Discrimination through Biased Memory*

Francesca Miserocchi†

May 17, 2026

Abstract

This paper shows that memory constraints amplify discrimination through limited and biased recall. I study science-track recommendations in Italian schools, where teachers evaluate students based on previously observed information. Gender gaps widen under higher evaluation load and disappear when teachers consult records. Two experiments isolate the mechanism: when relying on memory, teachers retrieve fewer student-specific signals and disproportionately recall stereotype-consistent ones, making girls 40 percent less likely than identical boys to receive a science recommendation. Allowing teachers to review records eliminates the gap. An experiment with U.S. evaluators generalizes the mechanism and documents stronger recall bias among male evaluators.

JEL Codes: J16 (Economics of Gender; Non-labor Discrimination); J71 (Discrimination); D91 (Role and Effects of Psychological, Emotional, Social, and Cognitive Factors); I24 (Education and Inequality); C91 (Laboratory, Group Behavior)

Keywords: discrimination; gender bias; memory; recall bias; cognitive load; STEM; educational tracking; stereotypes

*I thank the Unicredit Foundation and the Russell Sage Foundation for funding this project. I am grateful to Michela Carlana, Katie Coffman, Ben Enke, Ed Glaeser, Larry Katz, and Amanda Pallais for their support and guidance. I also thank Spencer Kwon and Andrei Shleifer for valuable feedback. This paper benefited from comments by Chiara Aina, John Conlon, Antonio Coran, Ross Mattheis, Raphael Raux, Maya Roy, Awa Ambra Seck, Jesse Shapiro, Alice Wu, and participants at the labor workshop and seminar at Harvard, the SITE Gender and Experimental Conference, and the HEC Ph.D. student conference. AEA RCT numbers: AEARCTR-0008921 and AEARCTR-0010756. IRB approval was obtained from Harvard University.

†Tilburg University. F.Miserocchi@tilburguniversity.edu

1 Introduction

Many high-stakes decisions require evaluators to forecast future performance using information observed in the past. These judgments are often made under cognitive constraints, making it difficult to retrieve all relevant signals. Managers assess employees' potential based on prior interactions (Benson, Li and Shue, 2021); judges evaluate defendants under heavy caseloads (Arnold, Dobbie and Yang, 2018), and teachers evaluate many students within a limited time. A large body of evidence from psychology shows that under cognitive strain, individuals struggle to retrieve information and are more likely to recall signals that align with social stereotypes (Macrae, Hewstone and Griffiths, 1993; Sherman and Frost, 2000). As a result, the information retrieved and used in evaluation may be incomplete and systematically skewed, creating scope for discrimination even when evaluators initially observed objective information about individuals.

This paper asks whether retrieval distortions amplify gender discrimination in high-stakes evaluations. I study the formal track recommendations that Italian middle school teachers issue at the end of eighth grade, a consequential decision shaping access to science-oriented education and, ultimately, entry into math-intensive careers. Using administrative data and surveys, I document that gender gaps in science-track recommendations widen when teachers face higher workload and when they rely more on memory rather than consulting student records. An experiment with middle school teachers provides causal evidence that memory constraints generate stereotype-consistent decisions even after observing ample, identical information: girls are 40 percent less likely than identical boys to receive a science recommendation when their teacher must rely on memory, and are steered toward humanities. The mechanism is limited and biased retrieval: teachers recall a limited set of student signals and disproportionately retrieve stereotype-consistent ones—for instance, that a student is passionate about poetry if she is a girl—and these recall distortions predict recommendation decisions. The effect is reversible: allowing teachers to review student records eliminates the gender gap. A large-scale online experiment with U.S. participants replicates the core mechanism in a different setting and population, and shows that male evaluators exhibit stronger stereotype-driven recall bias.

I first present observational evidence consistent with gender gaps in science-track recommendations widening under higher cognitive load, using administrative data on around 300 middle school teachers and their eighth-grade students. Track recommendations in Italy are assigned during class-specific meetings, when teachers evaluate all eighth-grade students within a short time window. This institutional setting generates variation in teachers' recommendation load across years: teachers who are assigned to more eighth-grade students—because classes are larger or because they are assigned multiple eighth-grade classes—must evaluate more students within the same time period. I exploit within-teacher variation across years in the number of students

each teacher must recommend to compare girls and boys with similar standardized math scores assigned to the same teacher in years of higher versus lower load. When the number of students to evaluate is higher, girls are significantly less likely than comparable boys to be recommended to the scientific track. On average, having a high rather than a low recommendation load—a difference of approximately seven students—reduces girls' probability of receiving a science recommendation by 4.2 percentage points relative to similar boys (around 21 percent). These patterns are not driven by changes in student composition or teacher selectivity.

Next, I conducted a survey of approximately 600 middle school teachers. I asked whether they consult the class register or rely on memory when assigning recommendations; nearly 60 percent report the latter. Relying on memory rather than consulting records entails a higher cognitive burden, as teachers must retrieve and process information about multiple students without external support. Linking survey responses to administrative records, I find that gender gaps in science-track recommendations emerge only among teachers who rely on memory. Girls with math ability comparable to boys are 4 percentage points (around 21 percent) less likely to receive a science recommendation from these teachers, while teachers who report consulting student records exhibit no gender gap. The difference is not explained by observable teacher characteristics, implicit stereotypes measured by an Implicit Association Test, school-level characteristics, or the types of students assigned to each group.

To isolate the causal role of memory constraints and identify biased retrieval as the mechanism, I conducted an experiment with 448 Italian middle school teachers. Teachers observed four hypothetical student profiles describing academic performance, interests, and background, then were asked to recall each student's characteristics and assign track recommendations. I randomized student gender, varied through first names, and teachers' memory constraints. Teachers in the *Memory* condition retrieved student characteristics without access to the profiles, while teachers in the *Baseline* condition could review the profiles before deciding, providing a perfect-memory benchmark. Since teachers observe the profiles of four students, signals about different students compete for retrieval, imposing memory constraints. Because student characteristics are held constant across gender conditions, any differences in recall reflect biased retrieval rather than differences in underlying attributes.

In the *Baseline* condition, girls are as likely as boys with identical profiles to be recommended for the science track. In the *Memory* condition, girls are 40 percent less likely than identical boys to receive a science recommendation and are disproportionately steered toward humanities tracks. With respect to the *Baseline* condition, memory constraints simultaneously under-allocate girls and over-allocate boys to the science track, implying approximately 11 percent fewer girls and 19 percent more boys enrolling in the science track. The mechanism operates through recall: relative to the *Baseline* condition, teachers in the *Memory* condition recall around half as many student-

specific signals and are 10 percentage points more likely to recall female-typed characteristics when evaluating a girl rather than an identical boy.

