2013, Carlana, 2019, Lavy, 2008, Lavy and Sand, 2018, Terrier, 2020), high school track choices (Carlana, 2019), and college attendance (Carrell, Page and West, 2010, Lavy and Megalokonomou, 2024, Reuben, Sapienza and Zingales, 2014). Prejudices about workers' abilities and productivity are often considered in workplace decisions concerning performance and job assignments (Ewens and Townsend, 2020, Glover, Pallais and Pariente, 2017, Goldin and Rouse, 2000, Sarsons, H., 2017). This study extends this literature by showing

[4] Gender biases and stereotypes differ from a cognitive-psychology standpoint. Kahneman and Tversky (1973) proposes a formal definition of stereotypes using the probability-judgments approach, as outlined by Bordalo et al. (2016).

that stereotypes affect school performance and influence educational decisions that are relevant to employment.

This article builds on labor market gender gap studies, exploring factors like childbearing (Berniell et al., 2021, Correll, Benard and Paik, 2007, Hardy and Kagy, 2018, Kleven et al., 2019) and occupational sorting (Francesconi and Parey, 2018). Despite addressing skills and education, an unexplained wage gap persists. Recent research examines psychological and behavioral causes, including preferences for schedule flexibility (Fleche, Lepinteur and Powdthavee, 2018, Wiswall and Zafar, 2018), commuting time (Le Barbanchon, Rathelot and Roulet, 2020), and competitiveness (Flory, Leibbrandt and List, 2015, Gneezy, Niederle and Rustichini, 2003, Niederle and Vesterlund, 2007). This study contributes to understanding how high school students internalize teachers' stereotyped assessments, suggesting that they shape their views on numeric and communication abilities as measured by their IATs. This internalization leads to gender disparities in labor markets, revealing teachers' biases as a previously undocumented source of disparities before workers enter the workforce. Lastly, I build on recent discrimination literature in labor and education (Kline and Walters, 2021), employing improved techniques for assessing individual-level and distributional estimates of gender stereotypes in education settings.

The paper's structure is as follows. Section 2 presents a value-added framework outlining the effects and mechanisms of teachers' stereotypical assessments. Section 3 introduces the empirical setting and data. Section 4 outlines the strategy for measuring gender differences in grading and their correlation with teachers' IAT scores. Methodology and results for assessing the long-term consequences of stereotyped assessments are discussed in Section 5. Section 6 examines mechanisms, and Section 7 concludes. Online Appendix A provides additional details on sample construction, outcomes, and survey design.

2 Value-Added Framework of Teachers' Stereotyped Assessments

This section introduces a student assessment model within a value-added framework, considering teachers' stereotypes and preconceived notions of students' abilities. Academic performance is assessed by two entities: one, influenced by the teacher's awareness of the student's group, and the other, conducted by external evaluators unaware of group affiliation.

2.1 Setup

Students ($i \in \mathcal{N}$) are assigned to teachers ($j \in \mathcal{J}$). In the benchmark case, groups are denoted as $h(i) \in \{h, h'\}$, with $G_i = 1$ for group h' and 0 otherwise. Each student has an unobserved innate ability ψ_i with identical ability distributions for each group. Students receive centrally assigned scores S_{ij}^B and teacher-assigned scores S_{ij}^T based on academic performance. All teachers harbor, to some extent, group stereotypes about the distribution of students' abilities. Students are unaware of these stereotypes, but they have an impact on their performance evaluation.

Centrally assigned scores are denoted as $\tilde{S}^B_{ij} = s_j(G_i, W_{ij}, \psi_i, \eta'_{ij})$, allowing other group-specific factors to influence test scores. Here, W_{ij} represents student and teacher covariates. The unobserved variable $\vartheta_{h(i),j}$ reflects the teacher's stereotypes about a student's actual ability based on group membership —that is, their ability stereotypes. Teacher-assigned scores are defined as $\tilde{S}^T_{ij} = c_i(G_i, W_{ij}, \psi_i, \vartheta_{h(i),j}, \epsilon'_{ij})$. Both η'_{ij} and ϵ'_{ij} are unobserved determinants of scores, assumed to have zero means and be identically distributed and pairwise independent for any pair (i, j). An additive structure is assumed for the scores.

$$E[S^B_{ij}|G_i, W_{ij}, \psi_i] = \alpha + \alpha_{g,j}G_i + W'_{ij}\alpha_{w,j} + \alpha_{\psi,j}\psi_i$$
$$E[S^T_{ij}|G_i, W_{ij}, \tilde{\psi}_{ij}] = \beta + \beta_{g,j}G_i + W'_{ij}\beta_{w,j} + \beta_{\psi,j}\tilde{\psi}_{ij}$$

The unobserved variable $\tilde{\psi}_{ij}$ captures teacher j's forecast of their student i's ability based on the student's observed group membership, with higher values indicating greater predicted ability, and $\lambda_{\psi,j}, \lambda_{\vartheta,j} \in (0, 1)$. The ability of student i of teacher j is forecast as,

$$\tilde{\psi}_{ij} = \lambda_j + \lambda_{\psi,j}\psi_i + \lambda_{\vartheta,j}\vartheta_{h(i),j} \tag{1}$$

Definition 1 (Ability stereotypes in teacher evaluations)**.** *Let $\underline{F}$ and $\bar{F}$ be continuous cumulative distribution functions of $\vartheta_{h,j}$ and $\vartheta_{h',j}$, respectively. A teacher holds detrimental stereotypes about group h and optimistic stereotypes about group h' iff $\underline{F}$ is first order stochastically dominated by $\bar{F}$. That is, $\underline{F}(\vartheta_{h,j}) < \bar{F}(\vartheta_{h',j})$ for all $\vartheta_{h,j}, \vartheta_{h',j} \in (0, 1)$.*

2.2 Empirical Implications

An implication is that the expected difference between predicted abilities of two groups of students for each teacher $j \in \mathcal{J}$ in the presence of ability stereotypes disfavoring group h and favoring h' is, $E[\vartheta_{h',j}] - E[\vartheta_{h,j}] \equiv \tilde{\vartheta}_j > 0$. Thus, teachers are more likely to assign lower values of forecast ability to students who belong to group h when making assessments based on stereotypes about this group, and the opposite is true for group h'. To empirically assess this implication, I compare the gender gaps on teacher-assigned and blindly graded examinations using a difference-in-differences research design. I use a value-added framework for achievement on both examinations (Rothstein, 2010, 2017) to disaggregate teachers' effects on students' academic performance (teacher and blindly graded) into value-added treatment effects that hold across all students regardless of their gender and gender-specific value-added treatment effects. I distinguish between two types of gender differences in teachers' value added. The first is unrelated to stereotypes and pertains to accommodating one group over another in teaching methods. The second is stereotype-driven and results from differential forecasting of students' abilities. The causal treatment effect, captured by the difference-in-differences parameter, considers both these factors.

Proposition 1. *Each teacher $j \in \mathcal{J}$ has a non-stereotype-driven differential assessment across student groups, Δ_j, and a stereotype-driven differential assessment across student groups, $\theta_j := \tilde{\beta}_j \tilde{\vartheta}_j$, with $\tilde{\beta}_j \in (0, 1)$. Both Δ_j and θ_j contribute to the differential assessment, disadvantaging group h and benefiting group h'. See proof in Online Appendix B.*

Assuming uniform non-stereotype-driven behavior by teachers across all their classrooms, this proposition enables estimating a lower bound for the impact of teachers' ability stereotypes, θ_j, on students' academic performance.

3 Institutional Background and Data

3.1 Public School System

The Peruvian public school system offers free education to 5.9 million pupils (74.5% of the student population in 2019) and employs 69.7% of full-time or part-time teachers. Schooling is mandatory for ages 5-17, with three levels: preschool (ages 3-5), primary (6-11), and secondary (12-17) (UNESCO, 2016). For secondary education, 1.9 million pupils were enrolled in 2019 between grades 7 and 11, the senior year. All school-aged children can enroll in public schools; however, those who enroll tend to come from underprivileged backgrounds.[5] Peruvian high schools do not offer academic, technical, or vocational courses or performance-based electives like advanced placement classes. All elementary, middle, and high schoolers learn the same subjects and curriculum. Students in a classroom have the same subject teachers.

Enrollment and classroom formation are regulated by the Ministry of Education.[6] Students' families must request admission during open enrollment (December–early February). Disabled students, siblings, and continuing students are prioritized for vacancies. Admitted students are placed in the first available section up to classroom capacity. When capacity is achieved, the remaining pupils move to the next classroom. Principals randomly or manually assign students to classes, balancing age, gender, disability, and discipline levels across classrooms.[7] In the weeks before classes start, a

[5] The 2018 National Census Evaluation of eighth graders shows that 43% of parents lack a high school diploma, while 27% have one. Census data indicates that 72% of districts nationwide have students below the 35th percentile of the student socioeconomic index distribution.

[6] The methods and criteria for enrollment are ratified by regulations such as R.S. 447-2020-MINEDU. Published annually, R.M. 193-2020-MINEDU specifies how principals use the enrollment record system (SIAGIE) to approve enrollment requests and classroom assignments.

[7] R.V. 307-2019-MINEDU restricts classrooms to 30 urban and 25 rural students, with the option for schools to adjust by 35 pupils per classroom based on infrastructure. Since 2014, SIAGIE, an enrollment record system, has randomly assigned classrooms. The principal can also employ manual assignments. The 2023 Metropolitan Lima Schools Direction survey of 42 principals revealed that 56% employ manual assignment balancing age, gender, disability, and class discipline, while 44% use random assignment.

school committee approves class formation for the student roster and presents it to the local education board. A Ministry of Education policy assigns high school teachers to schools through a process involving knowledge evaluation, interviews, class preparation, and simulation. Salary ranges depend on academic credentials, employment experience, and exam results (for instance, see Bertoni et al. (2022), Bobba et al. (2021)). The principal and school committee assign subject instructors to classrooms at the start of each academic year, prioritizing critical subjects during lecture hours. Legislation sets annual standards for instructor assignments.[8] Language arts and math teachers instruct for a duration of four to six hours per week for each class.

3.2 Students' Blindly Graded and Teacher-graded Examinations

Standardized tests (ECE) in high school are taken at the conclusion of the eighth grade year. Students demonstrate eighth-grade National Curriculum proficiency by completing four assessments over three days: math, reading, science, and social studies. The Peruvian National Bureau of Statistics (INEI) administers standardized tests to evaluate Ministry of Education policies and effectiveness. The test stakes are high for the students because their school principals, teachers, and families are informed of their individual-level score and distribution position compared to eighth graders in their region and nationwide. The math test includes numerical operations, problem-solving, and short-answer exercises, while the communications exam includes reading comprehension, short answers, incomplete sentences, and analogies. Responses are anonymously machine-graded by the Ministry of Education's Measurement of Quality Unit (UMC) following established rubrics and double-anonymized marking guidelines.[9] Teachers use year-end classroom evaluations to gauge students' proficiency in National Curriculum competencies. Exams, constructed with guidance from guidelines and textbooks, follow a curriculum-based grading system to minimize teacher preferences for specific evaluation methods (e.g., multiple-choice or essay writing). These assessments are crucial, determining students' promotion to the next grade, and scores are reported to parents. Both class assessments and standardized tests incorporate similar questions. Teachers use a national curriculum-based grading system for issuing grades. Additional information on examinations can be found in Online Appendix D.

The gender performance gap in Peru is typical of Latin American and developing countries, with boys performing better in mathematics and girls in language on standardized tests and Programme for International Student Assessment (PISA) scores (World Bank Group, 2017). High school dropout remains a challenge, especially for female students, despite increased education access between 2010 and 2019. Peru's child marriage

[8] Under R.V. 148-2023-MINEDU, teachers' time schedules determine homeroom and subject-specific instructor assignments. Seniority breaks ties.

[9] A natural disaster caused the cancellation of the 2017 exam, and ECE tests are private with limited public item releases. Former UMC assessment director L. Miranda discussed test items in December 2023.

rate and intimate partner violence prevalence are high (17% and 31%, respectively), contributing to high school dropout rates and negatively impacting female students' employment prospects (Nations, 2023). Recently implemented education reforms have focused on improving public school quality and opportunities (see, Bertoni et al. (2022), Bobba et al. (2021), Paxson (2002)) but little on closing the employment gender gap, with only the 2016 National Curriculum introducing gender equality guidelines (Minedu, 2016).

3.3 Data

Administrative Records and Outcomes.— This analysis links public school enrollment, university national registration, and Peruvian Ministry of Education and Labor employer-employee records. The longitudinal enrollment records (SIAGIE) of 7th to 11th grade public school students provide baseline demographic characteristics, test scores, teacher, and classroom assignments. From 12 to 22, I track these students' academic and job outcomes using public school enrollment, university, and employer data.[10] The national database of university student records (SIRIES) provides information on the status of college applications, admission, attendance, academic outcomes, intended majors, and declared majors in four-year college degrees. The National Tax Authority and Ministry of Labor's universe of tax-paying private firms (PLAME) with at least one contracted worker (including single-worker firms) provide employer and employee information. This monthly employment census characterizes all formal sector employees in the country, providing industry classifications for each job tenure at the month-worker-employee level. See Online Appendix A for details.

Two samples are created for empirical analysis using Peruvian student and employee administrative records. First, I estimate individual teachers' stereotyped evaluations using a teacher-student matched sample that combines results from eighth-grade standardized tests and teacher-graded classroom tests for cohorts in 2015, 2016, 2018, and 2019 (see Appendix Table XIII). A second analysis explores the impact of stereotypical grading on employment outcomes across five student cohorts. It also investigates its effects on other adult outcomes, including high school graduation. The analysis uses a longitudinal sample of 1.6 million public high school graduates from 2015 to 2019.[11]

The main set of outcomes comprises annual employment in the formal sector, earnings, and work hours, using monthly employment records—higher frequency than typical quarterly or annual measures in many administrative data sets. These outcomes are observed post-graduation, spanning ages 18 to 22. Workers are considered employed in the formal sector if they hold a job contract and earn positively with a primary employer

[10] In accordance with the data-access agreement, the Ministry of Education linked enrollment data to university data and the Ministry of Labor linked it to employer-employee data.

[11] Appendix Table IX, Panel A, illustrates the imbalanced data structure of the teacher-student matched sample, covering multiple high school grades in each student cohort. Panel B displays employment outcomes available for each cohort year in the longitudinal student sample.

for at least one month between 2015 and 2020. Primary employers contribute most of the quarterly earnings (Abowd, Lengermann and McKinney, 2003, Card, Heining and Kline, 2013, Sorkin, 2018). This definition calculates annualized pay and labor hours for an employee's primary employer, even with multiple employers. Mechanisms explored from administrative data encompass high school graduation, tracked over time to observe academic progress. Following Gray-Lobe, Pathak and Walters (2021), I project a student's graduation year based on the standard academic timeline for five high school years, starting from 7th-grade enrollment and assuming continuous progression to 11th-grade.

Table I summarizes the demographic and educational characteristics of 8th- through 11th-grade students in the (stacked) base and estimation samples. The base sample comprises all successful student-teacher matches, irrespective of having a stereotypical grading measure, while the total sample includes matches specifically with a stereotypical grading measure. Estimation sample students are more likely to speak Quechua and face high school retention compared to base sample students. Their parents are less likely to have attended college. After graduation, I examine students' 2015–20 workers' records, with each cohort observed for a specific time window based on their projected graduation year. Appendix Table XIV presents descriptive statistics for public school students matched to labor records (columns (4) and (5)) and a benchmark group of workers aged 18–25 (columns (1) to (3)). The first panel displays 2015–20 average earnings and hours, while the second and third panels detail workers' characteristics with their most recent primary annual employer. The regression sample comprises roughly 85,700 students who enter the laborforce with available records and projected graduation years from 2015 to 2019. Female workers in the regression sample earn about 14% less than males, mirroring a similar disparity in the benchmark sample. Both samples show comparable high school graduation rates for women and men, but men report higher rates of college and technical education.

Survey Design and IAT Data Collection.— This study conducted a nationwide survey of Peruvian teachers and high school students, utilizing a teachers' questionnaire and the Spanish-adapted IAT (data collection coverage in Appendix Figure 7). The Ministry of Education's Office of Monitoring and Strategic Evaluation executed data collection between September 2021 and September 2022 as part of its remote-learning program for high school students.[12] Survey questions and data collection procedures are described in the Online Appendix A. The analysis sample consisted of 2,541 mathematics and language arts teachers who completed a questionnaire and IAT, matching stereotyped administrative assessment data, excluding other subject teachers.

In an educational portal designed for this study, gender-science and gender-career IATs are administered in a random order. Self-reported gender attitudes and gender-related

[12] The data collection platform, "Opportunities for Everyone" (*Oportunidades para Todos* in Spanish), does not explicitly reference gender stereotypes. Teachers and students are invited to assist the ministry in developing inclusive learning tools and policies for all students.

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

	Full sample			Regression sample		
	Male (1)	Female (2)	Total (3)	Male (4)	Female (5)	Total (6)
A. Demographic characteristics						
Grade retention	0.03	0.02	0.03	0.04	0.02	0.03
Spanish	0.88	0.89	0.88	0.86	0.86	0.86
Quechua	0.10	0.09	0.09	0.11	0.11	0.11
Other language	0.02	0.02	0.02	0.02	0.02	0.02
Born in Lima	0.26	0.26	0.26	0.22	0.22	0.22
Parents Some College	0.12	0.12	0.12	0.09	0.08	0.08
Parents College +	0.08	0.08	0.08	0.03	0.03	0.03
B. HS graduation & college						
Graduated HS ever	0.87	0.90	0.88	0.85	0.88	0.87
Enrolled	0.22	0.26	0.24	0.14	0.17	0.16
Enrolled STEM	0.11	0.05	0.08	0.07	0.03	0.05
Observations	8,756,144	8,437,295	17,193,439	3,366,569	3,104,277	6,470,846
Num. of students	1,121,799	1,088,462	2,210,254	847,010	805,413	1,652,417
Num. of teachers			123,078			37,508

Note: This table presents summary statistics for 8th-11th-grade students (2015—2019) with available end-of-grade scores from the previous year. Columns (1-3) describe the base sample—students matched with teachers in math, language arts, or science. Columns (4-6) show statistics for the full regression sample, excluding base sample students with teachers lacking stereotypical assessment measure information. Observations are at a subject-by-grade level.

behaviors in the classroom are collected in the teachers' questionnaire, which supplements the IAT. Teachers associate science and the humanities with gender groups in the gender-science IAT and with career and family terms in the gender-career IAT. The gender-science IAT assesses teachers' strength of association between science, humanities concepts, and gender-specific terms (Greenwald, Nosek and Banaji, 2003, Greenwald, McGhee and Schwartz, 1998, Nosek, Greenwald and Banaji, 2007, Nosek et al., 2014). Participants make millisecond associations between words from each group (sciences and humanities) and gender terms (female and male). Quicker associations between, for example, the humanities and girls or the sciences and boys reveal a stronger latent connection, reflecting implicit gender stereotypes. IAT scores range from –2 to 2, with higher values indicating a stronger implicit preference for boys in science and girls in the humanities over boys in the humanities and girls in science.[13] Figure 1 depicts the IAT score distribution for educators in the analysis sample. Math instructors average 0.316, and language arts teachers average 0.307, indicating that teachers in these main high school subjects reinforce gender stereotypes about abilities. Unlike self-reported attitudes, IAT responses, measuring the speed of categorizing science and gender terms, are less influenced by cognitive processing. The IAT provides a more reliable assessment,

[13] The study uses an enhanced scoring method, as suggested by Lane et al. (2007) and Greenwald, Nosek and Banaji (2003), to eliminate teachers' excessively slow or fast responses, preventing fatigued random associations.

overcoming social desirability bias where participants may conform to cultural norms in their responses (De Houwer et al., 2009, Egloff and Schmukle, 2002, Steffens, 2004).

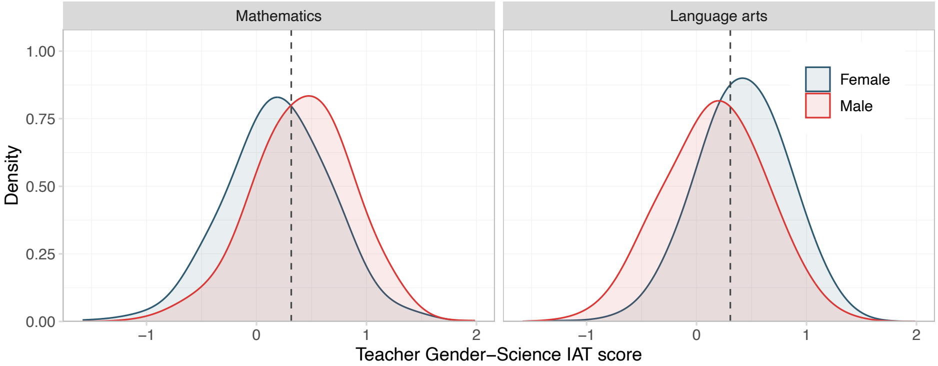

Figure 1: Distribution of Teachers' Gender-Science IAT Scores

Note: This figure displays the density of IAT scores for 2,541 teachers (Math=1,345; Language arts=1,196) successfully matched with administrative records, indicating teacher-level gender differences in grading. Following Nosek, Greenwald and Banaji (2007), scores below –0.15 favor girls, those between –0.15 and 0.15 suggest little to no stereotypes against girls, scores between 0.15 and 0.35 indicate slight stereotypes against girls, and scores above 0.35 suggest moderate to severe stereotypes against girls in math and science. Dashed lines represent the IAT score means.

Table II presents information on teachers with an available IAT and stereotyped grading measure. Gender stereotypes are prevalent, with 47.5% of math teachers and 45.9% of language arts teachers exhibiting moderate to severe or strong stereotypes against girls in science. Career stereotypes against women in professional careers are observed in 35.1% of math teachers and 41.1% of language arts instructors. Mathematics teachers are gender-balanced, while there's a female majority among language arts teachers. Demographics, class size, and weekly work hours are similar for both subjects. Public school teachers average 12 years of experience, 49% have tenure contracts, and 16–19% have encountered discrimination from colleagues or administrators.

4 Measuring Teachers' Stereotyped Assessments

Using eighth-grade student-teacher matched data, I calculate a parameter capturing systematic gender disparities in teachers' assessments across teacher- and blind-graded exams. Building on methods from existing studies (see, for instance, Lavy (2008), Lavy and Sand (2018), Lavy and Megalokonomou (2024), Terrier (2020)), this paper proposes an estimation approach to consider teachers' value added driving score differences between boys and girls across exams when calculating bias or stereotypes in assessments. Additionally, the approach recognizes that gender-related observable factors, rather than teachers' stereotyped grading, may contribute to systematic variations in grading between boys and girls on teacher-assigned tests. To address this, the study proposes estimating teachers' exam-specific value-added as random intercepts and adjusting systematic grading variations between genders across exam types using a comprehensive set of

Table II: Descriptive statistics for surveyed teachers with a stereotyped assessment measure

	Mathematics (1)	Language arts (2)	Total (3)
A. Demographic characteristics			
Female	0.45	0.57	0.51
Age, years	43.35	42.11	42.77
Mixed	0.66	0.66	0.66
Quechua	0.23	0.21	0.22
White	0.02	0.02	0.02
Afroperuvian	0.03	0.02	0.03
B. Job characteristics			
School in Lima	0.26	0.27	0.26
Teaching hours per week	28.54	28.08	28.32
Number of students per class	24.56	25.14	24.83
Position, tenure	0.47	0.45	0.46
Public sch experience, years	12.98	11.93	12.49
Private sch experience, years	2.84	3.04	2.93
College major, education	0.86	0.92	0.89
College	0.66	0.67	0.67
C. Discrimination and IAT			
Experienced discrimination	0.16	0.19	0.17
Witnessed discrimination	0.17	0.19	0.18
Gender-science stereotype (IAT score)	0.32	0.31	0.31
Gender-career stereotype (IAT score)	0.27	0.30	0.28
Observations	1,345	1,196	2,541

Note: This table reports descriptive statistics for surveyed teachers in the analysis sample matched with assessment bias. Age was reported between September 2021 and July 2020. Raw IAT scores presented are interpred as follows: strong bias' if score > 0.65, “moderate to severe bias if 0.65 score > 0.35, slight bias if 0.15 score > 0.35, little to no bias if -0.15 score 0.15, preference for girls if score < -0.15.

covariates.

4.1 Estimation Strategy

Following the model in Section 2, I define systematic gender differences in teacher j's assessment as the disparities between male and female students' mean gaps between teacher-assigned scores (S_{ij}^{T}) and blindly graded test scores (S_{ij}^{B}). The system of estimating equations presented as follows involves parameters of interest $(\alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \beta_{2,j})$.

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

$$S_{ij}^{T} = \beta_{1,j} + \beta_{2,j}\, G_{ij} + W_{ij}'\beta_3 + \epsilon_{ij} \tag{3}$$

The specification includes teacher-specific intercepts $(\alpha_{1,j}, \beta_{1,j})$ and gender effects on score assignment $(\alpha_{2,j}, \beta_{2,j})$. The indicator G_{ij} denotes 1 for male students and 0 for females. Adjusting for potential confounders, such as student discipline (Botelho, Madeira and Rangel, 2015) and teacher-match effects (Dee, 2005), an extensive set of student and teacher characteristics (W_{ij}) is included. These factors encompass age, ethnicity (proxied by language), grade retention, birthplace, parents' education, household status, teachers' gender and contract (seniority), school-location fixed effects, and lagged scores in physical education, social studies, mathematics, and language arts. Quadratic polynomials of lagged physical education and social studies scores proxy observed classroom behavior, as in Botelho, Madeira and Rangel (2015). Robust standard errors, clustered at the student level, are reported.

Using the parameters $\beta_{2,j}$ and $\alpha_{2,j}$ calculated by each teacher, I construct a measure of teacher-level systematic gender differences in assessing boys and girls. The differences between teacher-assigned and centrally assigned scores, attributed to the disclosure of students' gender in non-blind examinations, are denoted as $\theta_j := \beta_{2,j} - \alpha_{2,j}$. Under the identifying assumptions, θ_j is interpreted as the adjusted discrepancies in gender gaps between teacher-graded and centrally graded assessments, measured in standard deviations. Larger values of this underlying parameter for teacher j_1 compared to teacher j_2 indicate that j_1 is more likely to assign higher scores to boys than girls, reflecting the teacher's ability stereotypes against girls relative to the gender gap on blindly graded tests. This serves as a relative measure, facilitating comparisons of gender differences in assessment across teachers.

Identifying assumptions.— Two assumptions enable the identification and interpretation of θ_j as an informative metric for systematic gender differences in assessments due to gender stereotypes, facilitating comparisons among 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. This assumption is threatened by the possibility that teacher-assigned and centrally assigned scores on eighth-grade exams reflect the evaluation of slightly different abilities among boys and girls as test-taking skills may vary based on exam format.[14] Unobserved skill gaps between boys and girls could lead to performance differences in certain question types if this were true in this setting. Girls might struggle on tests with less favorable question types, not due to teacher stereotypes but because the format assesses skills boys prefer or excel in. To rule out the chance of teachers consistently favoring specific content or assessment strategies that benefit one gender, I compare the distribution of question contents and formats between classroom and centralized tests in Online Appendix D.[15] Appendix Table XXVIII confirms both math tests cover identical

[14] Liu and Wilson (2009) reveals males excel in complex multiple-choice and non-textbook-context questions, while Gallagher and De Lisi (1994) finds girls perform better in textbook-related questions. Klein et al. (1997) notes girls outperform in written performance assessments. However, Ben-Shakhar and Sinai (1991) states boys tend to guess more in multiple-choice exams, resulting in lower omission rates than girls.

[15] I examine the National Curricula for high school and eight-grade Ministry of Education exit exams, available at https://repositorio.minedu.gob.pe/handle/20.500.12799/7972.

materials and learning skills. Appendix Table XXIX reveals the consistent distribution of questions on each topic in eighth-grade exams. This content alignment reflects teachers' dedication to adhering to national curriculum standards.

In further evaluating the identifying assumption, I analyze the question types present in both exams. A pedagogy expert and an external annotator classified 192 test items from classroom and standardized exams to explore format differences. Classroom test items (102) were provided by the Direction of Metropolitan Lima Schools and drawn from Ministry of Education model examinations. In contrast, official technical documentation supplied 2019 standardized evaluation item formats (90). The results from the annotation analysis in Appendix Figure 12 reveal minor differences in question types between eighth-grade exams, suggesting classroom evaluations don't favor gender-specific question patterns.

Another challenge to the identification assumption is the potential for boys, in certain classes, to excel in skills needed for teacher-assigned tasks but not centrally graded tasks based on exam format. For example, cultural norms might lead boys in certain classes to focus on rapid problem-solving, earning higher marks in commonly asked classroom questions. This could shift gender discrepancies in teacher-assigned scores ($\beta_{2,j}$) but not in blindly graded scores ($\alpha_{2,j}$). Concerns about examination conditions favoring one gender over the other (Gneezy, Niederle and Rustichini, 2003, Niederle and Vesterlund, 2010) are alleviated as both exams have identical testing conditions for all genders.

The second identifying assumption is that, if there is any skill difference between boys and girls, it is uniform across all teachers' classrooms. This ensures a consistent comparison of skill levels between boys and girls taught by the same teacher in different classrooms. The assumption relies on the expectation that teachers maintain consistent evaluation practices across all their classrooms, like standardized grading or teaching approaches using formerly made lesson plans. For instance, a math teacher assessing problem-solving ability using the same criteria across classrooms would consistently show gender-based test differences. This supports the idea that teacher stereotypes, not student abilities, contribute to observed gender gaps in θ_j. Lastly, it is not assumed that gender stereotypes are responsible for the location of the distribution of θ_j because the mentioned assumptions do not require a cardinal but a ordinal interpretation of θ_j.

4.2 Estimates of Teacher-Level Gender Differences in Assessments

The teacher-level estimates $\hat{\theta}_1, \ldots, \hat{\theta}_J$ and associated standard errors $s_1, \ldots, s_J$ are obtained by estimating Equations (2) and (3). The scores for estimating $\hat{\theta}_j$ are standardized by year and subject. Equations (2) and (3) are estimated separately using a random-effects approach and the matched teacher-student sample containing classroom grades and test scores in the respective subjects. Figure 2 shows the distribution of $\hat{\theta}_j$ indicates that some teachers systematically assign higher scores to boys while others favor girls in mathematics and language arts.

Examining the heterogeneity of teacher-level estimates of gender differences in assessments depicted in Figure 2 is the focus of Appendix Table XV. The variance of the estimated θ_j distribution is likely upwardly biased due to sampling error. I report a student-weighted bias-corrected estimate of the underlying θ_j's variance, 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 $w_j = \mathcal{C}(j)/\mathcal{C}$ and $\mathcal{C}(j)$ is the size of teacher j's classroom for a given subject, and $\mathcal{C} = \sum_j \mathcal{C}(j)$ for that subject.[16] The results indicate substantial heterogeneity among teachers' gender differences in assessments, with mathematics teachers displaying less dispersion than language arts teachers. The preferred (student-weighted) variance measures show that systematic gender differences in grading disproportionately affect female students. Moving upward one standard deviation in the distribution of mathematics- and language arts-teacher gender grading gaps increases the score gap of boys versus girls by 0.05 and 0.07 test-score standard deviations, respectively.[17] The dispersion of gender-assessment differences in mathematics is smaller than in language arts, possibly due to limited discretion in awarding partial credit on numerical exercises or better student ability to determine correctness in assigning full credits per question.

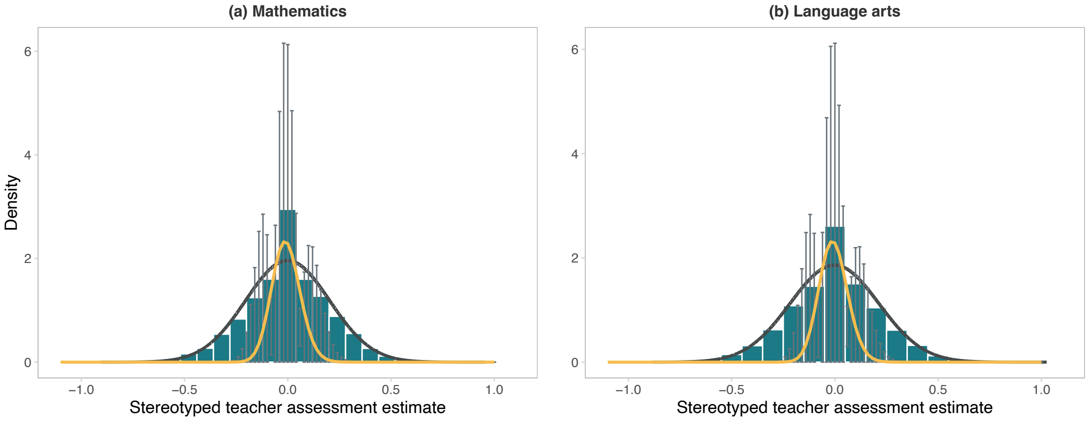

Figure 2: Teacher-level Stereotyped Assessment Posterior Means Under EB Deconvolution
Note: The figure displays empirical Bayes (EB) deconvolution estimates for the prior density of teacher-level stereotyped grading and a benchmark Gaussian density. The histogram shows the distribution of estimated teacher-level stereotyped assessments. The yellow line indicates the prior density of underlying population parameters, computed numerically through EB deconvolution from Efron (2016), using the deconvolveR package by Narasimhan and Efron (2020). The complexity penalty parameter was chosen to align the deconvolved density with the mean and bias-corrected variance of estimated teacher-level stereotyped grading. The parameter support ranges between -2 and 1 for mathematics and language arts. Gray bars indicate ± 1.96 estimated standard error, while the black line corresponds to the Gaussian density with the mean and variance of the posterior deconvolved distribution.

[16] Sampling variation can arise from factors like teachers' limited experience or classrooms with few students, making it challenging to reliably estimate $\hat{\theta}_j$ (see Kane and Staiger (2002) for an example). Methods for reducing systematic bias from observed variance, involving subtracting estimated standard errors s_j, are detailed in Aaronson, Barrow and Sander (2007), Abdulkadiroğlu et al. (2020) and Angrist et al. (2017). A bias-corrected estimate, weighted by school size, is also presented.

[17] In line with these findings, the dispersion of systematic grading differences between boys and girls by math teachers is smaller than that of language arts teachers (English and Hebrew) in Lavy and Sand's (2018) study.

Figure 2 displays the deconvolved density $\hat{g}(.)$ of teacher-level stereotyped assessment parameters θ_j, alongside the theoretical Gaussian density and observed distribution. Using Empirical Bayes (EB) methods detailed in Online Appendix E, I derive distributional estimates of $\hat{\theta}j$ and correct sampling error of individual level estimates, as in (Kline and Walters, 2021). Employing parametric EB (linear shrinkage estimation) and EB deconvolution, I utilize unbiased but noisy estimates, $\hat{\theta}j$, of the true underlying teacher-level stereotyped assessment parameters, θ_j. These posterior-mean estimates become the variable of interest for calculating the long-term effects of stereotyped grading on outcomes. In further analysis, I investigate the relationship between teacher characteristics and the estimated parameter $\hat{\theta}_j$. Appendix Table XVI examines how observed teacher characteristics and implicit gender-based ability stereotypes predict estimated gender differences in assessment.

4.3 Are Gender Differences in Assessments Correlated with Gender-Based Ability Stereotypes?

I empirically examine the extent to which teachers in this context hold gender-based prior beliefs about their student's abilities, as stated in Equation (1), reflected in the assessment-based gender stereotype measure. I achieve this by correlating gender differences in the assessment measure, $\hat{\theta}_j$, with a proxy measure for teachers' ability stereotypes in math and science relative to communications—the gender-science IAT score, IAT_j—using the following regression equation:

$$\hat{\theta}_j = \gamma_{1,s(j)} + IAT_j'\gamma_2 + X_j'\gamma_3 + \kappa_j \tag{4}$$

$\gamma_{1,s(j)}$ represents school-location fixed effects, and X_j is a vector of teacher characteristics, including demographic and job-related details from administrative records and surveys, with $s(j)$ denoting the school location where teacher j currently teaches. The estimated coefficient of interest, $\hat{\gamma}_2$, signifies the effect of a one standard deviation increase in math teachers' math and science stereotypes (measured by the IAT score) on their gender gap in grading mathematics.

I analyze a sample of 2,541 math and language arts teachers with IAT scores and available $\hat{\theta}_j$ estimated from administrative records using Equation (4). Table III presents evidence on the link between implicit ability stereotypes and grading gaps between female and male students. All specifications include school location fixed effects, and controls encompass teachers' gender, childbearing status, birth decade indicators, ethnicity, IAT association order, and the number of previous IATs. In columns (1) and (2), math teachers associating boys more with science and girls more with the humanities (per their IAT score) are likely to exhibit grading gaps favoring boys. In essence, teachers who perceive boys as better than girls in math- and science-related fields tend to give boys higher grades than deserved in mathematics, while giving girls lower grades than deserved.

This pattern persists with or without covariates. In columns (3) and (4), language arts teachers with such associations are less likely to generate gaps disadvantaging female students. Teachers awarding girls higher grades in language arts, where they perceive girls as more competent, imply that coefficients indicate a penalty for female students in mathematics (consistent with stereotypes) but not in language arts. The sign of coefficients aligns with subject-specific gender stereotypes, indicating grading differences reliably reflect gender-based math stereotypes.

Table III: Gender Differences in Assessment and Implicit Gender Stereotypes Relationship

	Mathematics teachers		Language arts teachers	
	(1)	(2)	(3)	(4)
IAT score	0.388**	0.210*	-0.649***	-0.520***
	(0.170)	(0.122)	(0.091)	(0.077)
Controls	No	Yes	No	Yes
R-squared	0.712	0.781	0.803	0.838
Observations	1,345	1,345	1,196	1,196

Note: This table presents regression estimates of implicit stereotypes (measured by IAT scores) on gender differences in the assessment of mathematics and language arts teachers, based on Equation (4). $\hat{\theta}_j$ is divided by the bias-corrected standard deviation, and IAT_j is standardized. Coefficients are in standard deviations. Observations are teacher-level, with robust clustered standard errors by school location in parentheses. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

5 Long-Term Effects of Teachers' Stereotypical Assessments

5.1 Estimation Strategy

The exposure to stereotyped assessments varies from teacher to teacher as students progress through grades. Let $\hat{\theta}^*_{j,-i}$ represent the leave-one-year-out estimator of the posterior mean for teacher j assessing student i in a specific grade. This estimator is normalized using the mean and standard deviation calculated per school year.[18] The main estimating equation to retrieve the causal relationship between teachers' stereotypical evaluations and student i's long-term outcomes variable, Y_i, is as follows:

$$Y_i = \delta_0 + \delta_1 \hat{\theta}^*_{j,-i} \cdot Female_i + \delta_2 Female_i + \delta_3 \hat{\theta}^*_{j,-i} + \delta_4 \mathbf{X}_i + u_i \qquad (5)$$

δ_1 measures the differential impact of a one standard deviation increase in teachers' stereotypical assessments of girls versus boys in a high school grade. δ_3 represents the effects of male students exposed to stereotypical assessments. $Female_i$ indicates the student's gender. The covariate vector $\mathbf{X}_i$ has four control groups. First, student-level controls include language, age, mother's education, repetition, and birthplace indicators for social norms.[19] Second, I include teacher characteristics such as

[18] The standard deviation for standardizing $\hat{\theta}j, -i$ is derived from bias-corrected measures in Table XV for the corresponding school year.

[19] By Carlana (2019), I consider student gender-related social norms to be related to the place of birth;

gender, age, type-of-appointment indicators (such as homeroom teacher, subject teacher, or administrative teacher), and, as a proxy for teacher experience, seniority-based contracts (for example, tenured or fixed-term contracts). As teachers' behavior may vary according to the students' gender, teacher characteristics interact with student gender. Third, to account for the compositional effects of assigned classes, classroom-level controls consider class size and averaged student-level characteristics. The fourth group comprises school-grade averages to address confounding variables across grades in the same school. Finally, the model also includes quadratic polynomials of prior grade classroom scores, school-grade means of these lagged scores, and fixed effects for cohort, grade, school–year, and school.

I utilize a leave-one-year-out teacher-level stereotyped assessments, $\hat{\theta}^{*}_{j,-i}$, computed with posterior-mean estimates (see Online Appendix E for details). This prevents potential correlation between $\hat{\theta}_j$ and long-term outcomes by avoiding using the same students for both teacher-stereotyped assessments and the dependent variable reflecting students' long-run outcomes. Consider calculating the leave-one-year-out teacher's stereotypical grading, the treatment variable for 2019 graduates. I exclude the data from all of the 2019 graduation cohort's examinations (that is, data on eighth-grade students who took the examinations in 2015) and use the remaining students' observations to calculate all teachers' stereotyped grading. This is repeated for other projected graduating cohorts, ensuring independence between teacher-stereotyped assessments and student outcomes.[20]

5.2 Identification of Long-Term Effects

The crucial identifying assumption is that, after conditioning on observed characteristics of students, teachers, classroom- and school-grade-level characteristics, and lagged student scores, a student's assignment to teacher j is orthogonal to unobserved determinants in levels and unobserved differences across genders that influence high school completion and labor market outcomes. The widely used selection-on-observables assumption, found in value-added literature (see Chetty, Friedman and Rockoff (2014*a*,*b*), Kane and Staiger (2008), Rothstein (2010, 2017)), addresses concerns of ability-based student sorting. Both the value-added literature (see, for instance, Abdulkadiroğlu et al. (2020), Angrist et al. (2017), Rothstein (2017)) and education production function literature highlight that incorporating student lagged test scores captures unobserved confounders related to socioeconomic factors and cognitive skills.