These recall distortions predict track recommendations: conditional on recalling the same total number of signals and evaluating the same student, recalling more female-typed rather than male-typed characteristics decreases the likelihood of science recommendations and increases the likelihood of humanities recommendations. Linking teachers in the experiment to administrative records on their past students, I find that teachers who exhibit larger gender gaps in their actual students' track recommendations also exhibit larger gender biases in recall when evaluating the hypothetical students. Importantly, the modal recommendation for each profile is the same across *Memory* conditions, indicating that memory constraints do not cause teachers to misperceive students or forget their characteristics entirely; rather, they shift which characteristics are retrieved in a stereotypical direction, altering decisions at the margin.

A large-scale online experiment with 1,010 U.S. participants recruited through Prolific replicates and extends these findings in a general evaluation setting. Evaluators observed a candidate, named either John or Susan, answering 10 trivia questions in a male-typed domain (sports) and a female-typed domain (pop culture), with both candidates performing identically (50 percent correct in each domain). Evaluators then completed a free-recall task in which they were asked to retrieve the questions answered correctly and incorrectly in the observed quiz (Quiz 1) and predicted the candidate's performance in a new quiz (Quiz 2). The same 2×2 design varied candidate gender and evaluators' memory constraints.

When required to rely on memory, evaluators predict that male candidates will outperform female candidates by 6.7 percentage points in sports and underperform by 1.8 percentage points in pop culture in Quiz 2, despite identical observed performance in Quiz 1. Under memory constraints, the gender gap is approximately 2.5 times as large as in the perfect-memory benchmark. Consistent with the teacher experiment, memory constraints reduce total recall volume by around half, and recall is biased in stereotype-consistent directions: evaluators recall more mistakes and fewer correct answers from the previously observed quiz when the candidate is female in a male-typed domain. The increase in discrimination under memory constraints is driven entirely by male evaluators, whose gender gaps nearly triple under higher cognitive load. Interestingly, only male evaluators exhibit disproportionate recall of stereotype-consistent characteristics, while female evaluators do not.

This paper contributes to several strands of literature. Foundational work in psychology and behavioral economics posits that individuals under cognitive strain rely more heavily on heuristics and implicit associations (Tversky and Kahneman (1974), Bertrand, Chugh and Mullainathan (2005)), yet empirical evidence on how cognitive strain shapes discriminatory outcomes and the mechanisms connecting them remains scarce. Singh and Venkataramani (2022) document that racial

disparities in hospital mortality widen as hospitals approach capacity, and Dube, MacArthur and Shah (2025) show that interventions expanding cognitive bandwidth among police officers reduce racial disparities in use-of-force decisions.

This paper provides the first experimental evidence directly linking memory constraints to group disparities in a high-stakes setting, identifying recall distortions as a novel mechanism through which cognitive constraints amplify discrimination. Cognitive load distorts what evaluators remember about previously observed candidates, skewing recalled characteristics in stereotype-consistent ways. This connects to recent work on memory and belief formation (Mullainathan, 2002; Bordalo et al., 2023, 2025) and to evidence that recall biases shape consequential economic decisions (Jiang et al., 2022; Charles, 2022). While differential treatment of otherwise identical candidates has traditionally been attributed to preferences (Becker, 1957) or beliefs (accurate, Phelps, 1972; Arrow, 1998; or distorted, Bohren, Imas and Rosenberg, 2019; Bohren et al., 2019), recent work highlights nontraditional cognitive channels, including contrast effects (Kessler, Low and Shan, 2022; Radbruch and Schiprowski, 2022), selective attention at exposure (Bartoš et al., 2016), and distorted signal perception driven by representativeness (Esponda, Oprea and Yuksel, 2023). This paper identifies limited and biased memory retrieval as a distinct cognitive channel.[1]

Second, the paper contributes to the literature on gender inequality in male-typed domains. Women receive lower credit for equivalent performance (Sarsons et al. (2021)), face harsher penalties for identical mistakes (Sarsons (2017), Egan, Matvos and Seru (2022)), and encounter steeper promotion barriers when advancement depends on discretionary assessments of potential (Benson, Li and Shue (2021)). This paper identifies a new mechanism behind such patterns: even when evaluators observe identical, ample information about male and female candidates, cognitive constraints introduce systematic asymmetries in what they subsequently recall, steering their assessments in stereotype-consistent directions.

Finally, this paper contributes to research on gender gaps in STEM and teacher decision-making. Prior work shows that teachers' biases affect student achievement (Carlana (2019), Lavy (2008)) and that boys are more likely than equally able girls to pursue science-oriented tracks (Buser, Niederle and Oosterbeek (2014)). This paper adds a new channel: gender disparities in educational tracking can emerge from cognitive constraints that bias information retrieval, even among evaluators who, when given full information access, treat boys and girls identically.

The paper is organized as follows. Section 2 describes the institutional setting. Sections 3 and 4 present the two pieces of observational evidence. Section 5 presents the teacher experiment, and Section 6 the Prolific experiment. Section 7 discusses the results and concludes.

[1] Seminal empirical studies on discrimination include Goldin and Rouse (2000), Bertrand and Mullainathan (2004), and List (2004). For a review of field evidence on discrimination, see Bertrand and Duflo (2017).

2 Institutional Setting

In the Italian education system, students select their high school track at the end of eighth grade. There are three main track types: academic, technical, and vocational. Within the academic category, the two most academically demanding tracks are the scientific and classical ones.[2] The scientific track emphasizes STEM subjects, while the classical track focuses on Latin and Greek and offers fewer math and science hours.

Track choices are highly consequential. Students attend classes in separate buildings, often in different neighborhoods, with limited interaction across tracks. Track choice predicts subsequent educational paths: the scientific and classical tracks lead to markedly different university fields, with the scientific track feeding predominantly into STEM degrees and the classical track into humanities and law.[3] Because track choice shapes university major and long-term earnings trajectories, the decision made at the end of eighth grade is an early and persistent determinant of labor market outcomes.

Teachers provide formal track recommendations to eighth-grade students and families via official letters, followed by parent-teacher meetings. Recommendations are assigned during class-specific teaching meetings held in November and December, where teachers jointly decide on a high school track for each student.[4] While all teachers contribute, math and Italian teachers play a central role: they teach the highest number of weekly hours per class, giving them more detailed knowledge of individual students, particularly relevant for scientific and classical track recommendations.[5] Recommendation meetings are typically concentrated in a single week, so teachers must evaluate many students within a short time window.

Though not binding, track recommendations strongly influence student choices. About 80 percent of students recommended for the scientific track select it, compared to 20 percent of those recommended for the humanities track. Conditional on math and Italian achievement and parental background, being recommended for the scientific track increases the likelihood of choosing it by more than 50 percentage points, suggesting that recommendations are a strong predictor of track choice even among observationally similar students.[6]

[2] Other academic tracks include linguistic, human sciences, and artistic high schools.