There may be concerns about schools not adhering to the procedures and guidelines provided by the Ministry for classroom assignments to teachers or that parents may

thus, I control for students' birthplace and gender interaction.

[20] For 2018 and 2019 cohorts, I create a leave-one-year-out measure for teachers' stereotypical grading by excluding 2015 and 2016 scores. Other cohorts get regular teacher stereotyped-assessment estimates, combining data from 2015, 2016, 2018, and 2019. Other studies employ different leave-out procedures, excluding students individually (Terrier, 2020) or classes Lavy and Megalokonomou (2024).

inadvertently attempt to alter them. For instance, parents with advanced degrees might want to change their daughters' math classes to those with less stereotypical teachers. However, subject-specific transfers are not allowed in the Peruvian system, making it unlikely for parents to change classes based on a single teacher's behavior. Therefore, parents are unlikely to make such changes based on a single subject teacher's perceived behavior or quality.

Complementary evidence supports the validity of the assumption. Table IV examines balance across math classrooms with diverse stereotyping levels, finding no systematic relationship with students' baseline characteristics. This analysis controls for fixed effects like cohort, grade, year, and school. No systematic differential assignment is observed between girls and boys in column (1). Column (2) shows no impact of students' ethnicity, migrant status, place of birth, or lagged test scores on allocation to stereotyped grading classrooms, with no distinct consequences for females. In columns (3) and (4), I analyze whether adolescents with higher-educated mothers are consistently less likely to be allocated to teachers who perpetuate negative stereotypes than those with less than a high school education. Students whose lagged test scores indicate higher proficiency in mathematics or language arts prior to the assignment are not less likely to be placed in classrooms where teachers use more biased grading criteria, as shown in column (5). Column (6) includes all characteristics, demonstrating that principals' formation procedures and parents' requests are unaffected by gender, academic performance, parents' education, or other demographics.

Next, I examine if classroom assignments resemble random allocation after accounting for baseline fixed effects (cohort, grade, years, and schools). I demonstrate that instructor characteristics are mostly unrelated to student pre-determined traits affecting assignments. Appendix Figure 9 displays the estimated effects of gender-interacted students' traits on teachers' observable characteristics, including contract type, experience, degree institution (university vs. technical institute), and promotion evaluation scores. Around 15% of the 136 coefficients in these regressions are statistically significant, aligning with expected chance results. Finally, in line with Bietenbeck (2020), I assessed if within-school variation in teacher-level stereotyped grading exposure aligns with nearly random teacher assignments across classrooms. Monte Carlo simulations replicated real data distributions for school size, stereotypical grading teachers, class capacity, and school size by randomly assigning students and teachers to school classes. After regressing teacher-level stereotyped assessment on school fixed effects in both actual and simulated data, I collected the residuals. Appendix Figure 10 displays root mean squared errors and visual similarity between simulated and actual residuals, providing additional evidence that simulations accurately reflect assignment patterns and variability in real data.

Table IV: Exogenous Assignment of Students to Mathematics Teachers with Varied Stereotypical Grading Levels

	(1)	(2)	(3)	(4)	(5)	(6)
A. Student characteristics						
Female	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Spanish-speaking		0.001		0.001		0.001
		(0.001)		(0.001)		(0.001)
Migrant		-0.001		-0.001		-0.001
		(0.003)		(0.003)		(0.003)
B. Parental traits						
Mother, college			0.001	0.001		0.001
			(0.001)	(0.001)		(0.001)
C. Previous grade test scores						
Math scores, $t-1$					0.000	0.000
					(0.000)	(0.001)
Lang. arts scores, $t-1$					0.000	0.000
					(0.000)	(0.001)
D. Characteristics interacted with female						
Spanish-speaking		0.000		0.000		0.000
		(0.001)		(0.001)		(0.001)
Migrant		0.002		0.003		0.003
		(0.004)		(0.004)		(0.004)
Mother, college			-0.002	-0.002		-0.002
			(0.001)	(0.001)		(0.001)
Math scores, $t-1$						-0.001
						(0.000)
Lang. arts scores, $t-1$						0.000
						(0.000)
R-squared	0.416	0.416	0.416	0.416	0.416	0.416
Observations	2,562,227	2,562,227	2,547,015	2,547,015	2,562,227	2,547,015

Note: This table reports estimates of the correlation between teacher-level stereotyped assessment practices and student pre-determined characteristics that are potential sources of student sorting. Mathematical and language arts lagged test scores are indicative of the student's performance in the preceding grade end-of-year class exam. As their information has been stacked across all grades in high school, the unit of observation is the student-grade. Baseline fixed effects are included at the level of grade, cohort, year, and school. Standard errors are clustered at the student and school levels.

5.3 Effects on Labor Market Outcomes

Employment in the formal sector.— This section discusses the impact of stereotyped grading exposure on labor market outcomes for cohorts graduating between 2015 and 2019. I track the trajectories of 17- to 22-year-olds who are now high school graduates, either working full-time in the formal sector or working part-time while attending college. Given the significance of mathematical knowledge and teachers in education literature, the focus is on the impact of math teachers' exposure to stereotypical assessments. Most jobs are non-professional entry-level positions due to the workforce's average age. Increased exposure to stereotypical grading has a significant, positive, and lasting effect on formal sector employment for 17- to 22-year-old boys. Effects for women are smaller and negative until the age of 18–19, becoming statistically insignificant afterward.

Table V, Panel A, shows estimated effects on the likelihood of formal sector employment using Equation (5). In each column head, the median age of students is shown for the first 5 years after graduation, a period when they typically enter the workforce. In column (1), 17-to-18-year-old girls have a 0.6 percentage point negative influence on formal sector employment. As minors, certain students in this age group may need parental or legal-guardian authorization before signing work contracts. Subsequent columns present preferred estimates for young job seekers legally eligible to work. In column (2), a 0.7 percentage point effect is shown, representing 21% of the mean outcome, on the gender gap in formal employment for 18–19-year-olds. The bottom rows show mean outcome values, indicating around 97% of girls in this age group are likely employed informally. Columns (3) through (5) continue to reveal negative and statistically significant effects for girls, persisting up to 5 years after graduation, with a 2 percentage point effect on the gender gap at ages 21–22.[21]

Paid hours worked.—On the intensive margin of labor supply, Panel B indicates that stereotyped grading has modest adverse effects on women's monthly paid work hours and significant positive effects on men's, thereby widening the gender gap. In columns (1) to (3), heightened exposure to a teacher with a one-standard-deviation more stereotypical assessment during one high school grade slightly reduces women's monthly paid work hours (0.5% to 0.9% of the mean), contributing to approximately 14% of the gender gap at this age. Columns (4) and (5) show that the unfavorable effects on women between 20 and 22 years old are no longer statistically significant four years after high school graduation. For men, columns (1) to (3) reveal a significant effect between ages 17–18 and 19–20, leading to increased paid work hours by 0.3% to 0.8% of the mean after exposure to more stereotypical evaluations.

Considering the total effect (sum of main and differential effects), teachers with stronger stereotyped grading practices boost outcomes for boys but have little to no effect on girls. Panel (a) in Figure 3, at ages 17–18 and 18–19, the total effect on girls' formal employment is initially statistically significant at –0.2 percentage points (6% of the mean). Boys at these ages experience an increased likelihood of formal sector employment by 0.4–0.5 percentage points (6%–12% of the mean). After nearly 3 to 5 years post-graduation, the total effect on girls becomes negligible while remaining substantial for boys, ranging from 0.6 to 2.2 percentage points (4%–12% of the mean). Panel (b) in Figure 3 indicates adverse and statistically significant total effects on women's paid working hours from ages 17 to 19, becoming statistically insignificant at age 20, aligning with total effects on women's likelihood of holding a formal sector job. For men exposed to a more stereotypical assessment teacher, there is an additional 2 hours of work between ages 21 and 22.

[21] In a supplementary analysis of these effects, I exclude students who applied or enrolled in college. The magnitudes of the effects are very similar to those in Table V, with a slight increase by age 19.

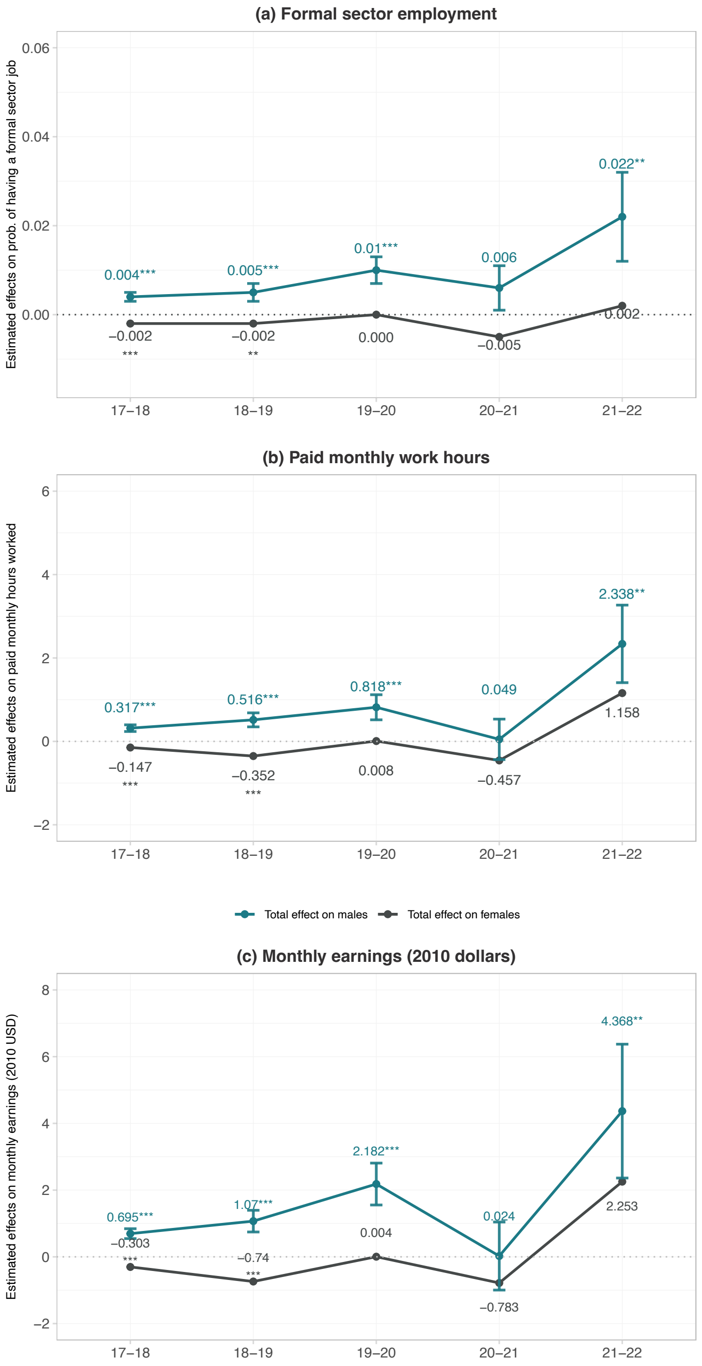

Figure 3: Effects of Teachers' Stereotypical Assessments on Employment Outcomes

Note: This graph displays the estimated total effects on labor market outcomes from a one-year exposure to a stereotypical assessment mathematics teacher. Green lines depict total effects on males, while dark grey lines show total effects on females, combining main and differential effects per Equation (5). Error bars indicate clustered standard errors by student and school for effects on males. The joint significance of total effects on females is denoted alongside the gray lines. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table V: Formal Sector Employment: Effects of Exposure to Mathematics Teachers' Stereotyped Assessments

	Age 17–18	Age 18–19	Age 19–20	Age 20–21	Age 21–22
	(1)	(2)	(3)	(4)	(5)
	A. Employed in the formal sector after high school graduation				
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.006***	-0.007***	-0.010***	-0.011*	-0.020**
	(0.001)	(0.002)	(0.004)	(0.005)	(0.010)
$\hat{\theta}^*_{j,-i}$	0.004***	0.005***	0.010***	0.006	0.022**
	(0.001)	(0.002)	(0.003)	(0.005)	(0.010)
Female	-0.039***	-0.072***	-0.092***	-0.112***	-0.092***
	(0.002)	(0.004)	(0.007)	(0.011)	(0.022)
$\bar{Y}$ female	0.010	0.033	0.060	0.084	0.102
$\bar{Y}$ male	0.031	0.075	0.119	0.156	0.182
Joint sig., p-val	0.001	0.027	0.922	0.267	0.855
R-squared	0.041	0.042	0.037	0.040	0.044
	B. Paid monthly work hours				
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.464***	-0.868***	-0.810**	-0.506	-1.180
	(0.097)	(0.213)	(0.357)	(0.547)	(1.030)
$\hat{\theta}^*_{j,-i}$	0.317***	0.516***	0.818***	0.049	2.338**
	(0.080)	(0.168)	(0.301)	(0.485)	(0.930)
Female	-2.854***	-5.525***	-8.229***	-10.964***	-7.132***
	(0.214)	(0.432)	(0.746)	(1.207)	(2.338)
$\bar{Y}$ female	93.681	95.912	97.337	99.804	100.242
$\bar{Y}$ male	101.148	103.529	104.422	104.647	104.202
$\bar{Y}$ female uncond.	0.600	2.411	4.726	6.723	7.746
$\bar{Y}$ male uncond.	2.202	6.071	10.154	13.229	14.578
Joint sig., p-val	0.009	0.004	0.972	0.253	0.140
R-squared	0.027	0.029	0.028	0.031	0.029
Observations	2,562,227	1,700,366	893,625	402,928	121,041

Note: This table presents estimated coefficients of the impact of increased gender-stereotypical teacher assessments in a specific high school grade on the probability of holding a formal sector job and the monthly paid work hours ages 17–22 (post-high school graduation). The outcome variables consider the contract with the dominant annual employer for at least one month during the specified age range. The treatment variable is the leave-one-year-out teacher-level stereotyped assessment, normalized by its standard deviation at the year-subject level. Student, teacher, lagged scores, classroom- and school-grade-level controls, place-of-birth fixed effects and gender interaction, cohort, grade, school, and year-fixed effects are covariates. The p-value for the joint significance of the total effect on girls is reported. The unit of observation is the student-grade, clustered standard errors by school are reported in parentheses. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

In the country's labor markets, men benefit more than women from stereotypical grading, with limited and shorter-lived effects for women in the formal economy. The substantial informal labor sector, constituting 64% of jobs between 2015 and 2019 (**?**), poses challenges for high school graduates without pre-age 18–19 work experience in securing formal sector employment. Despite these challenges, male students benefit from favorable treatment through teachers' stereotypical grading. Women, at 19–20 years old, can offset these effects with higher education achievements, given that 68% of workers have a high school diploma or less. These findings indicate that teachers who rely on gender stereotypes may inadvertently steer female students away from desirable career paths, causing a temporary delay in their entry into the formal sector.[22]

Monthly Earnings.— Table VI presents estimated monthly earnings losses (in 2010 USD) for recent high school graduates aged 17 to 22. Columns (1) to (3) indicate statistically significant effects on women between ages 17 and 20, exacerbating the gender pay gap. Having a teacher with one-standard-deviation more stereotypical grading in one high school grade leads to a monthly earnings loss of USD 1 for women relative to men at ages 18–19, increasing to USD 2.2 at ages 20–21. No differential exposure effects on girls are observed 4 and 5 years after high school completion (columns 4 and 5), suggesting dissipation of teachers' stereotypical grading effects on women as they gain work experience and potentially reach educational milestones. By ages 20–21, university or technical education students may have earned sufficient credits for legal paid apprenticeships supervised by their institutions.

Referring to total effects on male and female graduates, Panel (c) in Figure 3 indicates adverse effects on women from ages 17–18, fading out by ages 19–20. Exposure to stereotyped evaluations during one high school grade significantly affects women's annual earnings (USD 3.6–8.9), contributing to around 8% of the gender pay gap at age 18–19 (ranging from USD 4.1 to USD 9.4 per month). From age 19–20 onward, total effects on girls' monthly earnings are negligible. Thus, women experience catch-up in earnings three years after high school, temporarily leaving most female graduates at a comparative disadvantage in labor markets. Conversely, exposure to a more stereotypical teacher increases men's earnings up to 5 years after graduation. Back-of-the-envelope calculations suggest men gain about 0.3% of the monthly minimum wage at ages 17–18 and 1.8% at ages 21–22 due to a teacher one-standard deviation more stereotypical in grading.[23]

I assess whether the impact on earnings is driven by low- or high-paying jobs, focusing on the likelihood of obtaining high-paying positions. Using the research design in Equation (5), I use an indicator variable (set to 1 when a student's monthly earnings exceed a specified percentile) to analyze effects on the probability of securing jobs with earnings above each percentile for 18- to 22-year-old workers. Figure 4 illustrates women's main and differential effects of increased exposure to teachers' stereotypical grading on the likelihood of attaining high-earning jobs. Panel (a) in Figure 4 shows that girls below the 50th percentile in monthly earnings are less likely to secure high-paying jobs with a teacher one-standard deviation more stereotypical at ages 18–19. According to panel (b), only the bottom 25 percentiles of girls continue to be affected by exposure to stereotypical practices at ages 19–20. By the fourth year post-graduation, gender gap effects and

[22] 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. Education statistics (2015–19) show women enrolled in college (ages 17–24) at 32%, slightly exceeding men at 29%.

[23] The average monthly minimum wage between 2015 and 2019 was USD 240.

primary effects appear to balance (panel (c)). Similarly, in the fifth post-graduation year, imprecise point estimates make it challenging to discern a clear pattern for the total effect on high earners (panel (d)).

Table VI: Monthly Earnings: Effects of Exposure to Mathematics Teachers' Stereotyped Assessments

	Age 17–18	Age 18–19	Age 19–20	Age 20–21	Age 21–22
	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.998***	-1.81***	-2.178***	-0.807	-2.115
	(0.185)	(0.415)	(0.742)	(1.137)	(2.178)
$\hat{\theta}^*_{j,-i}$	0.695***	1.07***	2.182***	0.024	4.368**
	(0.148)	(0.325)	(0.628)	(1.02)	(2.006)
Female	-4.074***	-9.408***	-16.409***	-20.482***	-14.756***
	(0.397)	(0.867)	(1.534)	(2.533)	(4.875)
$\bar{Y}$ female	218.2	214.6	216.5	224.2	229.8
$\bar{Y}$ male	253.1	253	255.3	259.8	259.8
$\bar{Y}$ female uncond.	0.9	3.9	8.4	12.3	14.9
$\bar{Y}$ male uncond.	3.6	10.9	19.7	25.9	29.4
Joint sig., p-val	0.004	0.002	0.995	0.344	0.176
R-squared	0.022	0.026	0.026	0.028	0.027
Observations	2,562,227	1,700,366	893,625	402,928	121,041

Note: This table presents estimated coefficients of the impact of increased gender-stereotypical teacher assessments in a specific high school grade on monthly earnings at ages 17–22 (post-high school graduation). The outcome variables consider the contract with the dominant annual employer for at least one month during the specified age range. All of the treatment variables, covariates, and joint significance terms are defined similarly to those in Table V. The p-value for the joint significance of the total effect on girls is reported. The unit of observation is the student-grade, clustered standard errors by school are reported in parentheses. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

These findings indicate that teachers' gender-stereotyped evaluations significantly contribute to the gender pay gap by guiding women towards low-paying jobs in the formal economy and steering men into high-paying positions. The substantial total exposure effects, accounting for around 8% of the pay gap, can be partly attributed to weak enforcement of recent legislation for equal pay between men and women in Peru. Weak legal protection, common in middle- and low-income countries, results in a significant gender earnings gap in both formal and informal labor markets. [24] In 2015, the raw monthly earnings gap between male and female workers was 29.9%, with an adjusted earnings gap in hourly wages at 19% (Muller and Paz, 2018). The lack of robust equal-pay enforcement allows earning disparities generated in schools to persist. In summary, inadequate labor rights protection fosters an environment where educators can guide students toward low-paying jobs, contributing to substantial wage disparities.

[24] In 2017, the Peruvian government passed laws to prevent pay discrimination, but judicial action against pay discrimination started only in 2021.

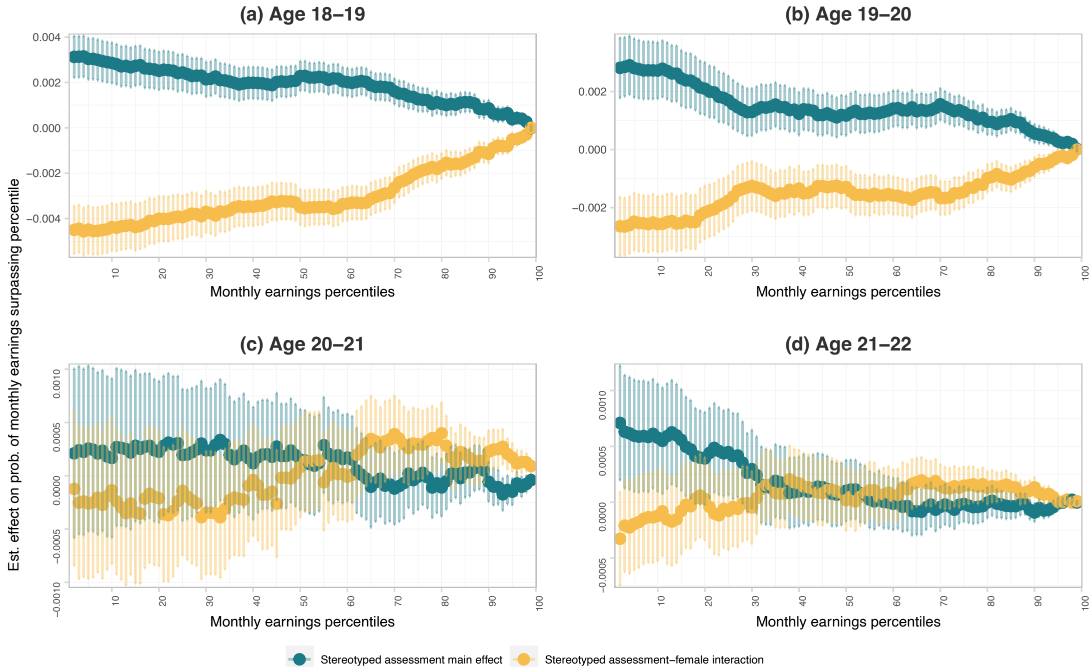

Figure 4: Effects on the Probability of Students' Earnings Exceeding a Percentile After High School Graduation (Ages 18–19)

Note: The graph displays the estimated effects of teachers' stereotypical assessments on the probability of student i's monthly earnings surpassing a given percentile, denoted $p \in (0, 100)$. Estimates are based on Equation (5), and percentiles are computed using the base sample's population of workers aged 16 to 25.

5.4 Beyond a Single Teacher: Effects of Average Exposure in High School

Figure 5 illustrates the impact on labor outcomes for individuals with varying average exposure to gender stereotyping assessments throughout high school across all grades.[25] The variable of interest is $\frac{1}{g}\sum_g \hat{\theta}^*_{j(g),-i}$, where $j(g)$ indicates the teacher assigned to student i in grade g. Panel (a) in Figure 5 shows that the overall impact of exposure to one standard deviation more stereotypical teacher across all high school grades is statistically significant and negative for 17-to-18-year-old girls (0.3 percentage points, equivalent to 10% of the mean). However, these effects become insignificant after this age. In contrast, a one-standard-deviation increase in average exposure to more stereotypical practices during high school has large positive effects on boys across all age groups. These effects range from 0.7 percentage points (8.8% of the mean) for boys aged 17–18 to 2.4 percentage points (9% of the group mean) for ages 21–22.

Panel (b) in Figure 5 suggests no conclusive evidence that female high school students exposed to more stereotypical assessment teachers have a statistically significant effect on working hours. However, average exposure effects on boys remain consistently positive, translating to an increase of up to 2.5 hours of work per month. Panel (c) in Figure 5, women exposed to teachers with one standard deviation higher incidence of stereotypical

[25] This estimation formally deviates from the assumption that students are as good as randomly assigned to one teacher; however, the reported effects are presented for policy relevance.

practices experience an annualized earnings loss of USD 8.4 at age 18–19 (2 years after graduation, equivalent to 18% of the mean) but are unaffected thereafter. Boys significantly benefit from increased high school exposure to stereotyped assessments, earning an additional USD 1.4 per month at age 17–18 and USD 4.7 per month at age 21–22. Overall, the effects appear substantial and favorable for men and neutral or slightly negative for women.

5.5 Academic Careers and High School Completion

Disparities in labor market outcomes are significantly influenced by high school completion. While graduating enhances employment prospects for boys, the connection between post-graduation work challenges for women and stereotypical assessment practices during their school years is mediated by female dropout rates. Female students assigned to math teachers with more stereotypical evaluations of girls' math performance are less likely to graduate high school on time or ever. Table VII presents grade-specific effects of exposure to math teachers with one standard deviation more stereotypical assessments against girls during one high school grade. The first panel shows the likelihood of graduating in the corresponding school year, while the second panel indicates the likelihood of graduating a calendar year after the projected date. The latter variable is defined differently for each cohort, ranging from one year after projected graduation for the youngest cohort (class of 2019) to four years for the oldest cohort (class of 2015) (see details in Appendix Table IX).

Columns (1) and (2) indicate a widening gender gap in high school graduation rates when 8th and 9th-grade students have math teachers with more stereotypical assessment practices. In 8th grade, the likelihood of a girl graduating drops by 5.2 percentage points for each standard deviation increase in stereotypical grading compared to boys according to the interaction coefficients in the first row. Parents receive a detailed report in 8th grade, comparing their child's performance in math, language arts, and science to the national average and students' performance by gender. This report may act as an ability signal, influencing parental encouragement for their children to pursue high school graduation. The results in the following columns suggest that having a more stereotypical teacher in 9th grade has a differential effect of 1.8 percentage points on girls relative to boys. However, this gap disappears by the 10th grade, suggesting that experiences in the first years of high school are more formative for students' long-term success. The point estimates follow a slightly decreasing trend over time as the students advance to more senior grades. Next, 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 stereotypical math teacher for one grade during high school, to the detriment of girls, widens the high school graduation gap by 1.5 percentage points (1.8% of the mean).

The results in the second panel analyze the outcome of ever graduating, including

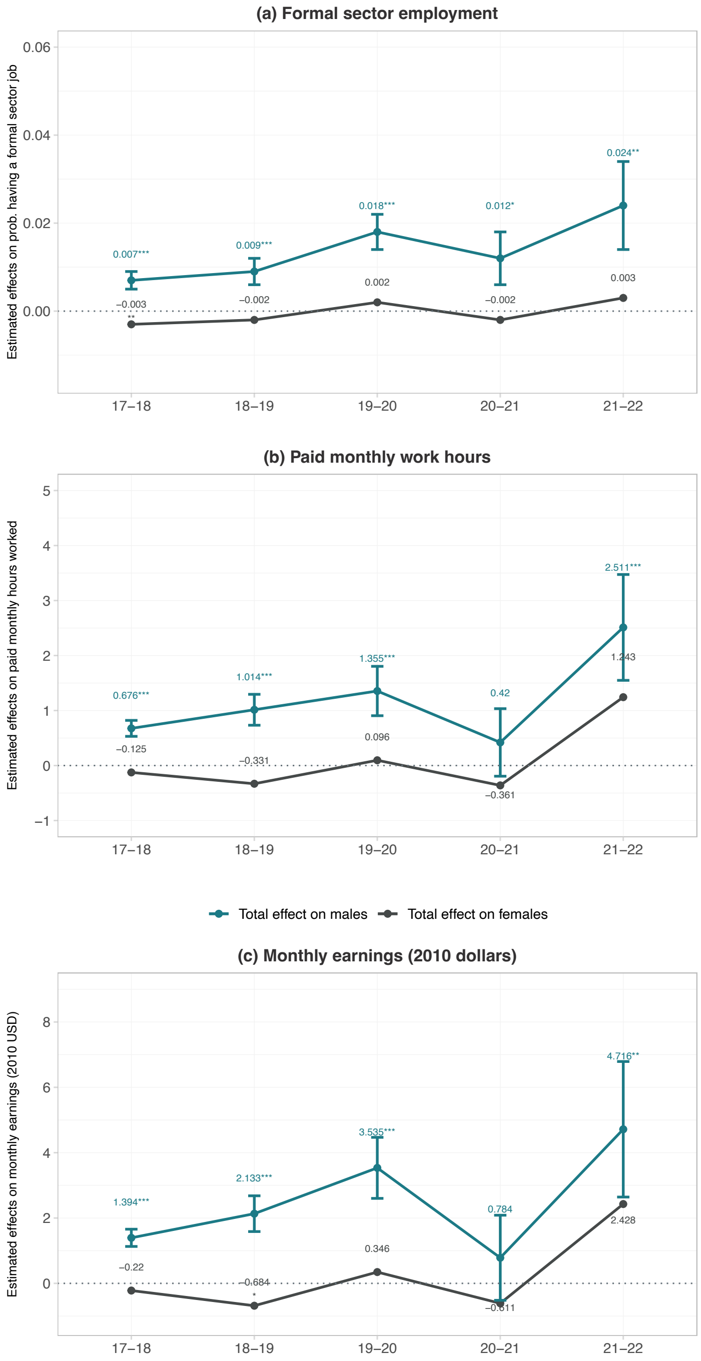

Figure 5: Effects of Average High School Exposure to Stereotypical Assessments on Student Outcomes

Note: The graph illustrates the mean effects of exposure to stereotypical math teachers on high school labor market outcomes, computed using Equation (5) with $\frac{1}{g}\sum_g \hat{\theta}^*_{j(g),-i}$ as the dependent variable. Error bars represent clustered standard errors by school for effects on males, with students as the unit of observation. The joint significance of total effects on females is denoted alongside the gray lines. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

late graduation. The summary of the point estimates is reported in column (1). In this case, a one-standard-deviation increase in stereotypical-evaluation practices is estimated to increase the gender gap in graduation disfavoring girls by 1.3 percentage points (1.5% of the mean). In contrast, for every one standard deviation that an educator's evaluation stereotypes are stronger than the average, boys' graduation rates improve by 0.8 percentage points (1% of the mean). The total effects on girls' high school graduation likelihood, calculated by summing main and interaction effects in the first two rows of stacked estimates in columns (4) and (8), are negative and statistically significant. This results in a reduction of on-time graduation likelihood by 0.6 percentage points (0.7% of the mean) and overall high school graduation by 0.5 percentage points (0.6% of the mean).

Finally, Appendix Table XVII demonstrates the stability of the coefficient, starting with only school by cohort fixed effects and progressively adding controls. Additionally, contrary to Lavy and Megalokonomou (2024), Appendix Table XVIII shows no significant impact of persistent stereotypical grading by math teachers on the gender gap in four-year college application, admission, or enrollment rates.

5.6 Robustness Checks

A potential identification threat arises if teachers' stereotypical evaluations of eighth-grade students differ from those in higher grades. This could occur if, for example, high school principals assign more conservative and stereotyping teachers to teach the upper grades of high school. To address this, I restrict the sample to graduating classes of 2018 and 2019, observed in both lower and upper grades. Appendix Table XIX shows statistically significant effects of math teachers' stereotyped assessments on graduation outcomes, mitigating concerns about this identification threat. Another concern is unobserved patterns in teacher assignments across grades. I examine this by defining a subsample of student cohorts with exposure to teachers' stereotypical grading throughout grades 7 to 11. Results in Appendix Tables XX do not indicate that this is a cause for concern.

Further evidence supporting the robustness of my results comes from language arts exposure. Unlike mathematics, language arts stereotyped grading doesn't involve negative stereotyping of students, and the parameter distribution is similar between math and language arts teachers (see Figure 2). Despite this, there is limited evidence that stereotyped grading in language arts directly affects academic performance for girls or boys across high school grades (Appendix Table XXI). Only exposure to more stereotypical language arts teachers in eighth grade significantly reduces the likelihood of high school graduation by 2.6 percentage points. Analyzing recurrent effects on girls' labor market outcomes (Appendix Tables XXII and XXIII), stereotypical language arts assessments impact the gender pay gap, paid hours up to a year post-graduation, and monthly earnings for up to two years post-graduation.

Finally, student tracking into sequences of more stereotypical teachers poses an

Table VII: High School Graduation Effects of Exposure to Mathematics Teacher Stereotyped Assessments

	8th grade (1)	9th grade (2)	10th grade (3)	All grades (4)
		A. Graduated on time		
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.052***	-0.018*	-0.001	-0.015***
	(0.010)	(0.009)	(0.007)	(0.004)
$\hat{\theta}^*_{j,-i}$	0.017	0.021*	0.003	0.009
	(0.011)	(0.012)	(0.012)	(0.005)
Female	0.021	-0.014	-0.005	-0.008
	(0.052)	(0.022)	(0.023)	(0.015)
$\bar{Y}$ female	0.754	0.793	0.858	0.824
$\bar{Y}$ male	0.683	0.732	0.808	0.772
Joint sig., p-val	0.000	0.599	0.625	0.012
R-squared	0.323	0.360	0.411	0.411
		B. Graduated ever		
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.048***	-0.019**	0.002	-0.013***
	(0.010)	(0.009)	(0.007)	(0.004)
$\hat{\theta}^*_{j,-i}$	0.024**	0.021*	-0.001	0.008
	(0.011)	(0.012)	(0.012)	(0.006)
Female	0.028	-0.022	-0.016	-0.006
	(0.039)	(0.028)	(0.017)	(0.014)
$\bar{Y}$ female	0.784	0.830	0.895	0.855
$\bar{Y}$ male	0.743	0.803	0.878	0.831
Joint sig., p-val	0.001	0.735	0.771	0.050
R-squared	0.257	0.265	0.242	0.267
Observations	579,867	580,869	675,845	2,562,561

Note: This table presents estimates of the impact of one-grade exposure to a math teacher's stereotypical assessment practices on high school graduation, both on time and ever. The analysis includes 8th–10th graders with projected graduation years between 2015 and 2019. The outcome variables consider the contract with the dominant annual employer for at least one month during the specified age range. Student, teacher, lagged scores, classroom- and school-grade-level controls, place-of-birth fixed effects and gender interaction are covariates. Grade-specific estimations (Columns (1)–(3) and Columns (5)–(7)) treat the student as the observation unit, with cohort, year, and school fixed effects and school-level clustered standard errors. Stacked grade estimations (Columns (4) and (8)) use the student grade as the observation unit, employing a sample of students stacked across grades and including grade, cohort, year, and school fixed effects, along with two-way clustered standard errors by student and school. The p-value for the joint significance of the total effect on girls is reported. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

identification threat. Early exposure to stereotypical grading may lead to negative impacts on performance and placement with lower value-added teachers with stronger stereotypical beliefs. In Peru, the absence of Advanced Placement classes reduces this concern, but tracking based on performance remains possible. Columns (1) to (3) in Appendix Table XXIV displays a trend of growing exposure to stereotypical evaluation practices impacting girls across grade levels, mainly attributed to recurrent teacher assignments.[26] Columns (4) through (6) indicate that restricting the sample to students with different teachers each year eliminates statistically significant differential tracking evidence.

6 Discussion of Mechanisms

6.1 Teachers' Stereotyped Assessments and Value Added

The mechanism receiving significant attention in the impact of teachers' gender stereotypes is its direct influence on student achievement, serving as a foundation for subsequent educational careers (Fryer and Levitt, 2010, Hyde and Mertz, 2009, **?**). Recent studies, such as Alesina et al. (2018), Carlana (2019), Lavy (2008), Lavy and Sand (2018), explore how gender stereotypes and biases contribute to achievement gaps. Carlana (2019) highlights the impact of teachers' implicit stereotypes, potentially affecting female students' choices in high school tracks. Similar effects are observed in literature examining the influence of racial and ethnic stereotypes on educational performance (Botelho, Madeira and Rangel, 2015, van den Bergh et al., 2010).

Appendix Table XXV outlines the effects of teachers' stereotyped assessments on gender score gaps in math and language arts. In eighth grade, gender-stereotyped assessments differentially decrease girls' scores by 0.26σ and 0.01σ in the senior grade. In language arts, teachers with stereotyped assessments lower students' scores by 0.013σ–0.041σ. To explore potential differential value-added, I calculate $\alpha_{1,j}$ as teacher j's value added toward girls and $\alpha_{1,j} + \alpha_{2,j}$ toward boys using Equations (2) and (3). Fixed-effects–SURE specification and joint covariance matrix calculations are employed for a formal analysis, essential for bias correction using sampling covariances derived from SURE estimation.

Bias-corrected estimates of the correlations in value-added and bias parameters across teachers are in Appendix Table XXVI (details in Online Appendix B). Consistent with Lavy and Megalokonomou (2024), math teachers with more stereotypical grading have lower value-added scores for both genders. Correlations in Appendix Table XXVI's Panel A (columns (2)–(3)) show a stronger negative association with value added toward boys, consistent in Panel B among language arts teachers. This suggests low-value-added teachers tend to have stronger stereotypes, emphasizing teachers' quality as a crucial mechanism in stereotypical assessments affecting high school completion and labor market outcomes.

6.2 Students' Internalization of Ability Stereotypes

I argue that confirming ability stereotypes —about what girls are expected to excel at and what they are expected to struggle with— through stereotyped grading may reinforce

[26] In this context, 8%-12% of students have the same teacher across multiple grades, contributing to repeated teacher assignments.

harmful gender stereotypes, leading to "internalized stereotypes" among boys and girls (David, Schroeder and Fernandez, 2019, Jones, C. P., 2000). Educational psychology studies show that teachers' stereotypes affect students' self-assessment and academic self-concept development (Ertl, Luttenberger and Paechter, 2017). Exley and Kessler's (2022) research indicates that female workers rate their performance lower in traditionally male-dominated fields. I test the mechanism using IAT scores from younger high school students expected to graduate between 2022 and 2025.[27] The estimation sample comprises 1,153 students matched with teachers with available IAT-based measures of stereotypes regarding students' abilities. Selection criteria for in-person sessions included school location, principal consent, and verified computer labs. Sessions, conducted during classroom time, involved educators with relevant qualifications.

I test whether students internalize teachers' gender stereotypes measured by the IAT, IAT_j, according to the following regression equation:

$$IAT_i = \tilde{\nu}_0 + \tilde{\nu}_1 \cdot IAT_j\, Female_i + \tilde{\nu}_2 Female_i + \tilde{\nu}_3 IAT_j + \tilde{\nu}_4 \mathbf{X}_i + \tilde{e}_i \qquad (6)$$

Table VIII: Mechanism Analysis of Internalized Math Teachers' Stereotyped Assessments and Student Survey-Collected Measures

		Reported employment prob.			Reported general interest		
	Student IAT	Eng.	STM	SS	Eng.	STM	SS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$IAT_j \cdot$ Female	0.136*	-0.034	-0.021	0.020	-0.010	-0.074	0.013
	(0.064)	(0.036)	(0.06)	(0.036)	(0.037)	(0.056)	(0.019)
IAT_j	-0.128*	0.036	0.021	-0.021	0.007	0.077	-0.012
	(0.066)	(0.037)	(0.058)	(0.035)	(0.037)	(0.057)	(0.018)
Female	0.093*	-0.033	-0.006	0.015	-0.023	-0.083*	-0.027*
	(0.046)	(0.03)	(0.041)	(0.024)	(0.029)	(0.038)	(0.014)
$\bar{Y}$ female	0.093	0.158	0.127	0.090	0.158	0.127	0.090
$\bar{Y}$ male	0.065	0.184	0.132	0.074	0.184	0.132	0.074
R-squared	0.022	0.011	0.005	0.004	0.009	0.022	0.009
Observations	1,153	1,153	1,153	1,153	1,153	1,153	1,153

Note: This table examines the association between a one-standard-deviation increase in teachers' IAT score, measuring gender bias against girls, and their students' Implicit Association Test. The sample consists of students with projected high school graduation years from 2020 to 2025, and both students' and teachers' IAT scores were standardized. Controls include student gender, age, self-reported number of previous IAT, IAT question order, and teachers' gender, age, and children status. Observations are at the student level, with grade and school fixed effects, and standard errors are clustered at the school level. STM = Science, Technology and Mathematics; SS= Social Sciences. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table VIII, Column (1), indicates that exposure to stereotyped teachers causes female students to internalize negative gender-based stereotypes in math classes, with girls' IAT scores increasing by 0.01σ when taught by a math teacher with an IAT one standard

[27] See descriptive statistics in Appendix Table XXVII and data collection timeline in Appendix Figure 11.

deviation higher than the average. This suggests a small negative impact on their self-perceptions of math and science abilities. Conversely, boys with teachers having higher IAT values show weaker associations with math and science, resulting in a decline of 0.13σ in their IAT scores. In line with social psychology literature that finds weak linkages between IAT scores and self-reported agreement with gender-related opinions (Egloff and Schmukle, 2002), columns (2) through (4) reveal no noticeable effects of stronger ability stereotypes on male or female students' self-reported beliefs about working in STEM versus social sciences. Additionally, in columns (6) through (7), exposure to teachers with stronger ability stereotypes does not significantly impact students' self-reported interest in STEM majors compared to social sciences. This underscores the influence of educators' stereotypes on adolescents' choices and domain-specific adequacy, supporting the earlier finding that stereotypically assessed teachers' value-added negatively affects girls less than boys.