[3] Among graduates in industrial and IT engineering, 67 percent attended the scientific track and only 3 percent the classical track; among humanities graduates, 36 percent attended the classical track and 12 percent the scientific track (AlmaLaurea, 2023).

[4] Teachers are assigned to classes and may teach multiple classes simultaneously. Classes are cohort-based groups of students who spend grades 6 to 8 together. A teacher assigned to class 3A, for instance, teaches in eighth grade, section A; a teacher assigned to both 3A and 3B must evaluate students from both classes.

[5] Math and Italian teachers spend substantially more time with students than other subject teachers: 6 and 9 weekly hours respectively, compared to 2–3 hours for other subjects, as established by the national curriculum (D.P.R. 89/2009, Art. 5).

[6] These estimates are based on administrative data described in Section 3. Full regression results appear in Table A3.

3 Gender Gaps Widen Under Higher Recommendation Load

As initial field evidence, I document that gender gaps in science-track recommendations widen when teachers face higher evaluation load, consistent with cognitive load amplifying discrimination through limited and biased recall of individual information. The experiments in Sections 5 and 6 provide causal evidence on the mechanism.

3.1 Data and measurement

I use data on eighth-grade students from cohorts between the 2011–12 and 2015–16 academic years in 92 Italian public middle schools, assigned to their math teachers, as in Carlana (2019).[7] A key feature of the data is that it includes students' standardized test scores in math and Italian in eighth grade, which are externally administered and graded by INVALSI (the Italian National Institute for the Evaluation of the Education System). These scores are not evaluated by the students' teachers, allowing comparisons among students of equivalent objective ability. The dataset includes multiple cohorts of students taught by the same math teachers in different years, enabling within-teacher comparisons across years of higher and lower recommendation load. The final dataset includes approximately 18,000 eighth-grade students taught by 313 math teachers. Table A1 reports summary statistics.

Measure of Teacher Workload. I measure teacher workload as the number of eighth-grade students a teacher must recommend in a given academic year. A higher recommendation load increases cognitive burden during the weeks when recommendations are made, making it more difficult to accurately recall individual student characteristics. Math teachers face a heavier recommendation load when they teach larger eighth-grade classes or are assigned multiple such classes. Typically, math teachers oversee three classes per year—one each in grades 6, 7, and 8— but they occasionally teach two eighth-grade classes to meet school staffing needs; around 10 percent of math teachers in the sample fall into this category. Class sizes average 22 students, ranging from 12 to 29. I exploit within-teacher variation in the number of eighth-grade students recommended across years, conditional on the total number of students assigned to the teacher across all grades, which arises from year-to-year changes in eighth-grade class size and the occasional assignment of a second eighth-grade class. This variation translates into an average difference of about five students per teacher, with a maximum difference of 28 students for those teaching two eighth-grade classes in a given year (Figure A1).

[7] The initial sample also includes the 2016–17 academic year, but information on track recommendations is unavailable for that year, as noted in Carlana (2019).

3.2 Empirical Strategy

To assess whether gender gaps widen when teachers face higher workloads, I compare track recommendations for girls and boys of similar objective ability assigned to the same teacher in years of higher versus lower recommendation load. Specifically, I estimate the following equation at the student level:

$$\begin{aligned} Y_{ijt} = \alpha_0 + \sum_{s=2}^{4} \beta_s \cdot 1[\text{WorkloadBin}_{jt} = s] + \sum_{s=2}^{4} \gamma_s \cdot 1[\text{WorkloadBin}_{jt} = s] \times \text{Female}_i \\ + \beta_1 \text{Female}_i + \mu_t + \nu_j + \delta_0 \mathbf{X}_{it} + \delta_1 \mathbf{Z}_{it} + \varepsilon_{ijt} \end{aligned} \tag{1}$$

where Y_{ijt} is a dummy equal to one if student i, taught by math teacher j in year t, is recommended for the scientific track, and Female_i is an indicator for female students. The model includes teacher fixed effects (ν_j), which absorb all time-invariant teacher characteristics, and year fixed effects (μ_t). Teacher workload is divided into four roughly equal bins: 19 or fewer, 20–24, 25–29, and 30 or more students to recommend.[8] Standard errors are clustered at the teacher level.

The vector $\mathbf{Z}_{it}$ includes standardized test scores in math and Italian, and $\mathbf{X}_{it}$ includes mother's education level, father's occupation, indicators for immigrant status, the number of years the student has been taught by the same math teacher (1, 2, or 3 years), the total number of students assigned to the teacher across all grades (grades 6 to 8), and average classmates' characteristics.[9] Controlling for the number of years spent with the same math teacher ensures that results are not driven by teachers knowing their students less well in high-workload years. Controlling for the total number of students across all grades isolates variation in recommendation workload (recommendations are made for eighth-grade students only) from overall teaching load.

I also estimate a more compact specification in which $\text{High Workload}_{jt}$ is a dummy equal to one if the number of students teacher j must recommend in year t exceeds the median:

$$\begin{aligned} Y_{ijt} = \alpha_0 + \beta \text{High Workload}_{jt} + \gamma \text{High Workload}_{jt} \cdot \text{Female}_i + \alpha_2 \text{Female}_i + \\ \mu_t + \nu_j + \delta_0 \mathbf{X}_{it} + \delta_1 \mathbf{Z}_{it} + \varepsilon_{ijt} \end{aligned} \tag{2}$$

The coefficient of interest, γ (and γ_s in equation 1), captures the additional gender gap in science-track recommendations when the teacher faces a high workload in a given year.

[8] Results are similar when using a continuous measure of workload or quintile-based bins; see Table A12.

[9] Classmates' characteristics include the share of female classmates, the share of immigrants, the share of classmates with highly educated mothers, the share with fathers in high-skilled occupations, and leave-out means of standardized math and Italian test scores.

An interpretation of the γ coefficients as effects of higher workload requires that within-teacher variation in recommendation load be orthogonal to unobserved factors that differentially affect recommendations for boys and girls. The main threat would be that girls systematically differ from boys—in ability, background, or classmate composition—in years when teachers have more students to recommend. Following Pei, Pischke and Schwandt (2019) and Singh and Venkataramani (2022), I assess this directly by regressing student covariates from the main specification—standardized test scores, socioeconomic background, and classmates' characteristics— on the High Workload indicator interacted with Female, including teacher and year fixed effects.

Table A4 shows that gender gaps in a wide range of observables do not vary with teacher workload: the boy-girl gap in math and Italian test scores, parental education and occupation, and classmate composition is stable across high- and low-workload years. Within-teacher variation in the number of eighth-grade students to recommend, conditional on the total number of students assigned to the teacher, is determined by enrollment fluctuations and the occasional assignment of a second eighth-grade class when staffing needs require it. Both sources are determined at the school-administrative level rather than by teacher choice or student characteristics.