6.3 Gender-Based Industry Sorting

Research on the gender wage gap underscores the significant contribution of occupational segregation to its persistence (Cortes and Pan, 2018, Kunze, 2018). Gender differences in occupational sorting, particularly impacting women without a four-year college degree, contribute to wage disparities (Blau and Kahn, 2017, Blau, Brummund and Liu, 2012). While Goldin (2014) notes favorable payment convergence in high-wage distribution occupations, there is limited understanding of factors influencing earnings gaps at the lower wage distribution tail, where high school graduates are prevalent. This section addresses this knowledge gap.

Figure 6 indicates that industry sorting is a probable channel for educators' impact on earnings trajectories among young low-skill workers in Peru. The longer a woman is in the workforce post high school, the more discouraged she is from entering male-dominated industries due to earlier exposure to stereotyped teachers. In Panels (a) and (b), the figures on the right indicate that teachers' stereotypical grading discourages women while encouraging men to pursue work in male-dominated industries at ages 18–19 and 21–22. Notably, the left panel demonstrates that teachers' stereotypical grading has little differential effect on women's participation in female-dominated industries such as 'finance and insurance," "educational services," "health care and social assistance," and "accommodation and food services." Similarly, it does not significantly impact men's participation in traditionally female-dominated industries. These results suggest that exposure to stereotypical teacher assessments causes significant sorting effects between female and male students, influencing the composition of workers in industries and contributing to the gender pay gap.

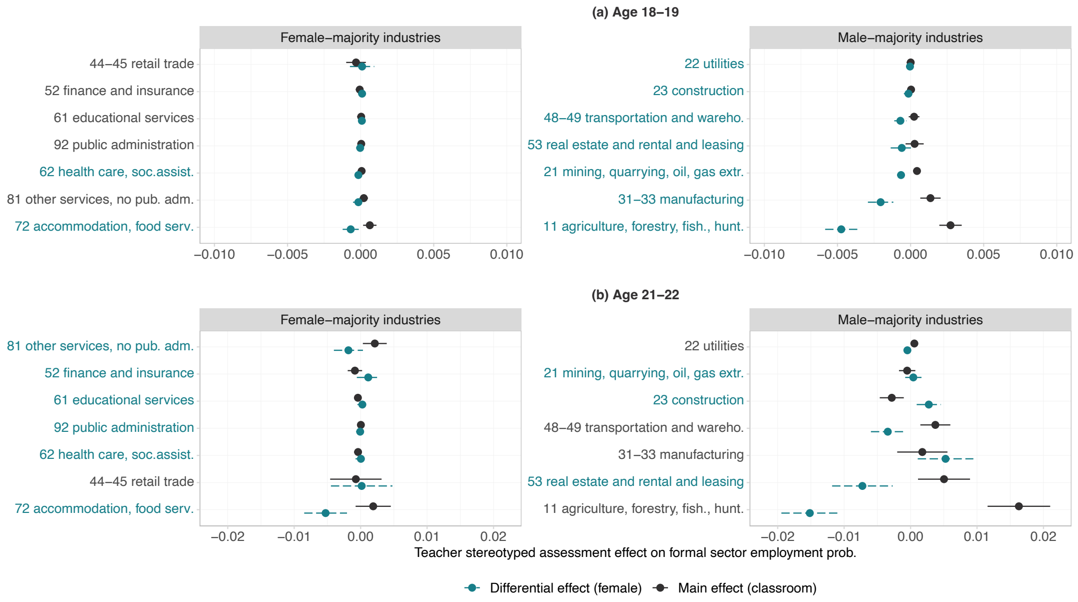

Figure 6: Effects on Probability of Employment in the Formal Sector by Female- vs. Male-Dominated Industries

Note: This graph illustrates the estimated impact of teachers' stereotyped assessments on the likelihood of formal employment in specific industries (k), post-graduation. Using the same research design and covariates as in Equation (5), each coefficient pair corresponds to a distinct regression. The left panels depict effects on female-dominated industries (with a higher proportion of 18–35-year-old female workers), while the right panels focus on male-dominated industries based on the pooled National Household Survey from 2015 to 2019. Industries that negatively affect teachers' stereotypical assessments of women are highlighted in green.

7 Conclusions

I examined the lasting impact of stereotypical teacher assessments on the formal employment and earnings of 1.6 million students from five graduating cohorts in Peruvian public high schools. Using nearly a decade of rich longitudinal data from administrative sources, I tracked students from 12 in eighth grade to 22 in higher education or the workforce. The study revealed a significant correlation between teachers' gendered grading practices and implicit gender-based ability stereotypes measured by the IAT. Exposure to stereotypical teacher evaluations intensifies gender gaps in earnings and formal employment up to five years post-high school graduation. Boys experience positive effects, while girls face small, statistically insignificant detrimental impacts. The consequences for females dissipate by the third year post-graduation, while positive effects persist for males over five years. Influenced by gender biases, high school dropout rates, and industry sorting, teachers' stereotypical assessments emerge as an undocumented source of gender gaps in earnings. Novel evidence suggests internalized ability stereotypes affecting students' views of math and science proficiency, and requirest more exploration. Given the lasting impact, policy interventions are crucial, including education for teachers and students to identify and mitigate biases. Implementing interventions like blind grading and educator training is essential to counteract adverse effects.

References

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

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

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

Alan, Sule, Seda Ertac, and Ipek Mumcu. 2018. “Gender Stereotypes in the Classroom and Effects on Achievement.” *The Review of Economics and Statistics*, 100(5): 876–890.

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

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

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

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

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

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

Biasi, Barbara, and Heather Sarsons. 2020. “Flexible Wages, Bargaining, and the Gender Gap.” National Bureau of Economic Research w27894, Cambridge, MA.

Bietenbeck, Jan. 2020. “The Long-Term Impacts of Low-Achieving Childhood Peers: Evidence from Project STAR.” *Journal of the European Economic Association*, 18(1): 392–426.

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

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

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

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

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

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

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

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

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

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014*b*. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633–2679.

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas*." *The Quarterly Journal of Economics*, 129(4): 1625–1660.

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

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

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

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

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

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

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

Efron, Bradley. 2016. "Empirical Bayes deconvolution estimates." *Biometrika*, 103(1): 1–20.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Greenwald, Anthony G., Brian A. Nosek, Mahzarin R. Banaji, and K. Christoph Klauer. 2005. "Validity of the salience asymmetry interpretation of the Implicit Association Test: Comment on Rothermund and Wentura (2004)." *Journal of Experimental Psychology: General*, 134(3): 420–425.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lavy, Victor, and Rigissa Megalokonomou. 2024. "The Short- and the Long-Run Impact of Gender-Biased Teachers." *American Economic Journal: Applied Economics*, 16(2): 176–218.

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

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

Minedu. 2016. "Currículo Nacional de la Educación Básica."

Ministry of Labor of Peru. 2019. "Annual report of women in the workplace, 2020." Lima, Peru.

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

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

Nations, United. 2023. "Global Database on Violence against Women."

Niederle, M., and L. Vesterlund. 2007. "Do Women Shy Away From Competition? Do Men Compete Too Much?" *The Quarterly Journal of Economics*, 122(3): 1067–1101.

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

Nosek, Brian A., Anthony G. Greenwald, and Mahzarin R. Banaji. 2007. "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., 265–292. New York, NY, US:Psychology Press.

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

OECD. 2022. "Gender wage gap."

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

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

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

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

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

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

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

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

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

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

Terrier, Camille. 2020. "Boys lag behind: How teachers' gender biases affect student achievement." *Economics of Education Review*, 77: 101981.

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

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

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

World Bank Group. 2017. *Peru Systematic Country Diagnostic.* World Bank, Washington, DC.

A Appendix A. Data Appendix

A.1 Sample construction

A.1.1 Longitudinal Student Sample

The study focuses on the outcomes of high school completion, college application, admission, and enrollment for the period of 2015–20. The linkage procedure for identifying public high school students on college records uses the unique anonymized student ID provided by the Ministry of Education. The matched employer-employee data set is processed in several steps to obtain outcome variables of employment in the formal sector, real earnings, and work hours. The data is then computed using the monthly exchange rate between PEN and USD and the CPI published by the Central Bank of Peru and the Bureau of Labor Statistics. The study considers three institutional background considerations: workers can have multiple employers providing additional sources of income that are not sustained over the whole calendar year, such as seasonal, extraordinary, or temporary jobs. Additionally, the employer-employee data set records earnings monthly, requiring an annualization procedure to avoid including earnings or hours measured with registry errors or misreported.

The study uses a procedure similar to those used by Abowd, Lengermann and McKinney (2003), Lachowska, Mas and Woodbury (2020) and Sorkin (2018) to determine the primary annual employer and associated earnings and hours worked during job tenure. The primary employer is determined for each quarter within each calendar year, preventing the inclusion of monthly administrative errors or contract status changes. The resulting annualized variables allow for the construction of outcome variables such as monthly earnings, working hours, annualized earnings, and log earnings. I code the quarters as *full* or *continuous* following Sorkin (2018), and I use the same procedure to define an annualized amount of earnings and hours worked. The resulting annualized variables allow me to construct the following outcome variables: (i) monthly earnings (in 2010 USD), which are composed of the workers' base salary and any supplemental salaries (for example, overtime, bonuses, and commissions); (ii) monthly working hours, which are contractual hours of work; (iii) annualized earnings, including supplemental salaries; and (ii) log earnings.

The sample is determined by a procedure that starts with the main student-teacher classroom assignments. Using Rothstein (2010, 2017), I limit student-teacher matches based on these criteria: I drop special education, multilingual or bilingual classes that teach in native languages, and other nonregular classes (border schools or public schools run by the military). I also drop classes with more than 60 sample students and classes from a sample school. Teachers with more than 200 pupils in a grade and numerous schools in a school year are excluded. Students with valid teacher matches whose stereotypical evaluations are not absent for all grades in the sample, individually for mathematics, language, and science, are kept. For 2014 cohorts and grades below eighth grade, I drop pupils without lagging scores. I only sample cohorts I to V that have long-term outcomes since the time span over which I can examine outcome variables varies.

After joining long-term outcomes data, I apply constraints. Following Card, Heining and Kline (2013), I remove a person's history if they meet any of the following criteria: (i) if their contract's starting date is empty, inconsistent with the matched employer-employee's date of birth or date of registered work, or inconsistent across observations; (ii) if the log hourly wage change from one quarter to the next is greater than one. Workers between 16 and 25 with high school, technical, college, or other higher education were kept in the sample. This filter is imposed because the oldest cohort (Cohort I in the second panel of Table IX), which we can observe up to five years after high school graduation, would graduate from college in the last year we observed the matched employer-employee records, but they would not be registered until the first semester of 2021.

A.1.2 Teacher-Student Matched Sample

Eighth-grade students' standardized exams and classroom test scores are used to estimate teacher-level stereotypical evaluations. The data set includes 2015, 2015, 2018, and 2019 Student Census Assessment (ECE) blindly scored test results. The cross section with student-subject scores is further restricted using the aforesaid criteria to depurate student-teacher classroom assignments. I especially drop special-education, bilingual, and other non regular classes, students from classrooms with more than 60 students or fewer than 5 students, single education classes because gender score gaps cannot be calculated, teachers who have more than 200 students in a grade for a given school year, and students without classroom test scores for the year.

A.2 Surveys

Teachers' Survey. Before data collection, instruments and website functionality underwent pilot testing with around 30 teachers through individual interview sessions. Due to the COVID-19 pandemic, in-person school activities were limited during data collection. The sampling frame included teachers whose contact information was obtained from Ministry of Education records. The Ministry sent text messages inviting teachers to participate, followed by email reminders for unfinished activities (Implicit Association Test and survey). Online informational meetings were conducted at the start of the 2022 school year for sampled teachers and principals. Teachers received two data collection instruments: an Implicit Association Test and a Questionnaire.

The web-based IAT used in the study was pre-tested in Spanish for readability, clarity, and functionality with 26 teachers in Lima public high schools. The IAT includes seven blocks of associations, each with 20 (practice) or 40 (scored) associations. In the

Table IX: Distribution of Student Cohorts in Estimation Samples

Cohort	High school grades					Projected graduation year	Sample
	7th	8th	9th	10th	11th		
I				░	░	2015	F
II			░	░	░	2016	F
III		░	░	░	░	2017	F
IV	▓	▓	▓	▓	▓	2018	F, B
V	▓	▓	▓	▓	▓	2019	F,B

Cohort	Projected graduation Year	Employer-employee records				
		2016	2017	2018	2019	2020
I	2015	░	░	░	▓	▓
II	2016		░	░	░	▓
III	2017			░	░	░
IV	2018				░	░
V	2019					░

Note: This table depicts the sample used for estimating effects on high school completion (first panel) and labor market outcomes (second panel). Colored cells represent grades with observable exposure to stereotyped teacher assessments for analysis. F = Full sample in main specifications, B = Balanced by cohort sample in robustness checks. The second panel illustrates the post-high school graduation years for evaluating the effects of stereotypical grading practices on labor market outcomes. Each colored cell represents a year of available information.

Table X: Association blocks of gender-science IAT

Block	Num. associations	Function	Items, left-key response	Items, right-key response
1	20	Practice	Male	Female
2	20	Practice	Science	Humanities
3	20	Practice	Science-Male	Humanities-Female
4	40	Test	Science-Male	Humanities-Female
5	20	Practice	Humanities	Science
6	20	Practice	Humanities-Male	Science-Female
7	40	Test	Humanities-Male	Science-Female
	200	Total		

In this version of the Implicit Association Test (IAT) compatible order, the following terminology is used: Target A = Male, Pos = Sciences, Target B = Female, Neg = Humanities. The association for a stronger implicit preference for Male in science in relation to Female in science is given by B6-B3 and B7-B4.

data collection version, teachers must correctly associate terms before proceeding. An example sequence, following Greenwald, McGhee and Schwartz's (1998) notation, is presented in Table X, with target words (male and female) and attributes (science and humanities). The words for association pairs in the Gender-Science IAT are in Table XI, and for the Gender-Career IAT, refer to Table XII. The analysis categorizes IAT results into: *preference for girls*, *little to no stereotypes*, *slight stereotypes*, *moderate to severe stereotypes*, and *strong stereotypes*, using break points proposed by Greenwald, Nosek and Banaji (2003), Greenwald et al. (2005) and the improved scoring algorithm by Greenwald, Nosek and Banaji (2003), Lane et al. (2007).

The questions to measure gender attitudes that are used in this study are the following:

1. On a scale of 1 (=Strongly disagree) to 4 (=Strongly agree), how much do you agree with the following statements.[28]

[28] In all Likert scale questions, the options are "Strongly disagree", "Disagree", "Agree", and "Strongly

Table XI: Concepts and items of gender-science IAT

Concepts	Items
Humanities	Psychology, Philosophy, Humanities, Literature, History, Education, Arts
Sciences	Biology, Physics, Chemistry, Mathematics, Astronomy, Engineering, Geology
Male	Man, Him, Men, Boy, Gentleman, Father, Son, Husband
Female	Woman, Her, Women, Girl, Lady, Mother, Daughter, Wife

From the original set of items in the gender-science Implicit Association Test (IAT) proposed by Greenwald, McGhee and Schwartz (1998), I replaced the item “Music” with “Psychology,” as it was unequivocally recognized as part of the “Humanities” category during the pilot and follow-up focus-group sessions for instrument testing with the pilot’s participants.

Table XII: Concepts and items of gender-career IAT

Concepts	Items
Family	Garden, Kitchen, Marriage, Laundry, Home, Children, Family
Career	Office, Manager, Salary, Job, Business, Profession, Employees
Male	Man, Him, Men, Boy, Gentleman, Father, Son, Husband
Female	Woman, Her, Women, Girl, Lady, Mother, Daughter, Wife

Note: The set of items corresponds to the set of items in the gender-career Implicit Association Test (IAT) proposed by Greenwald, McGhee and Schwartz (1998).

Statements: When a mother has wage labor, her children suffer; It is more important that boys perform better in school than girls; Generally speaking, men are better political leaders than women; A college education is more important for a man than a woman; Generally speaking, men are better for business than women; Being a homemaker is as satisfying as working for a salary; Men are more intelligent for science and technology careers than women.

2. How much do you agree with the opinion of some teachers who say that women have the same potential as men for the following careers?

   Careers: Systems and technology engineering, Law and Political Science, Marketing y Business, Psychology and Education, Social Sciences, Statistics and Informatics, Gastronomy and culinary arts, Accounting and Financial Sciences, Arts and Design

3. The following statements describe attitudes that some teachers have about the role of women and men in society. How much do you agree with each of the following statements? There are no right or wrong answers, just different opinions.

   Statements: When jobs are scarce, men have more rights to a job than women; When jobs are scarce, employers must prioritize Peruvians over immigrants; It is problematic when women earn more money than their husbands; Gay couples are just as good parents as other couples; It is a duty towards society to have children; Adult children have the duty to provide long-term care for their parents; People who don not work become lazy; Work is a duty towards society; Work should always come first, even if it means less free time.

agree”.

Students' Survey. Data, collected in Metropolitan Lima through remote and in-person sessions, include 5,139 students in grades 8–11 from 266 classrooms. Educators led workshops, ensuring student participation, and administered the gender-science IAT, adapted and previously tested for 13-to-16-year-olds.

1. How interesting do you find the following majors, on a scale of 0 to 10 (including those numbers)? Respond by rating one of the options from 0 to 10, with 0 meaning "No chance" and 10 meaning "Very high chance." (Same Career options as above).

2. How many chances do you believe the following majors will give you, on a scale of 0 to 10 (including those figures), to land a job? Respond by rating one of the options from 0 to 10, with 0 meaning "No chance" and 10 meaning "Very high chance." (Same Career options as above).

3. On a scale of 1 (=Strongly disagree) to 4 (=Strongly agree), how much do you agree with the following statements.

   Statements: When a mother has wage labor, her children suffer; It is more important that boys perform better in school than girls; Generally speaking, men are better political leaders than women; A college education is more important for a man than a woman; Generally speaking, men are better for business than women; Being a homemaker is as satisfying as working for a salary; Men are more intelligent for science and technology careers than women; Boys and girls have the same talent for mathematics; In Peru, men can work in nursing.

B Appendix B. Technical Appendix

B.1 Omitted Proof of Proposition 1

Replacing equation determining ψ_j on the $E(S^T_{ij}|G_{ij}, W_{ij}, \tilde{\psi}_{ij})$,

$$\begin{aligned} E(S^T|G_i, W_{ij}, \tilde{\psi}_{ij}) &= \beta + \beta_{g,j}G_{ij} + W'_{ij}\beta_{w,j} + \beta_{\psi,j}(\lambda_j + \lambda_{\psi,j}\psi_i + \lambda_{\vartheta,j}\vartheta_{h(i),j}) \\ &= \beta_j + (\beta_{g,j}G_{ij} + \beta_{\psi,j}\lambda_{\vartheta,j}\,\vartheta_{h(i),j})G_{ij} + W'_{ij}\beta_{w,j} \\ &= \beta_j + [\beta_{g,j} + \tilde{\beta}_j\vartheta_{h(i),j}]G_{ij} + W'_{ij}\beta_{w,j} \end{aligned}$$

where in the second line I replaced $\beta_j = \beta + \beta_{\psi,j}(\lambda_j + \lambda_{\psi,j}\psi_i)$ and in the last line I denote $\tilde{\beta}_j := \beta_{\psi,j}\lambda_{\vartheta,j}$. Similarly, replacing $\alpha_j = \alpha + \alpha_{\psi,h}\psi_i$ in $E[S^B_{ij}|G_{ij}, W_{ij}, \psi_{ij}]$, I determine, $E[S^B_{ij}|G_{ij}, W_{ij}, \psi_{ij}] = \alpha_j + \alpha_{g,j}G_{ij} + W'_{ij}\alpha_{w,j}$. If $\beta_{\psi,j} \in (0,1)$, teachers assign higher grades to students with higher forecasted ability, $\tilde{\beta}_j > 0$. Next, I define the difference in difference parameter with the following terms, $E(S^T|G_{ij} = 1, W_{ij}) - E(S^T|G_{ij} = 0, W_{ij}) = \beta_{g,j} + \tilde{\beta}_j\tilde{\vartheta}_j$, and, $E(S^B|G_{ij} = 1, W_{ij}) - E(S^B|G_{ij} = 0, W_{ij}) = \alpha_{g,j}$ resulting in, $\underbrace{(\alpha_{g,j} - \beta_{g,j})}_{:=\Delta_j} + \underbrace{\tilde{\beta}_j\tilde{\vartheta}_j}_{:=\theta_j}$ Assuming uniform influence of non-stereotype-driven gender differences in assessments ($\alpha_{g,j}$ and $\alpha_{g,j}$) across all classrooms of teacher j leads to $\Delta_j = 0$.

B.2 Benchmark Definition of Non-stereotyped Teacher

Definition 2. *Let ϑ be distributed according to the cumulative distribution function $\tilde{G}$ given that the student is a member of any group $h, h' \in \mathcal{G}$. A teacher does not hold stereotypes toward groups h and h' if the feedback they assign to any student i is drawn from a cumulative distribution function $\tilde{G}$ that is the same for each group h, h'.*

B.3 Fixed-Effect–SURE Estimation

In matrix notation, I can write Equations (2) and (3) as $\mathbf{S}_{ij} = \mathbf{X}_{ij}\mathbf{A}_j + \mathbf{U}_{ij}$, where $\mathbf{S}_{ij} = (S^B_{ij}, S^T_{ij})'$, $\mathbf{X}_{ij} = diag(\mathbf{X}_{1,ij}, \mathbf{X}_{2,ij})$ with $\mathbf{X}_{1,ij} = (\mathbf{1}_{ij}, G_i)', \mathbf{X}_{2,ij} = (\mathbf{1}_{ij}, G_i)'$, $\mathbf{1}_{ij}$ is an takes the value of one when teacher j teaches student i, G_i is an indicator variable taking value of 1 when the student i is male, and $\mathbf{u}_{ij} = (\eta_{ij}, \epsilon_{ij})'$. The teacher parameter vector $\mathbf{A}_j \equiv (\alpha_j, \beta_j)$ with $\alpha_j := (\alpha_{1,j}, \alpha_{2,j})'$ and $\beta_j := (\beta_{1,j}, \beta_{2,j})'$ with $\alpha_j, \beta_j \in \mathbb{R}, j \in \{1, \ldots, J\}$. Moreover, I assume that $\mathbb{E}[\mathbf{u}_{ij}\mathbf{u}'_{ij}|\mathbf{X}_{it}] \equiv V_j > 0$, where V_j is a 2×2 covariance matrix.

After estimating the teacher parameters by generalized least squares, I have the following $(\hat{\alpha}_j, \hat{\beta}_j) \sim \mathcal{N}((\alpha_j, \beta_j), \mathbf{\Sigma}_j)$. The bias-correct variance estimator of $(\hat{\alpha}_j, \hat{\beta}_j)$ is a matrix $\hat{\Omega} = J^{-1}\sum_j \left[((\hat{\alpha}, \hat{\beta}_j) - (\bar{\alpha}_j, \bar{\beta}_j))((\hat{\alpha}_j, \hat{\beta}_j) - (\bar{\alpha}_j, \bar{\beta}_j))' - \Sigma_j)\right]$. A feasible estimator of $\mathbf{\Sigma}_j, j \in \{1, \ldots, J\}$ under conditional homoskedasticity is as follows:

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_j =& \left\{ |\hat{V}_j|^{-1} \begin{bmatrix} \hat{\sigma}^2_{\eta_j} \mathbf{X}^2_{2,\cdot j} & -\hat{\sigma}_{\eta_j \epsilon_j} \mathbf{X}_{1,\cdot j} \mathbf{X}_{2,\cdot j} \\ -\hat{\sigma}_{\eta_j \epsilon_j} \mathbf{X}_{1,\cdot j} \mathbf{X}_{2,\cdot j} & \hat{\sigma}^2_{\epsilon_j} \mathbf{X}^2_{1,\cdot j} \end{bmatrix} \right\}^{-1} \\
=& \left\{ |\hat{V}_j|^{-1} \begin{bmatrix} \hat{\sigma}^2_{\eta_j} \mathcal{D} & -\hat{\sigma}_{\eta_j \epsilon_j} \mathcal{F} \\ -\hat{\sigma}_{\eta_j \epsilon_j} \mathcal{F}' & \hat{\sigma}^2_{\epsilon_j} \mathbf{1}_{\cdot j} \end{bmatrix} \right\}^{-1} \qquad (7)
\end{aligned}$$

Let N_j denote the number of students being taught by teacher j—that is, $N = \sum_{j=1}^{J} N_j$—so that *within*-teacher variance estimators take the form $\hat{\sigma}^2_{\eta_j} = N^{-1} \sum_{i=1}^{N_j} \hat{\eta}_j$ with *within* residuals $\hat{\eta}_j, \hat{\epsilon}_j$ calculated for teacher j. $\mathcal{D} = \begin{pmatrix} \mathbf{1}_{\cdot j} & G_i \\ M_{\cdot j} & M^2_{\cdot j} \end{pmatrix}$, $\mathcal{F} = \begin{pmatrix} \mathbf{1}_{\cdot j} & M_{\cdot j} \end{pmatrix}$ with subscript $\cdot j$ indicating calculations *within* teacher j. Based on elements of $\hat{\Omega}$, it can be shown that the correlations of interest, $Corr(\hat{\alpha}_j, \hat{\beta}_j)$, can be written as follows:

$$Cov(\hat{\alpha}_{1,j}, \hat{\beta}_{1,j}) = J^{-1} \sum_j \left[\hat{\sigma}_{\hat{\alpha}_{1,j} \hat{\beta}_{1,j}} + \mathcal{C}\, \hat{\sigma}_{\eta\epsilon} G_i^2 \hat{\sigma}_{\eta_j} (1 + \mathbf{1}_{\cdot j}) \right] \qquad (8)$$

$$Cov(\hat{\alpha}_{2,j}, \hat{\beta}_{1,j}) = J^{-1} \sum_j \left[\hat{\sigma}_{\hat{\alpha}_{2,j} \hat{\beta}_{1,j}} + \mathcal{C}\, \hat{\sigma}_{\eta\epsilon} G_i^2 \hat{\sigma}_{\eta_j} (1 + \mathbf{1}_{\cdot j}) \right] \qquad (9)$$

Equations (8) and (9) are analogous for $\hat{\beta}_{2,j}$. Moreover, the variances are as follows:

$$Var(\hat{\alpha}_{1,j}) = J^{-1} \sum_j \left[\hat{\sigma}^2_{\hat{\alpha}_{1,j}} - \mathcal{C}\, G_i^2 (\hat{\sigma}^2_{\eta_j} \hat{\sigma}^2_{\epsilon_j} \mathbf{1}_{\cdot j} - \hat{\sigma}^2_{\eta_j \epsilon_j}) \right] \qquad (10)$$

$$Var(\hat{\alpha}_{2,j}) = J^{-1} \sum_j \left[\hat{\sigma}^2_{\hat{\alpha}_{2,j}} - \mathcal{C}\, G_i^2 (\hat{\sigma}^2_{\eta_j} \hat{\sigma}^2_{\epsilon_j} \mathbf{1}_{\cdot j} - \hat{\sigma}^2_{\eta_j \epsilon_j}) \right] \qquad (11)$$

$$Var(\hat{\beta}_{1,j}) = J^{-1} \sum_j \left[\hat{\sigma}^2_{\hat{\beta}_{1,j}} - \mathcal{C}\, (\hat{\sigma}^2_{\eta_j})^2 G_i^2 (\mathbf{1}_{\cdot j} - 1) \right] \qquad (12)$$

$$Var(\hat{\beta}_{2,j}) = J^{-1} \sum_j \left[\hat{\sigma}^2_{\hat{\beta}_{2,j}} - \mathcal{C}\, (\hat{\sigma}^2_{\eta_j})^2 G_i^2 (\mathbf{1}_{\cdot j} - 1) \right] \qquad (13)$$

Here, $\mathcal{C} \equiv |\hat{V}_j|^{-1} \cdot |\hat{\Sigma}_j|^{-1}$.

C Appendix C. Additional Figures and Tables

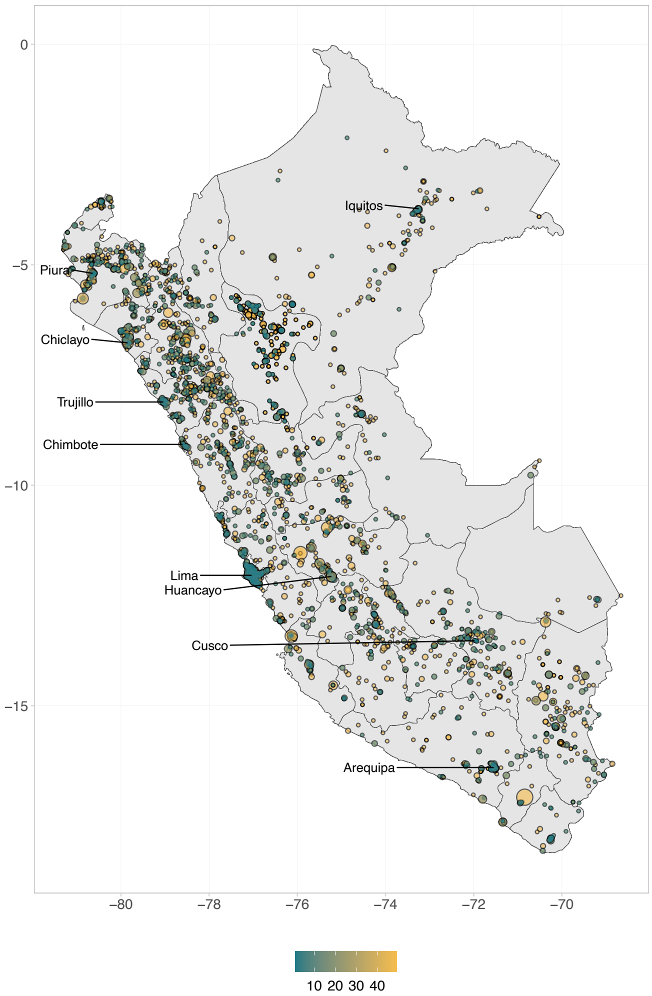

Figure 7: Map of Teachers' Survey Completion

Note: The map displays the count of teachers registered and those who completed the survey, binned at the school level based on school names as of 2019–20. Out of 2,115 schools with registered teachers, 1,673 had teachers who completed the survey. Among survey respondents, 46.9% were from schools with one participating teacher, 51.7% from schools with 2–10 participating teachers, and 1.3% from schools with 11–49 participating teachers. Yellow circles denote teachers who signed up but did not complete the survey, while green circles represent teachers who successfully completed the survey.

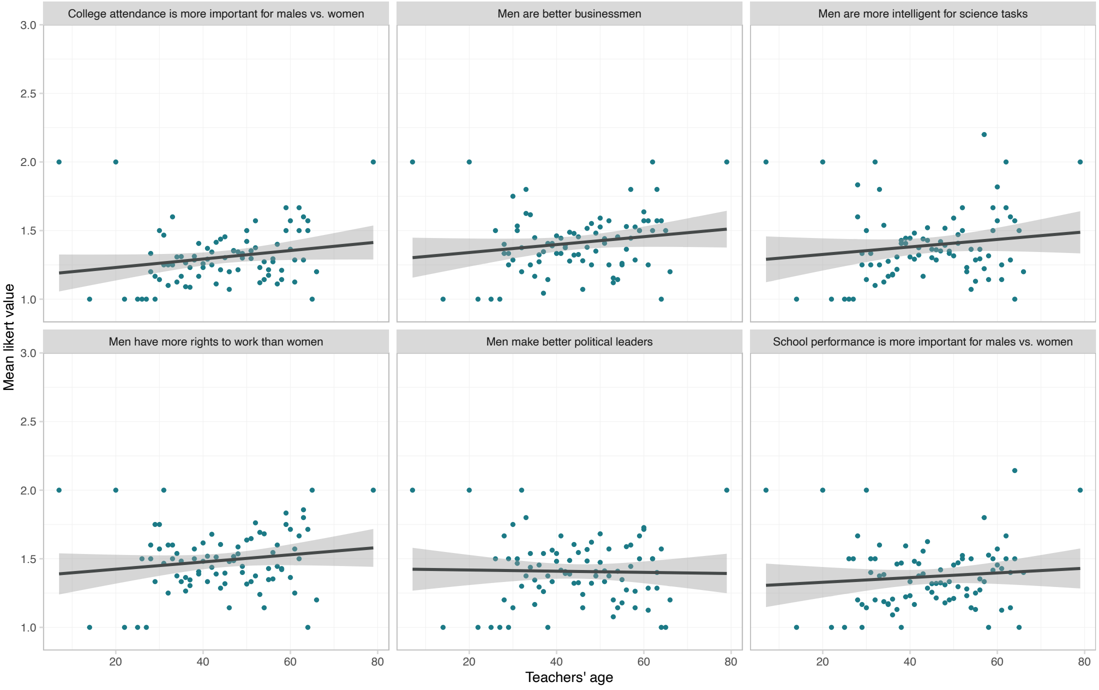

Figure 8: Mathematics Teachers' Self-Reported Gender Attitudes Collected by Survey

Note: This graph illustrates the relationship between teachers' age and their agreement levels with statements probing self-reported gender attitudes. The sample comprises 1,345 mathematics teachers with available IAT scores and stereotyped assessment measures. Each graph includes labels for the corresponding statements, and answers are coded to represent higher values for increased agreement, indicating more unfavorable gender attitudes towards women. The questions are phrased as, "On a scale of 1 (Strongly disagree) to 4 (Strongly agree), how much do you agree with the following statements."

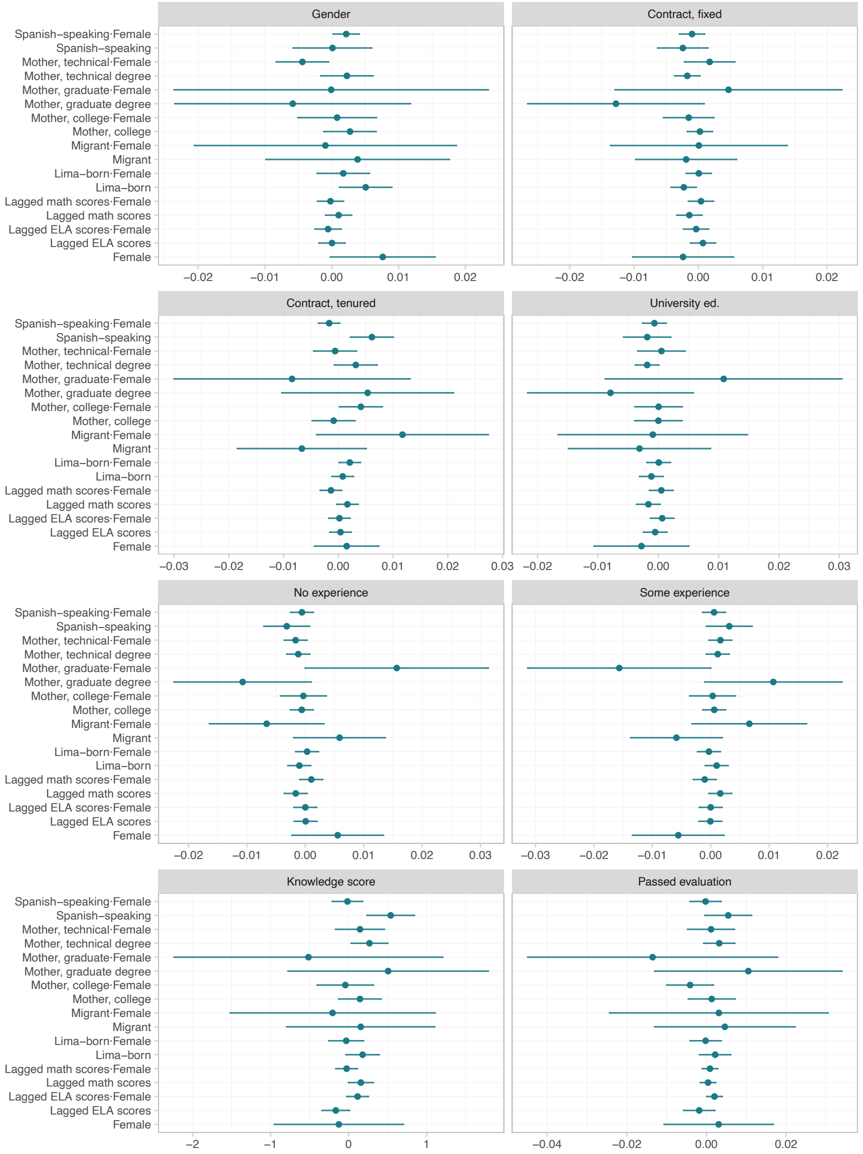

Figure 9: Exogeneity Test for the Relationship Between Observable Characteristics of Math Teachers and Pre-Determined Traits of Students

Note: This figure shows correlations between teacher characteristics and student pre-determined characteristics, adjusting for cohort, grade, year, and school fixed effects. The variable "University ed." distinguishes teachers with university degrees from those with technical ones. Experience denotes teaching experience in private schools. "Knowledge score" represents the average score in centralized promotion examinations taken by teachers from 2017 to 2019. "Passed evaluation" is a dummy variable indicating whether the teacher passed the test. Error bars show ±1.96 estimated standard errors, with standard errors clustered at the school level.

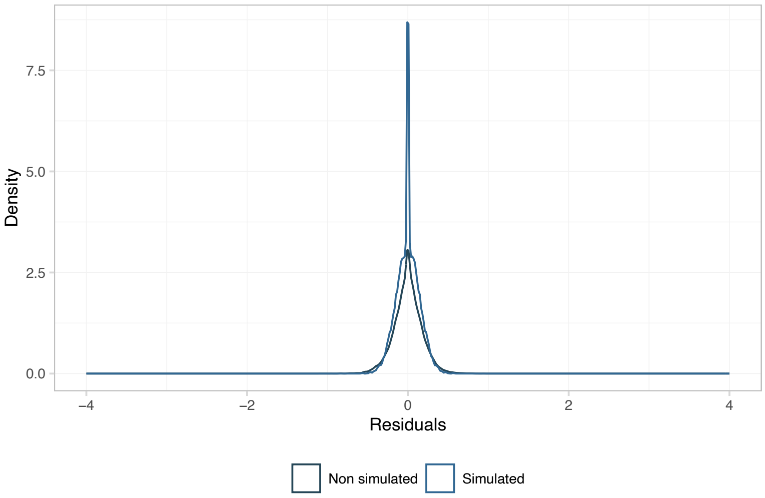

Figure 10: Actual and Simulated Variation in Teacher-Level Stereotyped Assessment Exposure

Note: The graph shows the Kernel density of residuals from teacher-level stereotyped class assessments on school fixed effects are shown. I calculated the residuals using 1,000 regressions with simulated data where students were randomly assigned to classrooms within a grade per school, considering school and classroom capacity constraints. Density calculations use an optimal bandwidth and an Epanechnikov kernel. Actual residuals have a 0.17 root mean squared error (RMSE), while simulated residuals have a 0.14 RMSE.