A higher recommendation load might affect gender gaps through channels other than cognitive load. In the next paragraphs, I examine how gender gaps vary with recommendation load and assess possible explanations.

3.3 Results

Panel (a) of Figure **1** shows the probability of being recommended for the science track by gender and teacher workload, accounting for teacher and year fixed effects and baseline controls. The gender gap widens with recommendation load: girls are about 8 percentage points less likely than similar boys to receive a science recommendation when their teacher has 25–29 rather than 10–19 students to recommend. The right panel shows the γ_s coefficients from equation 1; the pattern is robust to including student and classmate controls. Panel (b) shows that gender differences in standardized math scores do not widen with teacher workload, suggesting that the recommendation gap is not driven by changes in student ability composition across high- and low-workload years.

Table **1** quantifies the effect using specification 2. Girls are 4.2 percentage points (around 21%) less likely than similar boys to be recommended to the science track in high- versus low-workload years (columns 2–3), with results stable to the full set of controls. Importantly, high workload does not reduce the average science-recommendation rate (column 1), ruling out a capacity constraint. Results are stronger for high-achieving students in columns 4-6, who are most likely to excel in science tracks (gender gap increase of 8.3–8.2 percentage points in high workload years).

When teachers face higher loads, the "missing science" girls are reallocated to the top-classical

track (Table A7), a track of similar prestige and difficulty as the scientific one, but with curriculum focused on humanities—a pattern consistent with teachers falling back on stereotype-consistent decisions rather than relying on student-specific information.

Alternative Explanations. Three alternative mechanisms could, in principle, generate the widening gender gap. First, girls may participate less when classes are larger, reducing their visibility. This would predict widening gender gaps in teacher-assigned math grades as well; Table A5 shows no such pattern. Moreover, gender gaps widen when a teacher is assigned an additional eighth-grade class—increasing recommendation load holding class size constant (Table A8)—ruling out a within-class visibility mechanism. Second, teachers may direct more instructional effort toward boys under higher loads. This would predict higher teacher-assigned grades for boys in high-load years; Table A5 shows no such pattern. Third, the effect may reflect general annual workload rather than recommendation-specific load. Students in grades 6 and 7 do not generate recommendations, so their total number should not drive the pattern if the mechanism operates at the decision stage. Table A9 confirms that the number of sixth- and seventh-grade students assigned to a teacher does not predict gender gaps in recommendations.

Moreover, teachers with stronger implicit math–male associations show larger science-track penalties for girls in high-workload years (Table A10), consistent with higher cognitive load leading teachers to fall back on gender stereotypes.[10]

Appendix Section A.2 presents sensitivity and robustness checks. Results are robust across alternative specifications and sample definitions (Tables A12, A2, and A15) and are not driven by capacity constraints or rank-based decision rules (Tables A13 and A14).

4 Gender Gaps Widen When Teachers Rely More on Memory

A second piece of field evidence, directly linking memory constraints to gender gaps, comes from teachers' own reports on how they conduct recommendation meetings. This section documents that gender gaps are larger among teachers who rely more on memory, consistent with the mechanism identified in the experiments in Sections 5 and 6. When teachers do not consult class registers and instead rely more on memory, they face a higher cognitive burden and may recall a limited and biased set of student signals, which could lead them to penalize girls in math.

[10]Teachers in this sample completed an Implicit Association Test (Greenwald, McGhee and Schwartz 1998) linking boys to science and girls to humanities, a widely used measure of implicit gender stereotypes in economics (Carlana 2019; Lowes et al. 2015; Glover, Pallais and Pariente 2017; Kline, Rose and Walters 2022).

4.1 Data

Between February and March 2023, I surveyed approximately 600 Italian public middle school teachers. Surveyed teachers first participated in the experiment described in Section 5 and then answered questions about their recommendation practices.[11] Their responses were matched with administrative records of their eighth-grade students from 2016 to 2022.[12] Among the 609 surveyed teachers with eighth-grade classes, 285 taught core subjects (math and Italian). I classify teachers as relying on memory if they report remembering students' performance without consulting the class register or notes when assigning track recommendations, and as checking records otherwise.[13] Approximately 60 percent of core-subject teachers reported relying on memory during the meeting where track recommendations are assigned (Figure B3). Appendix Figure B1 compares survey teachers with the math teachers in Section 3, and Figure B2 shows the geographical distribution of survey teachers' schools.

4.2 Results

Figure **2** shows recommendation rates by gender and math ability decile, separately for core subject teachers who check records and those who rely on memory.[14] A gender gap emerges in the upper ability deciles only among teachers who rely on memory; teachers who check records show no gap at any ability level. Table **2** quantifies this pattern. I estimate:

$$\begin{aligned} \mathrm{Y}_{ij} = \beta_0 + \beta_1 \mathrm{Memory}_j + \beta_2 \mathrm{Female}_i + \beta_3 \mathrm{Female}_i \times \mathrm{Memory}_j \\ + \gamma_1 \mathbf{Z}_{ij} + \gamma_2 \mathbf{X}_j + \nu_{s(i)} + \mu_t + \varepsilon_{ij} \end{aligned} \tag{3}$$

where Memory_j is a dummy equal to one if teacher j reports relying on recollection rather than checking records. The coefficient of interest is β_3, the differential gender gap for memory-reliant teachers. The specification includes school fixed effects ($\nu_{s(i)}$), year fixed effects (μ_t), student-level

[11] A full description of the survey questions is available in Appendix C.1.

[12] I exclude the 2020–21 cohort because Italian middle schools faced widespread COVID-19 disruptions that year and recommendation practices differed substantially, as most schools still operated under remote instruction. The 2019–20 cohort is unaffected: recommendations were issued in November 2019 and COVID-19 disruptions began in March 2020.

[13] Specifically, teachers were asked about the actions they typically take in the class meetings where track recommendations are assigned, from a predetermined list including: *"I usually check my students' performance in the class registers and notes," "Another teacher checks students' performance,"* and *"I remember my students' performance without checking the class register."* In robustness checks, I treat the small group reporting both relying on memory and checking records as a separate category; results are similar (Table B4).

[14] I focus on core-subject teachers (math and Italian) since they play a primary role in track recommendation decisions. Compared to Section 3, I include Italian teachers here to increase statistical power given the smaller survey sample; results are similar when restricted to math teachers only (Table B3).

controls $\mathbf{Z}_{ij}$ (standardized test scores, immigrant status, mother's education, father's occupation, and average classmates' characteristics), and teacher controls $\mathbf{X}_j$ (subject, age, gender, contract type, dummy equal to one if born in the south). Standard errors are clustered at the teacher level.

Girls are 3.9 percentage points (around 21 percent) less likely to receive a science recommendation relative to similar boys if their teacher relies on memory. The gap is stable when baseline student and teacher controls are added in columns 2 to 4 of Table **2**. Teachers who check records show no significant gender gap.

One concern is that teachers who rely more on memory may differ along other dimensions that affect gender gaps. As a first step, I compare memory-reliant teachers with those who check records along observable characteristics (Table B1). Teachers who rely on memory tend to be older and more likely to have been born in Central or Southern Italy. Teachers also completed an Implicit Association Test at the end of the survey, which allows me to assess whether reliance on memory correlates with stronger implicit gender stereotypes. Teachers who rely more on memory do not exhibit stronger implicit bias, and they report similar beliefs about the importance of grades, attitudes, and parental background in shaping track recommendations.[15] To address the concern that remaining differences in teacher characteristics drive the results, I include controls for teachers' attitudes and IAT scores in column 5; the coefficient remains unchanged. The pattern is robust to alternative coding of the memory measure and interacting all controls with gender (Tables B4–B5).

Taken together, the two pieces of field evidence are consistent with the hypothesis that cognitive constraints amplify gender disparities: gender gaps in science-track recommendations widen when teachers operate under greater workloads and when they rely on memory rather than consulting records, and are absent when information is readily accessible. These patterns are robust to a wide set of controls and falsification tests, and the most plausible alternative explanations are unlikely to account for them. This evidence cannot rule out all confounders, nor identify the precise channel through which cognitive load amplifies discrimination. The experiments in the next two sections provide causal evidence on the mechanism.

5 Experiment with Teachers: Causal Evidence

The teacher experiment has two goals: to causally estimate the effect of memory constraints on gender gaps in track recommendations, and to test directly whether the mechanism operates through limited and biased retrieval of individual-level information. The conceptual framework in

[15] Table B2 extends these balancing tests to student characteristics by regressing student covariates on teachers' reliance on memory in the specification of equation 3. The analysis considers standardized test scores, teacher-assigned grades in math and Italian, parental background, and immigration status. The results indicate that teachers who rely more on memory are not systematically assigned to girls who differ from boys in ways that would mechanically widen gender gaps.

Appendix F formalizes two distinct channels: limited recall of student information leads evaluators to fall back on their priors (the fall-back-on-prior channel), while disproportionate recall of stereotype-consistent characteristics distorts the average signal retrieved from memory (the biased-recall channel). By eliciting what decision-makers remember under different cognitive constraints, the experimental design provides direct evidence on both channels.

5.1 Sample and Experimental Design

Italian middle school teachers who participated in the survey described in Section 4 first took part in the experiment. Teachers were informed that they would provide track recommendations for a series of hypothetical student profiles. They then observed four student profiles, one at a time. Each profile contained information on (i) academic performance in mathematics and Italian, (ii) interests, and (iii) additional background characteristics. I designed these profiles to include two students with "extreme" signals, one excelling in humanities but struggling in math, and the other exhibiting the opposite pattern, and two students with "mixed" signals: one excelling in both subjects and one with average performance in both. Detailed descriptions of the four profiles are provided in Appendix C.1.1. Figure **3** summarizes the experimental flow.

After viewing all four profiles, teachers were asked to report their recollections of each student's academic performance and interests in open-ended responses, and to provide a track recommendation. The recall task was financially incentivized: teachers participated in a lottery for a bonus prize, with the probability of winning increasing with the number of characteristics recalled. Recollections were elicited using a *free recall* task, a standard method in cognitive psychology that measures memory retrieval without prompts (Macrae, Hewstone and Griffiths 1993; Fyock and Stangor 1994).[16] By eliciting free recall, the experiment directly measures which student characteristics come most readily to mind and how memory constraints shape retrieval of previously observed student information. Teachers then provided a track recommendation for each student, choosing among the scientific track, the classical track, other humanities tracks (linguistic, artistic, and human sciences), and technical and vocational tracks, reflecting the options available in the Italian system.

Experimental Variation. Memory constraints were randomized at the teacher level, while student gender was randomized at the student-profile level. Gender was randomized by altering students' first names; in three of the four profiles, I changed gender by modifying only the last letter of

[16]Free recall tasks differ fundamentally from recognition tasks, where respondents choose among a list of options: whereas recognition captures whether information is familiar when re-encountered, free recall captures whether it is *accessible* and spontaneously retrieved. This distinction is critical for the mechanism being tested: the paper's hypothesis concerns which information comes to mind at the decision stage, not whether teachers would recognize a characteristic as belonging to a student if prompted.

the name, holding all other name attributes constant (e.g., religious origin, potential associations with socioeconomic status). Memory constraints were randomized across teachers by assigning them to either a *Memory* group or a *Baseline* group. Teachers in both groups initially observed identical information about the students. After reviewing the profiles, teachers in the *Memory* group were required to report their recollections and provide recommendations without re-examining the profiles; teachers in the *Baseline* group were permitted to review the profiles again before completing both tasks. Figure C2 displays the recall screen for both conditions.

The *Memory* condition mimics the naturalistic decision environment documented in Section 4, where most teachers rely on recollection rather than records. The *Baseline* condition provides a clean benchmark for recommendations made without memory constraints, a situation that is rarely achievable in practice, but allows the experiment to isolate the causal effect of memory load on gender gaps. The design thus maps directly onto the real-world contrast between teachers who check records and those who rely more on memory, lending external validity to the experimental comparisons.[17]

Sample and Balance. A total of 503 teachers were randomized into the experiment, and 448 completed the recommendation task. Early in the field period, the survey did not require teachers to submit recommendations, generating attrition and differential attrition across arms.[18] Differential attrition could bias the results if correlated with teacher characteristics affecting the outcome. I address this concern in two ways. First, Figure C3 shows that while differential attrition decreases over time and disappears after the forced-response feature was introduced on February 17, the treatment effect estimate remains stable throughout and shows no corresponding trend. Restricting the sample to teachers surveyed after February 17 yields quantitatively identical estimates (Appendix C.2). Second, baseline characteristics are well balanced across the *Memory* and *Baseline* groups (Table **3**): teachers in both groups are similar in education, socioeconomic background, age, gender, and subject taught; they also report comparable attitudes toward recommendation criteria, exhibit similar IAT scores, and display similar historical gender gaps in recommendations to actual students (constructed from the administrative data). The close match on historical gender gaps is reassuring: teachers with a history of higher gender gaps, both overall and among high-ability students, are not systematically over-represented in either condition. Following Imbens and Rubin

[17] Originally, I had preregistered an additional treatment, the *Memory and Higher Cognitive Load* condition, involving a second randomization of memory constraints through an extra cognitive load manipulation. I implemented only the *Memory* treatment to preserve statistical power following higher-than-expected attrition. The simple memory treatment proved effective: teachers were unaccustomed to participating in lab-style experiments, and the survey design required a manipulation of memory load that did not appear artificial.