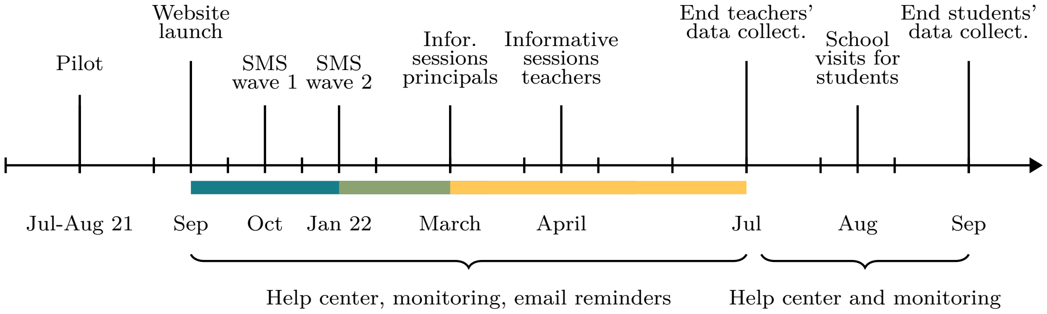
Website launch
Pilot
SMS wave 1
SMS wave 2
Infor. sessions principals
Informative sessions teachers
End teachers' data collect.
School visits for students
End students' data collect.
Jul-Aug 21
Sep
Oct
Jan 22
March
April
Jul
Aug
Sep
Help center, monitoring, email reminders
Help center and monitoring

Figure 11: Data Collection Timeline

Note: This figure illustrates data-collection phases for teachers spanning from September 2021 to September 2022 and for students from August 2022 to September 2022. Data collection covered the fourth quarter of 2021 and the first three quarters of 2022, with the color bar representing the last quarter of the 2021 school year (summer vacation) and the first two quarters of the 2022 school year. Student data collection occurred in August and September 2022, with enumerators assisting students in completing the IAT and survey during homeroom teachers' tutoring hours in Lima. A help center, accessible via email, phone, and WhatsApp, supported teachers and students throughout remote data collection and website operation.

Table XIII: Descriptive Statistics for 8th-Grade Students with Teacher-Assessed and Blindly Graded Test Scores

	Female	Male	Total
	(1)	(2)	(3)
A. Demographic characteristics			
Age, years	12.755	12.864	12.812
Spanish	0.591	0.595	0.593
Indigenous language (Quechua)	0.077	0.076	0.076
Other	0.015	0.015	0.015
Missing language	0.317	0.315	0.316
Born in Lima	0.205	0.206	0.206
Low-income household	0.229	0.233	0.231
Parents Some College	0.104	0.106	0.105
Parents College +	0.103	0.110	0.107
B. School attributes			
School in Lima	0.235	0.232	0.233
School in urban area	0.888	0.884	0.886
C. Students' educational aspirations			
HS	0.059	0.075	0.067
Technical	0.060	0.109	0.086
College	0.294	0.294	0.294
Graduate	0.250	0.181	0.214
Missing educational aspiration	0.337	0.340	0.338
Num. of students	605,902	641,961	1,247,863
Num. of schools			8,831
Num. of teachers			38,799
Observations			2,889,905

Note: This table presents summary statistics for eighth-grade students in 2015, 2016, 2018, and 2019. The data includes standardized exam and teacher-assigned scores in math, language arts, and science. Students are matched with their subject teachers. The bottom rows show the number of schools and matched teachers. Observations are at the student-subject level.

Table XIV: Descriptive Statistics on Matched High School Graduates' Employment Records

	Benchmark sample			Regression sample	
	Female	Male	Non reported	Female	Male
	(1)	(2)	(3)	(4)	(5)
A. Average earnings 2015-2020					
Monthly earnings (2010 USD)	218.113	254.862	321.517	213.648	247.902
Hourly wage (2010 USD)	2.759	2.979	3.421	2.830	3.114
B. Worker characteristics					
Worker age, years	19.202	19.244	19.437	19.160	19.203
Special education	0.000	0.000	0.000	0.001	0.001
Less than HS	0.046	0.064	0.163	0.034	0.058
HS graduate	0.621	0.691	0.779	0.603	0.704
Some college or technical	0.195	0.118	0.039	0.224	0.120
Technical degree	0.051	0.046	0.019	0.057	0.046
Bachelors degree	0.000	0.000	0.000	0.002	0.001
Masters, PhD	0.000	0.000	0.000	0.000	0.000
No information	0.088	0.082	0.000	0.079	0.070
C. NAICS industry					
11 agriculture, forestry, fishing	0.218	0.224	0.270	0.127	0.183
21-23 mining, utilities, construction	0.017	0.082	0.073	0.012	0.054
31-33 manufacturing	0.180	0.179	0.253	0.158	0.175
42-49 trade, transportation	0.222	0.213	0.150	0.276	0.264
51-59 information, finance, prof. services	0.183	0.167	0.124	0.209	0.169
61-62 educational and health care services	0.019	0.008	0.006	0.021	0.007
71-72 arts, recreation, hospitality services	0.099	0.074	0.058	0.127	0.091
81 other services	0.057	0.046	0.060	0.065	0.050
92-99 public adm. and unclass.	0.005	0.007	0.004	0.005	0.007
Number of employees	176,963	281,511	2,014,575	29,541	56,155
Number of firms	30,214	45,798	190,055	8,423	15,503

Note: This table presents descriptive statistics comparing the characteristics of recently graduated workers and public high school graduates from the estimation sample with available matched employer-employee records. Columns (1)–(3) show means for the benchmark sample of 16- to 25-year-old workers, aligning with the work experience of recent high school graduates in the regression sample. Columns (4)–(5) display means for students with projected graduation years of 2015–20 who are part of the estimation sample and have been employed for at least one month between 2015 and 2020. Panel A shows monthly earnings, paid working hours, and hourly wages. Panels B and C detail worker characteristics and occupational categories based on their most recent contact with a dominant annual employer. All exclusion filters outlined in Appendix A are met by workers in the benchmark sample.

Table XV: Variation in Estimates of Teacher-Level Stereotypical Assessments

	Teachers' subjects	
	Mathematics (1)	Language arts (2)
Mean		
Student-weighted	-0.005	-0.005
Bias-Corrected Standard Deviation		
School-size-weighted	0.168	0.177
Student-weighted	0.048	0.070
Num. of teachers	16,243	16,102

Note: This table presents estimated means and standard deviations of stereotyped teacher assessments, $\hat{\theta}_j$, derived from Equations (2) and (3). Covariates are detailed in Section 4.1. The first row displays the means, $\hat{\mu}$, of $\hat{\theta}_j$ weighted by student weights ($w_j = \mathcal{C}_j/\mathcal{C}$), where $\mathcal{C}(j)$ is students assigned to teacher j and $\mathcal{C} = \sum_j \mathcal{C}(j)$. The second and third rows show bias-corrected variance estimates, addressing sampling error in $\hat{\theta}_j$ using standard errors s_j. The second row presents school-size-weighted bias-corrected variance, $\hat{\sigma}_S$, and the third row shows student-weighted variance, $\hat{\sigma}_W$, computed with student weights, w_j. Observations are at the teacher level, with standard errors calculated using student-level clusters.

Table XVI: Relationship Between Gender Differences in Assessment and Mathematics Teachers' Characteristics

	(1)	(2)	(3)
Demographic characteristics			
Female	-3.558***	-3.545***	-3.347***
	(0.253)	(0.254)	(0.603)
Age, older than median	0.051	0.049	0.467
	(0.223)	(0.235)	(0.812)
Higher ed., university	0.331	0.364	0.16
	(0.292)	(0.291)	(0.696)
Teaching experience in private schools, years			
Less than 2 yrs		-0.405	-1.153
		(0.419)	(0.904)
2-5 yrs		-0.052	-0.589
		(0.356)	(0.755)
6-10 yrs		-0.08	-0.943
		(0.440)	(0.912)
More than 10 yrs		0.492	-0.598
		(0.657)	(1.354)
Teaching experience in public schools, years			
Less than 2 yrs		-0.838	0.076
		(0.958)	(3.996)
2-5 yrs		-1.7**	-0.781
		(0.815)	(3.625)
6-10 yrs		-1.311	-0.598
		(0.820)	(3.690)
More than 10 yrs		-1.504*	-1.143
		(0.859)	(3.729)
Teacher evaluation performance			
Skill & knowledge test, z score			0.086
			(0.474)
Passing status			-0.647
			(1.160)
R-squared	0.0258	0.0259	0.0419
Observations	15,741	15,741	5,308

Note: This table displays the estimated relationship between mathematics teachers' stereotypical assessment estimates and covariates. Coefficients are obtained from the estimating equation $\hat{\theta}_j = \phi_1 + X_j'\phi_2 + u_j$, where $\hat{\theta}_j$ is the graded estimate for teacher j divided by the bias-corrected standard deviation. Coefficients are weighted by the inverse of the associated standard error, s_j^2, for precision. As of 2019, teachers' median age is 45. Indicators for higher education use *Technical Institute* as the base category. *Teaching experience* variables, based on 2019 records, have *No experience* as the base category. National Teacher Evaluations (2015-2019) provide performance covariates. *Skills & knowledge test z score* is the standardized score, and *Passing status* indicates exam approval. Observations are at the teacher level, including school-fixed effects and missing-value dummies. Standard errors are clustered at the school level. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XVII: Coefficient Stability: High School Graduation Effects of Math Teacher Stereotyped Assessments

	(1)	(2)	(3)	(4)	(5)
			A. Graduated on time		
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.010**	-0.010*	-0.010**	-0.010**	-0.008
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$\hat{\theta}^*_{j,-i}$	-0.006	-0.003	-0.002	-0.002	-0.003
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Female	0.018***	0.011***	-0.014	-0.014	-0.032
	(0.001)	(0.001)	(0.015)	(0.016)	(0.019)
$\bar{Y}$ female	0.826	0.826	0.826	0.826	0.826
$\bar{Y}$ male	0.774	0.774	0.774	0.774	0.774
Joint sig., p-val	0.000	0.000	0.000	0.000	0.000
R-squared	0.373	0.402	0.402	0.403	0.414
			B. Graduated ever		
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.011**	-0.010**	-0.010**	-0.011**	-0.010**
	0.005	0.005	0.005	0.005	0.005
$\hat{\theta}^*_{j,-i}$	-0.003	0.000	0.003	0.003	0.002
	0.007	0.006	0.007	0.006	0.006
Female	0.003***	-0.005***	-0.013	-0.016	-0.029
	0.001	0.001	0.019	0.016	0.018
$\bar{Y}$ female	0.857	0.857	0.857	0.857	0.857
$\bar{Y}$ male	0.833	0.833	0.833	0.833	0.833
Joint sig., p-val	0.000	0.000	0.005	0.003	0.004
R-squared	0.178	0.229	0.229	0.253	0.259
Observations	2,583,863	2,583,863	2,583,863	2,583,863	2,583,863

Note: This table presents estimates of the impact of one-grade exposure to a mathematics teacher's stereotypical assessment on timely and overall high school graduation. The analysis uses the full regression sample of 8th–10th graders with projected graduation years from 2015 to 2019. The teacher-level stereotyped assessment is a leave-one-year-out estimate normalized by its year-subject-level standard deviation. Columns progressively introduce controls, with baseline covariates in columns (1) and (6) (grade, cohort, year, and fixed effects); adding student-level controls in columns (2) and (7), teacher-level controls in columns (3) and (8), and lagged student test scores in columns (4) and (9). All of the treatment variables, covariates, and joint significance terms are defined similarly to those in Table VII. The p-value for the joint significance of the total effect on girls is reported. The observation unit is the student grade; this specification uses two-way clustered standard errors by student and school. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XVIII: College Attendance Effects of Exposure to Mathematics Teacher Stereotyped Assessment

	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	0.001	-0.001	-0.006	-0.004	-0.003	-0.002
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
$\hat{\theta}^{*}_{j,-i}$	0.003	0.003	0.002	0.000	-0.001	-0.001
	(0.005)	(0.006)	(0.005)	(0.005)	(0.004)	(0.005)
Female	-0.013	0.008	0.018	0.02	0.012	0.012
	(0.015)	(0.012)	(0.012)	(0.011)	(0.014)	(0.014)
$\bar{Y}$ female	0.250	0.34	0.204	0.213	0.186	0.189
$\bar{Y}$ male	0.199	0.285	0.168	0.177	0.154	0.157
Joint sig., p-val	0.170	0.476	0.144	0.214	0.214	0.279
R-squared	0.116	0.138	0.090	0.098	0.094	0.097
Observations	2,562,561	2,562,561	2,562,561	2,562,561	2,562,561	2,562,561

Note: This table provides estimates of the impact of one-grade exposure to a math teacher's stereotypical assessment practices on college outcomes. The analysis includes a stacked sample of 8th–11th graders with projected graduation years from 2015 to 2019, utilizing the full regression sample. Controls for covariates, such as grade, cohort, year, and school fixed effects, are the same as those described in Table VII. Standard errors are two-way clustered by student and school (in parentheses). The p-value for the joint significance of the total effect on girls is presented. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XIX: Robustness Exercise: High School Completion Effects of Exposure to Mathematics Teachers' Stereotyped Assessments in a Subsample of Students Observed in Upper and Lower Grades

	8th grade (1)	9th grade (2)	10th grade (3)	All grades (4)
		A. Graduated on time		
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.053***	-0.026**	-0.008	-0.03***
	(0.009)	(0.01)	(0.008)	(0.005)
$\hat{\theta}^*_{j,-i}$	0.017	0.009	0.016	0.014*
	(0.007)	(0.008)	(0.007)	(0.004)
Female	0.018	0.003	-0.026	-0.013
	(0.053)	(0.039)	(0.032)	(0.021)
$\bar{Y}$ female	0.765	0.803	0.861	0.78
$\bar{Y}$ male	0.704	0.744	0.808	0.72
Joint sig., p-val	0	0.053	0.277	0
R-squared	0.294	0.365	0.456	0.367
		B. Graduated ever		
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.046***	-0.032***	-0.007	-0.025***
	0.009	0.01	0.008	0.005
$\hat{\theta}^*_{j,-i}$	(0.02*)	(0.017)	(0.018)	(0.01)
	0.007	0.008	0.007	0.003
Female	(0.022)	(-0.023)	(-0.003)	(-0.006)
	0.052	0.038	0.031	0.021
$\bar{Y}$ female	(0.79)	(0.829)	(0.889)	(0.804)
$\bar{Y}$ male	0.751	0.795	0.86	0.767
Joint sig., p-val	0	0.073	0.131	0
R-squared	0.253	0.303	0.35	0.285
Observations	565,976	399,759	353,313	1,613,248

Note: This table presents estimates of the impact of one-grade exposure to a math teacher's stereotypical assessment practices on high school graduation, both on time and ever. The analysis includes 8th–10th graders with projected graduation years between 2015 and 2019. The description of outcome variables, covariates and fixed effects is the same as in Table VII. The p-value for the joint significance of the total effect on girls is presented. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XX: Robustness exercise, high school –completion effects of exposure to mathematics teachers' stereotyped assessments in a subsample of students with teachers' stereotyped assessments measures available for all grades

	8th grade	9th grade	10th grade	All grades
	(1)	(2)	(3)	(4)
		A. Graduated on time		
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.052***	-0.018*	-0.001	-0.015***
	(0.010)	(0.009)	(0.007)	(0.004)
$\hat{\theta}^*_{j,-i}$	0.017	0.021*	0.003	0.009
	(0.011)	(0.012)	(0.012)	(0.005)
Female	0.021	-0.014	-0.005	-0.008
	(0.052)	(0.022)	(0.023)	(0.015)
$\bar{Y}$ female	0.754	0.793	0.858	0.824
$\bar{Y}$ male	0.683	0.732	0.808	0.772
Joint sig., p-val	0.000	0.599	0.625	0.012
R-squared	0.323	0.36	0.411	0.411
		B. Graduated ever		
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.048***	-0.019**	0.002	-0.013***
	0.010	0.009	0.007	0.004
$\hat{\theta}^*_{j,-i}$	(0.024**)	(0.0210*)	(-0.001)	(0.008)
	0.011	0.012	0.012	0.006
Female	(0.028)	(-0.022)	(-0.016)	(-0.006)
	0.039	0.028	0.017	0.014
$\bar{Y}$ female	(0.784)	(0.83)	(0.895)	(0.855)
$\bar{Y}$ male	0.743	0.803	0.878	0.831
Joint sig., p-val	0.001	0.735	0.771	0.050
R-squared	0.257	0.265	0.242	0.267
Observations	579,867	580,869	675,845	2,562,561

Note: This table presents estimates of the impact of one-grade exposure to a math teacher's stereotypical assessment practices on high school graduation, both on time and ever. The analysis includes 8th–10th graders with projected graduation years between 2015 and 2019. The description of outcome variables, covariates and fixed effects is the same as in Table VII. The p-value for the joint significance of the total effect on girls is presented. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XXI: High School Graduation Effects of Exposure to Language Arts Teachers' Stereotyped Assessments

	8th grade	9th grade	10th grade	All grades
	(1)	(2)	(3)	(4)
		A. Graduated on time		
$\hat{\theta}^{*}_{j,-i} \cdot$ Female	-0.015	-0.009	-0.007	-0.008
	(0.011)	(0.01)	(0.008)	(0.005)
$\hat{\theta}^{*}_{j,-i}$	-0.001	-0.008	-0.009	-0.003
	(0.013)	(0.013)	(0.013)	(0.006)
Female	-0.089	-0.021	-0.015	-0.032
	(0.038)	(0.053)	(0.024)	(0.019)
$\bar{Y}$ female	0.753	0.790	0.859	0.826
$\bar{Y}$ male	0.683	0.732	0.810	0.774
Joint sigificance, p-val	0.033	0.023	0.005	0.000
R-squared	0.325	0.361	0.420	0.414
		B. Graduated ever		
$\hat{\theta}^{*}_{j,-i} \cdot$ Female	-0.026**	0.002	-0.005	-0.009*
	(0.011)	(0.01)	(0.008)	(0.005)
$\hat{\theta}^{*}_{j,-i}$	0.012	-0.007	-0.005	0.002
	(0.013)	(0.013)	(0.013)	(0.006)
Female	-0.077	-0.019	0.015	-0.028
	(0.042)	(0.045)	(0.040)	(0.019)
$\bar{Y}$ female	0.784	0.828	0.896	0.857
$\bar{Y}$ male	0.744	0.803	0.880	0.833
Joint sig., p-val	0.072	0.491	0.043	0.012
R-squared	0.258	0.266	0.250	0.270
Observations	579,199	589,851	672,218	2,583,863

Note: This table presents estimates of the impact of one-grade exposure to a language arts teacher's stereotypical assessment practices on high school graduation, both on time and ever. The analysis includes 8th–10th graders with projected graduation years between 2015 and 2019. The outcome variables consider the contract with the dominant annual employer for at least one month during the specified age range. Student, teacher, lagged scores, classroom- and school-grade-level controls, place-of-birth fixed effects and gender interaction are covariates. Grade-specific estimations (Columns (1)–(3) and Columns (5)–(7)) treat the student as the observation unit, with cohort, year, and school fixed effects and school-level clustered standard errors. Stacked grade estimations (Columns (4) and (8)) use the student grade as the observation unit, employing a sample of students stacked across grades and including grade, cohort, year, and school fixed effects, along with two-way clustered standard errors by student and school. The p-value for the joint significance of the total effect on girls is reported. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XXII: Formal Sector Employment: Effects of Exposure to Language Arts Teachers' Stereotyped Assessments

	Age 17–18	Age 18–19	Age 19–20	Age 20–21	Age 21–22
	(1)	(2)	(3)	(4)	(5)
	A. Employed in the formal sector after high school graduation				
$\hat{\theta}^{*}_{j,-i} \cdot$ Female	-0.003***	-0.002	-0.004	-0.004	0.000
	(0.001)	(0.002)	(0.003)	(0.005)	(0.010)
M_js	0.002***	0.002	0.005*	0.004	0.009
	(0.001)	(0.002)	(0.003)	(0.005)	(0.010)
Female	-0.038***	-0.067***	-0.082***	-0.101***	-0.089***
	(0.002)	(0.004)	(0.007)	(0.011)	(0.023)
$\bar{Y}$ female	0.010	0.033	0.060	0.084	0.103
$\bar{Y}$ male	0.030	0.072	0.115	0.152	0.178
Joint sigificance, p-val	0.188	0.615	0.491	0.891	0.214
R-squared	0.040	0.040	0.036	0.039	0.042
	B. Paid monthly work hours				
$\hat{\theta}^{*}_{j,-i} \cdot$ Female	-0.323***	-0.205	-0.295	-0.209	0.754
	(0.09)	(0.196)	(0.342)	(0.555)	(1.082)
$\hat{\theta}^{*}_{j,-i}$	0.262***	0.212	0.477*	0.056	-0.463
	(0.075)	(0.162)	(0.289)	(0.472)	(0.98)
Female	-2.596***	-5.174***	-7.355***	-10.031***	-6.414***
	(0.209)	(0.434)	(0.745)	(1.167)	(2.219)
$\bar{Y}$ female	93.207	95.13	97.857	99.651	100.11
$\bar{Y}$ male	100.238	102.676	103.784	103.611	102.859
$\bar{Y}$ female uncond.	0.592	2.380	4.762	6.712	100.11
$\bar{Y}$ male uncond.	2.098	5.858	9.777	12.826	102.859
Joint sigificance, p-val	0.231	0.947	0.389	0.681	0.691
R-squared	0.027	0.029	0.027	0.030	0.030
Observations	2,556,134	1,689,262	881,845	393,186	114,020