[18] Fifty-five teachers viewed the profiles but did not complete the recommendation task (11 percent attrition): 39 in the *Memory* group and 16 in the *Baseline* group, implying differential attrition of 23 teachers (about 5 percent of the randomized sample). As soon as I acknowledged the presence of differential attrition, I introduced the forced-response feature and differential attrition disappeared (C.2).

(2015), all normalized differences are below 0.25 standard deviations. A joint F-test of the null that treatment is unrelated to all baseline covariates is also not significant.

5.2 Methodology

I estimate the following equation at the student-teacher level:

$$Y_{ij} = \beta_0 + \beta_1 \text{Memory}_j + \beta_2 \text{Female}_{ij} + \beta_3 \text{Memory}_j \times \text{Female}_{ij} + \mu_i + \beta_4 \mathbf{X}_j + \varepsilon_{ij} \quad (4)$$

One observation is a student profile i evaluated by teacher j. The outcomes Y_{ij} are either a dummy for a science-track recommendation, a dummy for humanities-track recommendations, or indicators for whether the teacher recalled female- or male-typed characteristics. The main outcome is science-track recommendation, given its importance for subsequent educational and labor market outcomes and its consistency with the administrative analysis, where individual teacher influence is most cleanly identified for the scientific track. Humanities-track recommendations serve as the complementary outcome, capturing reallocation away from science; they correspond to female-typed recommendations, defined as tracks disproportionately recommended to girls in the administrative data.[19]

Student fixed effects μ_i absorb all profile-level variation, so identification comes from within-profile comparisons across teachers in different memory conditions evaluating students of different genders. The vector $\mathbf{X}_j$ contains teacher-level controls: birth year, gender, subject taught, father's education, contract type, school region, and region of birth. Standard errors are clustered at the teacher level.

The main hypothesis is $\beta_3 < 0$ for science-track recommendations and $\beta_3 > 0$ for humanities recommendations: memory constraints amplify gender gaps by making teachers more likely to steer girls toward stereotype-consistent tracks. The coefficient β_2 captures the gender gap in the *Baseline* condition; if full information access at the time of decision eliminates bias, β_2 should be close to zero.

[19] Tracks are classified as male- or female-typed based on the relative frequency of track recommendation by gender in the administrative data, following the pre-registration (Table C3): female-typed tracks are classical and other academic tracks (linguistic, artistic, human sciences), which are disproportionately recommended to female students; male-typed tracks are scientific, technical, and vocational, which are disproportionately recommended to male students. Since all tracks are classified as either male- or female-typed, results for male-typed recommendations are mechanically the mirror image of results for female-typed recommendations—the coefficients are equal in magnitude and opposite in sign—and are therefore not reported separately; the pre-registered outcomes are thus fully recoverable from Table **4**. Figure C5 shows treatment effects for each track separately.

Classifying Student Characteristics. I classify student characteristics as *female-typed* if they are disproportionately associated with females and as *male-typed* if they are disproportionately associated with males, in line with the definition of stereotypes as exaggerations of attributes more distinctive of a group (Bordalo et al. 2019). Specifically, I classify characteristics as female-typed if they relate to humanities, reading, writing, and emotional sensitivity, and as male-typed if they relate to mathematics, science, technical skills, competition, and intuitive talent (Table C2).

This classification follows a large literature documenting gendered language in evaluative contexts: descriptions of women disproportionately emphasize relational and humanities-oriented traits (Trix and Psenka 2003; Schmader, Whitehead and Wysocki 2007), while descriptions of men emphasize technical competence and brilliance rather than diligence (Leslie et al. 2015; Eberhardt, Facchini and Rueda 2023; Baltrunaite, Casarico and Rizzica 2024). Despite the prevalence of such patterns in observational data, existing studies cannot disentangle whether they reflect biased retrieval or underlying gender differences in these attributes. The experimental design addresses this directly: because student profiles are held constant across conditions and genders, any gender differences in what teachers recall must reflect distortions in retrieval rather than differences in what was observed.

To construct the recall measures, I code teachers' open-ended free-recall responses by identifying mentions of all pre-determined characteristics from the student profiles (Table C2 lists all student profiles' characteristics). The main recall measure codes a student characteristic as recalled when thinking about student i if the teacher's response mentions it, regardless of accuracy, and thus captures both *accurate recall*—correctly remembering a characteristic belonging to the student—and *false memories*—attributing to a student a characteristic from a different profile. If teachers form impressions based on all the information they retrieve, whether accurate or not—both forms of distortion affect decisions. I construct supplementary measures to separately identify accurate recall and false memories: false memories for student i are student characteristics retrieved as belonging to student i which truly belonged to a different student j, while accurate recall for student i measures whether a characteristic of student i is retrieved by the teacher as belonging to student i.

5.3 Main Treatment Effects on Recommendations

Figure **4** displays the share of students recommended to scientific and humanities tracks by student gender and memory condition. Among students with identical profiles, girls are 11.3 percentage points less likely than boys to be recommended for the science track when teachers must rely on memory. This gap disappears entirely when teachers can review student information: in the *Baseline* condition, girls are no less likely than boys to receive a science recommendation. Interestingly, in the memory condition, the share of girls recommended to the science track falls while the share of boys rises, consistent with teachers reverting to stereotypes that steer girls away from

science and boys toward it. A symmetric pattern appears for humanities tracks: memory constraints disproportionately steer girls toward humanities. Taken together, these results indicate that gender bias in track recommendations is not a fixed characteristic of teachers but a state-dependent response to the cognitive environment at the time of decision.

Table **4** quantifies these effects. Columns (1) and (4) show average gender gaps in science and humanities-track recommendations among students with identical profiles, controlling for student fixed effects: girls are 5.6 percentage points less likely than boys to be recommended for the science track. Columns (2) and (5) add the memory interaction. In the *Baseline* condition, the gender gap in science-track recommendations is small and statistically indistinguishable from zero: with information freely accessible, girls are as likely as boys to receive a science recommendation. In the *Memory* condition, large gender gaps emerge: girls are 12.5 percentage points less likely to receive a science recommendation—corresponding to a 40 percent reduction relative to boys in the *Memory* condition—and 17.5 percentage points more likely to be directed toward humanities tracks. Adding teacher controls in columns (3) and (6) leaves the estimates almost unchanged.

These effects imply a meaningful misallocation of talent. Relative to girls in the *Baseline* condition, memory constraints reduce the probability of girls being allocated to the science track by approximately 20 percent, redirecting them toward humanities tracks. At the same time, boys are disproportionately steered toward science: their probability of receiving a science recommendation increases by approximately 34 percent relative to boys in the *Baseline* condition. Memory constraints thus simultaneously under-allocate girls and over-allocate boys to the science track. As a suggestive back-of-the-envelope calculation, using the estimated effect of receiving a science recommendation on science-track enrollment (56 percentage points, with no significant gender difference; Table A3), these recommendation gaps would translate into roughly 11 percent fewer girls and 19 percent more boys enrolling in the science track.

The experimental effect of 12.5 percentage points is larger than the effects documented in the field: a 4.2 percentage point increase in the gender gap under high workload in the administrative data, and a 3.9 percentage point increase for teachers who self-report relying more on memory. This attenuation is expected: the experimental manipulation imposes a sharp memory constraint on information that teachers just observed, whereas in the field, teachers have months of direct observation of their students, making memory constraints less binding.

A natural concern is whether teachers' behavior toward hypothetical profiles reflects their real-world recommendation practices. Figures C9–C12 address this directly by comparing experimental recommendations for each profile with historical recommendations for actual students with similar grades. In all cases, the modal track recommendation in the experiment matches that in the administrative data; where deviations occur, they are consistent with the additional information on students' interests and attitudes provided in the experimental profiles but not observed in

the administrative records—factors that teachers themselves report as among the most important determinants of track recommendations (Figure B5). Importantly, the modal recommendation for each profile is also the same across the *Memory* and *Baseline* conditions (Appendix C.6): memory constraints do not cause teachers to misperceive student profiles entirely, but rather to form more stereotype-consistent impressions of otherwise identical students depending on their gender.

Moreover, Figure C4 shows that teachers with larger historical gender gaps in science-track recommendations in the administrative data exhibit larger gender gaps in the experiment under the *Memory* condition, while no such correlation emerges in the *Baseline* condition. This pattern confirms that experimental gender gaps mirror teachers' real-world behavior, and that even teachers with large historical gender gaps display no bias when information is freely accessible—suggesting that discrimination is not a fixed trait but depends on the cognitive constraints teachers face at the time of the decision.

5.4 Mechanism: Limited and Biased Retrieval of Student Information

This section provides direct evidence on how memory constraints distort the retrieval of student-specific signals using free-recall data. Because student profiles are held constant across conditions, any gender differences in what teachers remember must reflect distortions in retrieval rather than differences in what was observed.[20]

The hypothesis is that, under memory constraints, individuals recall a limited and potentially distorted set of student signals. The direction of recall distortions is a priori ambiguous. Individuals may disproportionately retrieve stereotype-consistent information because the candidate's gender acts as a retrieval cue, increasing the probability of recalling signals more similar to the group stereotype —consistent with the similarity-based retrieval mechanism formalized in Bordalo et al. (2023), and with evidence from social psychology on stereotypes as retrieval schemas (Macrae, Hewstone and Griffiths, 1993; Sherman and Frost, 2000). Alternatively, teachers may disproportionately retrieve stereotype-inconsistent features due to surprise effects, as expectancy-violating information may be more memorable (Stangor and McMillan (1992); Hastie and Kumar (1979)). Stereotype-consistent retrieval would exacerbate discrimination, whereas disproportionate recall of inconsistent information could attenuate or even reverse it for highly counter-stereotypical candidates.

Figure **4**, panel (b), documents two effects of memory constraints on recall. First, teachers in the *Memory* condition recall fewer student characteristics overall—approximately half as many as in the *Baseline* condition—consistent with cognitive load reducing access to stored information. Second, and more distinctively, the characteristics they recall are distorted in a stereotype-consistent direction:

[20]The recall task was completed before teachers provided track recommendations, reducing concerns that recalled characteristics reflect ex-post rationalization of a recommendation already made.

teachers are 9.1 percentage points ($p = 0.003$) more likely to recall female-typed characteristics when evaluating girls rather than identical boys, and 5 percentage points ($p = 0.087$) less likely to recall male-typed characteristics. Neither gap is present in the *Baseline* condition, where recall is both more complete and unaffected by student gender.

Table C4 quantifies these effects using equation 4 with the recall indicator of female- or male-typed characteristics as the outcome. Teachers in the *Memory* condition are 10.6 percentage points (21 percent relative to the baseline mean) more likely to recall female-typed characteristics when evaluating a girl rather than an identical boy, relative to teachers in the *Baseline* condition. This estimate is robust to the inclusion of student fixed effects and teacher controls. A symmetric pattern emerges for male-typed characteristics, though the estimates are smaller and statistically insignificant.

Figure **5** disaggregates these patterns by individual profile characteristics. Teachers in the *Memory* condition disproportionately recall humanities- and emotion-related characteristics when evaluating girls—for instance, that the student won a poetry award, loves writing, reads fiction, and is passionate about languages—and science- and technology-related characteristics when evaluating boys—for instance, that the student is intuitive with technology, loves scientific subjects, and enjoys sports—despite identical profiles. Teachers are also more likely to recall background and social characteristics for female students, consistent with Wu (2018), who shows that women are more frequently described in terms of personal attributes rather than competence.

Beyond qualitative characteristics, Table C5 reports analogous analyses for teachers' recall of students' grades. Teachers in the *Memory* condition are less likely to recall any grade information overall. This reduced recall of objective grade information may itself contribute to the recommendation gap through a reversion-to-prior channel: when teachers cannot remember a student's grades, they fall back on prior beliefs about the distribution of ability by gender. Conditional on recalling grades, teachers tend to recall slightly higher literature grades for girls than for identical boys, though these differences are statistically insignificant.

Stereotype-consistent recall is driven by students with mixed signals — those who perform well in both humanities and science, or average in both — where individual-level information is most ambiguous (Table C6). For students with extreme signals, exceptionally strong in either math and technology or humanities, surprise effects may partially offset stereotype-consistent recall. Figure **5** illustrates this: the characteristic "fixes bikes," which belongs to a student with extremely male-typed characteristics, is disproportionately recalled for girls under memory constraints, consistent with this characteristic being highly unexpected for a female student. This heterogeneity is consistent with the administrative data, where higher workload amplifies gender gaps most strongly among students with strong performance in both math and Italian — precisely those for whom signals are most ambiguous and surprise effects are least likely to counteract stereotype-driven recall.

Together, these patterns suggest that memory constraints operate through two reinforcing channels: reduced recall of objective information, which weakens the anchor that ties recommendations to individual performance and leads teachers to fall back on prior beliefs, and biased retrieval of qualitative characteristics, which actively tilts the information set in a stereotype-consistent direction.

Notably, the recall biases may reflect selective memory of stereotype-consistent signals under memory load, or false memories from misattribution errors at retrieval. To disentangle the two, I construct two additional measures: *accurate recall*—retrieval of characteristics that genuinely belong to the student—and *false memories*—misattribution of characteristics from another profile. Both are captured in the main recall measure. Appendix Figures C7 and C8 show that both contribute to the overall pattern, suggesting that memory constraints lead both to the selective retrieval of stored information and to biased reconstruction at the decision stage. The presence of false memories, which involves the misattribution of characteristics across profiles, is difficult to reconcile with selective attention at encoding alone and points instead to active stereotype-based reconstruction at retrieval.