Note: This table presents estimated coefficients of the impact of increased gender-stereotypical teacher assessments in a specific high school grade on the probability of holding a formal sector job and the monthly paid work hours ages 17–22 (post-high school graduation). The description of outcome variables and covariates is the same as in Table V. The p-value for the joint significance of the total effect on girls is reported. The unit of observation is the student-grade, clustered standard errors by school are reported in parentheses. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XXIII: Monthly Earnings: Effects of Exposure to Language Arts Teachers' Stereotyped Assessments

	Age 18–19 (1)	Age 19–20 (2)	Age 20–21 (3)	Age 21–22 (4)	Age 22–23 (5)
$\hat{\theta}^*_{j,-i}$ · Female	-0.483***	-0.659*	-0.378	-0.632	1.367
	(0.17)	(0.385)	(0.717)	(1.14)	(2.272)
$\hat{\theta}^*_{j,-i}$	0.361**	0.587*	0.717	0.253	-0.406
	(0.141)	(0.317)	(0.599)	(0.975)	(2.056)
Female	-3.828***	-9.144***	-14.159***	-20.944***	-11.279**
	(0.382)	(0.855)	(1.503)	(2.393)	(4.428)
$\bar{Y}$ female	216.3	211.9	215.2	222.2	226.6
$\bar{Y}$ male	246.9	247.1	248.9	252.8	251.6
$\bar{Y}$ female uncond.	0.9	3.9	8.4	12.2	14.8
$\bar{Y}$ male uncond.	3.4	10.4	18.7	24.8	28.1
Joint sig., p-val	0.194	0.737	0.425	0.615	0.523
R-squared	0.022	0.025	0.024	0.027	0.028
Observations	2,556,134	1,689,262	881,845	393,186	114,020

Note: This table presents estimated coefficients of the impact of increased gender-stereotypical teacher assessments in a specific high school grade on monthly earnings at ages 17–22 (post-high school graduation). The outcome variables consider the contract with the dominant annual employer for at least one month during the specified age range. All of the treatment variables, covariates, and joint significance terms are defined similarly to those in Table V. The p-value for the joint significance of the total effect on girls is reported. The unit of observation is the student-grade, clustered standard errors by school are reported in parentheses. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XXIV: Student Tracking to Stereotyped Teachers' Sequences

	Full sample			Subset sample: no repeated teacher assignment		
	9th grade (1)	10th grade (2)	11th grade (3)	9th grade (4)	10th grade (5)	11th grade (6)
Grade 8 · Female	0.042***			0.014		
	0.012			0.015		
9th grade · Female		0.013			-0.003	
		0.012			0.022	
10th grade · Female			0.023*			0.021
			0.012			0.015
Grade 8	0.217***			-0.195***		
	(0.03)			(0.035)		
9th grade		0.267***			0.163***	
		(0.035)			(0.057)	
10th grade			0.359***			0.289***
			(0.037)			(0.085)
Female	-0.01	-0.367	0.263	0.014	-0.13	-0.026
	0.27	0.282	0.346	0.333	0.326	0.364
R-squared	0.015	0.015	0.024	0.009	0.013	0.041
Observations	325,750	323,036	360,838	143,779	85,564	49,087

Note: The table displays the estimated relationship between stereotyped assessments students were exposed to in grade g and the assigned assessments in the subsequent grade, across high school grades. Coefficients are from $\hat{\theta}_{j(i),g+1} = \lambda_0 + \lambda_1 \hat{\theta}^*_{j(i),g} + \lambda_3 X_I + e_i$, where $\hat{\theta}_{j(i),g+1}$ is the unshrunk teacher-level stereotyped assessment divided by the bias-corrected (student-weighted) standard deviation. $\hat{\theta}^*_{j(i),g}$ is the standardized leave-one-out posterior mean under linear shrinkage. Each column reports the coefficient estimate for the indicated grade, with observations at the student level. School and classroom fixed effects are included, along with missing value dummies. Standard errors, clustered at the school-classroom level, are in parentheses. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XXV: Classroom Scores: Effects of Exposure to Teachers' Stereotyped Assessments

	Mathematics teachers				Language arts teachers			
	Grade 8 (1)	Grade 9 (2)	Grade 10 (3)	Grade 11 (4)	Grade 8 (5)	Grade 9 (6)	Grade 10 (7)	Grade 11 (8)
	A. Mathematics classroom scores							
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.026***	-0.008***	-0.01***	-0.011***	-0.01***	0.001	0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
$\hat{\theta}^*_{j,-i}$	0.008***	0.005**	-0.003	-0.004*	0.009***	0.000	0.000	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Female	-0.05	0.106	-0.018	0.158	-0.078	0.015	-0.142*	0.02
	(0.072)	(0.097)	(0.064)	(0.169)	(0.127)	(0.074)	(0.065)	(0.131)
$\bar{Y}$ female	-0.053	-0.037	-0.008	0.029	-0.051	-0.022	0.001	0.023
$\bar{Y}$ male	-0.19	-0.183	-0.164	-0.139	-0.189	-0.175	-0.17	-0.142
R-squared	0.437	0.442	0.446	0.466	0.429	0.433	0.441	0.459
	B. Language arts classroom scores							
$\hat{\theta}^*_{j,-i} \cdot$ Female	-0.005***	0.002	-0.001	0.000	-0.041***	-0.016***	-0.013***	-0.013***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
$\hat{\theta}^*_{j,-i}$	0.001	0.002	0.005***	-0.001	0.021***	0.005	0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)
Female	0.181	0.261	-0.007	-0.02	0.114	0.112	0.078	0.042
	(0.222)	(0.144)	(0.063)	(0.165)	(0.086)	(0.082)	(0.056)	(0.101)
$\bar{Y}$ female	0.042	0.069	0.09	0.119	0.047	0.075	0.095	0.121
$\bar{Y}$ male	-0.286	-0.293	-0.289	-0.269	-0.287	-0.295	-0.292	-0.272
R-squared	0.4	0.393	0.401	0.42	0.402	0.404	0.41	0.427
Observations	579,199	589,851	672,218	742,595	579,867	580,869	675,845	725,980

Note: This table presents estimates of the impact of one-grade exposure to mathematics and language arts teachers' assessment-based gender stereotypes on students' classroom scores. The analysis involves students in grades 8 to 11, with projected graduation years from 2015 to 2019. The teacher-level stereotyped assessment is a leave-one-year-out estimate for each student cohort based on their projected graduation year. Controls include covariates detailed in Table VII notes. The unit of observation is the student, with cohort, year, and school fixed effects and clustered standard errors at the school level. $*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table XXVI: Bias-corrected correlation matrix of parameters of teacher value-added and stereotyped teacher assessments

		Teacher VA parameters	
	Stereotyped assessment (1)	VA towards females (2)	VA towards males (3)
		A. Mathematics	
Stereotyped assessment	1	-0.116	-0.214
VA towards females		1	0.988
VA towards males			1
Standard deviation	0.116	0.576	0.599
		B. Language arts	
Stereotyped assessment	1	-0.364	-0.658
VA towards females		1	0.867
VA towards males			1
Standard deviation	0.124	0.126	0.179

Note: This table presents the correlation matrix between teacher-level stereotyped assessment measures ($\theta_j = \beta_{2,j} - \alpha_{2,j}$) and value-added parameters. Value-added for male students is $\alpha_{1,j} + \alpha_{2,j}$, and for females, it's $\alpha_{1,j}$. Parameters were estimated using a fixed-effects-SURE specification in Equations (2) and (3). The SURE-calculated sampling covariances adjusted the variance-covariance matrix of joint parameters for sampling error. Estimation equations consider quadratic polynomial lagged scores in language and mathematics, baseline controls at student and teacher levels, classroom and school-grade means of baseline covariates, and fixed effects for cohort, grade, school, and year.

Table XXVII: Descriptive Statistics for Surveyed Students with a Stereotyped Assessment Measure

	Female	Male	Total
	(1)	(2)	(3)
A. Demographic characteristics			
Age (years)	15.38	15.42	15.40
Mixed	0.15	0.13	0.14
Quechua	0.03	0.03	0.03
White	0.03	0.03	0.03
Afroperuvian	0.01	0.02	0.01
Other	0.03	0.04	0.04
Mother ed., high school or less	0.16	0.16	0.16
College +	0.10	0.10	0.10
Missing	0.73	0.74	0.73
School in Lima	0.82	0.80	0.81
Born in Lima	0.73	0.74	0.73
B. Postsecondary plans			
Higher ed., yes	0.87	0.78	0.84
Work, only working	0.01	0.02	0.01
Work, working and studying	0.68	0.68	0.68
Work, only studying	0.27	0.23	0.25
C. Discrimination and IAT			
Discriminated by teachers	0.08	0.10	0.09
Discriminated by peers	0.09	0.12	0.10
Previously taken IAT, None	0.74	0.68	0.72
1	0.11	0.14	0.12
2	0.06	0.09	0.07
3	0.04	0.04	0.04
4	0.03	0.03	0.03
5+	0.02	0.02	0.02
Gender-science stereotype (IAT score)	0.10	0.07	0.09
Number of students	3,069	2,055	5,139

Note: This table reports descriptive statistics for the analysis sample of high school students (grades 7–11) who were surveyed and matched with their stereotypical assessment estimates. Students' ages and postsecondary plans were reported between September 2022 and September 2021. Experience and witnessing discrimination is a self-reported measure of discriminatory behaviors by teachers or colleagues in 2021–2022, based on race, immigrant status, socioeconomic level, gender, sexual identity, and religion. IAT raw scores (non-standardized) are reported. IAT score interpretation is as stated in Table II.

D Appendix D. Analysis of Blindly Graded- and Classroom Examinations

Table XXVIII: Comparison of mathematics classroom and standardized test contents

Standardized test	Teacher-graded tests
A. Quantity	
Uses rational numbers	Represents a rational number
Compares rational numbers	Evaluates an affirmation linked to the equivalences between discrete, successive percentages and justifies its position
Equivalence between rational numbers	Select units to measure or estimate mass
Problem-solving involving additive and multiplicative notions in Q	Select and use strategies to operate with rational numbers
B. Regularity, Equivalence and Change	
Inference of additive, multiplicative or repetition patterns	Uses strategies to solve equations and inequalities of the first degree
Use and interpretation of algebraic expressions	Establishes relationships between data and conditions of situations that involve generalizing a pattern and representing it using an algebraic expression
Planning and solving equations and inequalities	Evaluates the validity of statements related to situations involving relationships between two equally proportional magnitudes
Interpretation of directly and inversely proportional magnitudes	Expresses understanding of a related function
Meaning of linear function and affine function	
C. Shape, Movement and Location	
Characteristics and properties of plane figures	Makes statements about the relationships and properties it discovers between objects and geometric shapes
Visualization and use of solid properties	Establishes relations between the characteristics and measurable attributes of real or imaginary objects
Calculation, estimation, and the relationship between the perimeter, the area, and the volume of geometric figures	Expresses, with drawings, constructions with a ruler and compass, concrete material and geometric language, his understanding of the relationship of similarity between two-dimensional forms.
Interpretation of geometric transformations in the plane	
D. Data Management and Uncertainty	
Interpretation of statistical tables and graphs, with grouped and ungrouped data	Collects data on nominal or ordinal qualitative variables and discrete or continuous quantitative variables through surveys or other methods
Interpretation of measures of central tendency with pooled and unpooled data	Use procedures to determine the median, mode, and mean of discrete data
Interpretation and use of the notion of probability	Expresses understanding of the meaning of probability value to characterize events.

Note: Table comparing question distribution in standardized math tests and benchmark teacher-graded tests. Percentages in each cell are based on the total number of questions. Standardized test contents sourced from Ministry of Education reports. Teacher-graded tests include eighth-grade benchmark entrance and exit exams and National Curriculum assessments.

Table XXIX: Mathematics questions by topic in classroom and standardized tests

	Standardized test		Teacher-graded test
Learning skills	2015	2016	
A. Quantity			
Use of rational numbers	0.06	0.04	0.04
Comparison of rational numbers	0.03	0.03	0.07
Equivalence between rational numbers	0.02	0.02	0.07
Problem-solving involving additive and multiplicative notions of rational numbers	0.14	0.16	0.11
B. Regularity, equivalence and change			
Inference of additive, multiplicative or repetition patterns	0.06	0.04	0.04
Use and interpretation of algebraic expressions	0.03	0.04	0.07
Planning and solving equations and inequalities	0.11	0.12	0.11
Interpretation of directly and inversely proportional magnitudes	0.03	0.04	0.04
Meaning of linear function and affine function	0.07	0.06	0.04
C. Shape, Movement and Location			
Characteristics and properties of plane figures	0.08	0.09	0.07
Visualization and use of properties of solids	0.06	0.00	0.00
Calculation, estimation, and the relationship between the perimeter, the area, and the volume of geometric figures	0.09	0.13	0.07
Interpretation of geometric transformations in the plane	0.02	0.02	0.11
D. Data Management and Uncertainty			
Interpretation of statistical tables and graphs, with grouped and ungrouped data	0.07	0.07	0.00
Interpretation of measures of central tendency with pooled and unpooled data	0.06	0.04	0.11
Interpretation and use of the notion of probability	0.08	0.08	0.04

Note: Comparison of question distribution in standardized math tests and benchmark teacher-graded tests. Percentages in each cell are calculated based on the total number of questions. Standardized test contents sourced from Ministry of Education reports. Teacher-graded tests include eighth-grade benchmark entrance and exit exams and the National Curriculum.

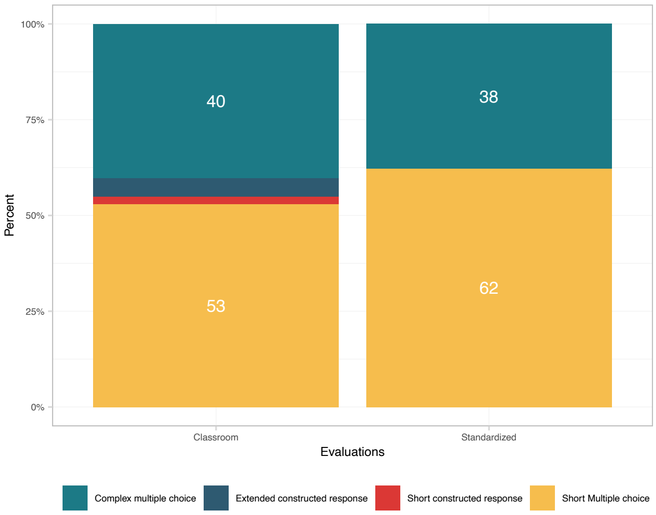

Figure 12: Distribution of test item formats in standardized assessments and classrooms

Note: The graph displays the percentage of 192 items categorized by format. The analysis includes two types of test items. Questions (items) from two tests are considered: Math end-of-year exams in four randomly chosen Metropolitan Lima schools and Ministry of Education model exams for teachers, comprising 102 "Classroom tests" items. Due to confidentiality, 90 standardized exam items were coded from technical public documentation. Classroom test items were classified by two external annotators, while standardized test items were heuristically classified as follows: items with difficulty levels 1 and 2 were coded "short multiple choice," whereas levels 3 and 4 were coded "complex multiple choice."

E Appendix E. Distributional Estimates of Teachers' Stereotyped Assessments

I use Empirical Bayes (EB) methods to leverage unbiased but noisy estimates, $\hat{\theta}_j$, of the true underlying teacher-level stereotyped assessment parameters, θ_j. These methods provide estimators for aspects of the population distribution of θ_j beyond mean and variance, even though these values are not observed (Efron and Morris, 1972, Stein, 1964). Estimators in this framework have desired properties such as minimizing Bayes risk and minimizing quadratic error loss (Efron and Morris, 1972). I present precise estimates from two EB methods: parametric EB (linear shrinkage estimation; see, for example, Chetty, Friedman and Rockoff (2014*b*), Gilraine, Gu and McMillan (2020), Kane and Staiger (2008)) and EB deconvolution (proposed by Efron (2010, 2012, 2016)).

E.1 Parametric EB estimator

The problem is modeled hierarchically by treating teacher-level stereotyped grading estimates as a random sample from a common prior population distribution of θ_j. Assuming that the population parameters θ_j follow a normal prior distribution with hyperparameters (μ, ϕ), where $\theta_j|\mu, \phi \overset{ind}{\sim} \mathcal{N}(\mu, \phi^2)$ for $j = 1, \ldots, J$. In J experiments (classroom assignments), teacher assignment j estimates the parameter θ_j from a class size $\mathcal{C}(j)$ that is independently distributed, written as $\hat{\theta}_j|\theta_j \overset{ind}{\sim} \mathcal{N}(\theta_j, s_j^2)$ for $j = 1, \ldots, J$, where σ_θ^2 denotes the variance of $\hat{\theta}_j$. The focus is on the joint conditional posterior distribution $p(\theta|\mu, \phi, \hat{\theta})$ of the underlying parameters θ_j, which takes a specific form in this case, $\theta_j|\mu, \phi, \hat{\theta} \sim \mathcal{N}(\hat{\theta}_j^*, \Sigma_j)$.

Moreover, the focus is on the posterior mean estimator $\hat{\theta}_j^* = (\frac{1}{\sigma_j^2}\hat{\theta}_j + \frac{1}{\phi^2}\mu)/(\frac{1}{\sigma_j^2} + \frac{1}{\phi^2})$, where $\sigma_j^2 = \sigma_\theta^2/\mathcal{C}(j)$ represents the sampling variance. The parametric Empirical Bayes (EB) is a precision-weighted average of the prior population mean and the estimator $\hat{\theta}_j$, shrinking the estimated stereotypical grading measure toward the respective sample mean. A feasible version of this posterior mean estimator only requires the standard error of the associated $\hat{\theta}_j$. The distribution of posterior means under linear shrinkage, shown in Appendix Figure 13, exhibits lower variability as it shrinks toward the sample mean.[29]

E.2 Application of EB Deconvolution Methods

The parametric Empirical Bayes (EB) setup relies on a Gaussian prior distribution for the true parameter θ_j, leading to a Gaussian posterior distribution. To assess the impact of this assumption, I explore a more flexible model for θ_j's prior distribution, following Efron (2016). This approach employs an exponential family of densities, assuming an unknown prior density $g(\mu)$ with population parameters distributed as $\theta_j \overset{ind}{\sim} G(\mu)$ for

[29] The independence assumption of the estimated $\hat{\theta}_j$ is not directly addressed here. Gilraine, Gu and McMillan (2020) propose a test to formally assess this assumption.

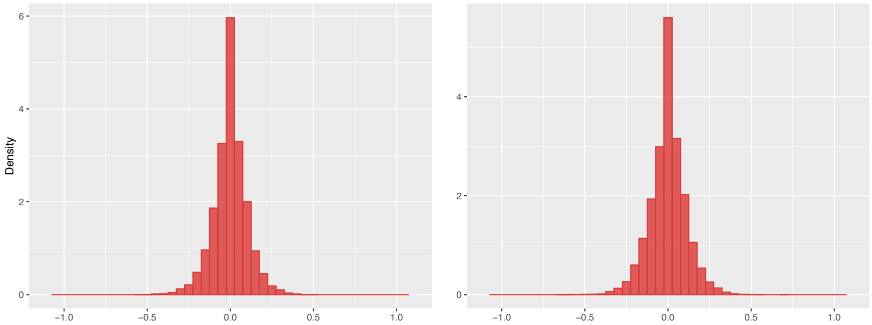

Figure 13: Distribution of linear shrinkage posterior means of teacher-level stereotyped assessment estimates

Note: The graph shows the distributions of posterior means, $\hat{\theta}_j^*$, of teacher-level stereotyped assessment estimates $\hat{\theta}_j$. The teacher-level stereotyped assessments estimates are shrunk toward the group mean. For visualization, the estimates are limited to ± 1.1.

$j = 1, \ldots, J$. Each θ_j independently draws an observed $\hat{\theta}_j$ from a known exponential probability-density family f_j, denoted as $\hat{\theta}_j \overset{ind}{\sim} f_j(\hat{\theta}_j|\theta_j)$. Deconvolved density $\hat{g}(.)$ of the teacher-level stereotyped assessment parameters θ_j is shown in Figure 2. Following Kline, Rose and Walters (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 Appendix Table XV. Visual inspection, following Kline, Rose and Walters (2021), aligns the deconvolved density well with the theoretical Gaussian density and its observed distribution. These results suggest that a linear shrinkage estimator approximates the posterior means as if computed using the true underlying population, particularly for stereotyped assessments against girls and boys located at opposite ends of the distribution (see Gilraine, Gu and McMillan (2020)).