5.5 Biases in Recall and Biases in Decisions

A key assumption underlying the mechanism is that teachers form assessments based on what they recall and make decisions accordingly. If teachers were aware of their recall biases, they might discount their recollections, and disproportionate recall of stereotype-consistent information would not generate discrimination. I therefore examine whether limited recall quantity and biased recalled content are associated with gender gaps in recommendation decisions, along the two channels in the conceptual framework (Appendix F). Both tests use within-teacher variation in recall content across profiles: teacher fixed effects ensure that estimates compare how the same teacher recommends differently depending on what she happens to recall about each student, rather than reflecting stable differences in bias across teachers.

Panel (a) of Figure **6** documents the *fall-back-on-prior* channel. It shows how gender gaps in recommendations vary with the number of individual-level signals recalled, controlling for student and teacher fixed effects and the gender gap in recalled characteristics. Teachers who recall fewer student-specific signals make more stereotypical recommendations, and the gender gap in science recommendations widens as recall decreases, consistent with teachers reverting to gender priors when individual-level information is scarce.

Panel (b) documents the *biased signal recall* channel. The binned scatterplot shows the relationship between the gender gap in recalled characteristics—the difference between female-typed and male-typed characteristics recalled—and the probability of a science recommendation,

controlling for teacher and student fixed effects, and total recall volume. Recalling one additional female-typed rather than male-typed characteristic is associated with a 5 percentage point reduction in the probability of a science recommendation, indicating that even small distortions in recall are associated with sizable gaps in decisions.

Figure 7 connects the experimental recall patterns to real-world behavior. Among teachers who face memory constraints in the experiment, those with larger historical gender gaps in actual science-track recommendations—measured from administrative data prior to the experiment—also exhibit more gender-biased recall in the *Memory* condition. No such correlation emerges in the *Baseline* condition. This suggests that the biased retrieval documented in the experiment is connected not only to experimental decisions but also to gender gaps in actual recommendations.

Decomposition of Recall Channels. Panels (a) and (b) of Figure 6 suggest that the documented recall biases account for a large share of the Memory-induced recommendation gap. I quantify the contribution of each channel via an accounting exercise that maps directly to these panels (Appendix D).[21] The *Memory* condition shifts teachers toward lower recall bins; Panel (a) shows that gender gaps in recommendations are larger at lower recall levels, consistent with greater reliance on gender priors when less individual information is retrieved. This *fall-back-on-prior* channel accounts for approximately 44 percent of the Memory-induced gap. The *Memory* condition also shifts the composition of recalled signals toward stereotype-consistent characteristics; Panel (b) shows that recalling more female-typed than male-typed signals increases the probability of a female-typed recommendation, conditional on total recall. This *biased-recall* channel accounts for a further 28 percent. Together, the two channels explain approximately 72 percent of the gap, with the remainder potentially reflecting direct stereotype activation at the decision stage, independent of recalled content, or measurement error in the recall data.

6 Generalizing the Mechanism: A Large-Scale Experiment with U.S. Evaluators

The teacher experiment shows that memory constraints causally amplify gender discrimination and identifies limited and biased retrieval as the mechanism in a real-world, high-stakes setting. Three important questions remain that the teacher experiment cannot answer. First, does the mechanism extend beyond education to other evaluative contexts and populations? Second, the classification of student characteristics as female- or male-typed in Section 5 is based on evidence on commonly

[21] The decomposition uses the probability of a female-typed recommendation as the outcome; see Appendix D for details.

held stereotypes, but to better understand the mechanism of stereotype-driven recall, it is important to examine the recall patterns with a more objective, data-driven signal classification of stereotype-consistent signals. Third, the teacher sample is nearly 80 percent female, which precludes analysis of whether the mechanism operates differently across evaluator gender. The Prolific experiment addresses all three.

I recruited 1,010 U.S. evaluators through Prolific in a 2×2 design that varies candidate gender and memory constraints. Signals are trivia questions with verifiable correct answers. Following Bordalo et al. (2019), I classify domains as male- or female-typed based on the observed gender gap in correct answer rates in a reference population of 400 participants who completed the quiz prior to the main experiment—a more objective, data-driven classification unavailable in the teacher experiment. Finally, I recruited a gender-balanced evaluator sample, enabling heterogeneity analysis by evaluator gender that the teacher sample does not permit.

6.1 Sample and Experimental Design

The experiment follows a 2×2 between-subjects design, varying candidate gender (John or Susan) and memory constraints (*Memory* vs. *Baseline* condition). Figure E1 provides an overview.

Evaluators are told that a large number of candidates completed 80 trivia questions across two domains, sports and pop culture, and that they will need to predict the performance of one candidate in these domains. These domains were chosen following Bordalo et al. (2019), who document that males outperform females in sports trivia and females outperform males in pop-culture trivia—making each domain stereotype-consistent for one gender and stereotype-inconsistent for the other. Before observing any performance, evaluators report their prior beliefs about how many out of 10 questions the candidate answered correctly in each domain.

Evaluators then observe their candidate answering Quiz 1, which consists of 10 questions per domain, presented one at a time, along with the candidate's answer and the correct answer (Figure E2 provides an example). The signals are drawn from real candidates whose performance was identical on the 20-question subset shown to evaluators—both answering exactly 50 percent correct in each domain. From the evaluator's perspective, candidates are identical except for gender.

After observing Quiz 1, evaluators complete a free recall task, listing all questions they remember the candidate answering correctly and incorrectly in separate open-ended boxes (Figure E3).[22] Evaluators then predict the candidate's performance on a new, similar Quiz 2. Evaluators in the *Baseline* condition receive a link to review the candidate's responses before the recall task, serving as a perfect-memory benchmark. Evaluators in the *Memory* condition must rely solely on memory and

[22]Evaluators may report the question title, the candidate's answer, or a uniquely identifiable topic. The task is explicitly designed as an open-ended free recall exercise to mirror real-world decision environments, paralleling the design of the teacher experiment.

are not permitted to review the responses. All evaluators receive financial incentives for accurately predicting candidate performance.[23]

Two research assistants independently coded the open-ended recall responses to identify which questions evaluators recalled, and whether each was placed in the correct or incorrect answers box.[24] A question was coded as recalled only if both assistants agreed on its classification; the two assistants agreed on 98 percent of question-evaluator pairs.[25]

The main recall outcomes at the evaluator by domain level are the total number of questions recalled as answered correctly (incorrectly) by the candidate, and the share of questions recalled as correct out of all questions recalled, all in domain d (sports/pop-culture). Correspondingly, at the question level, the main recall outcomes are whether a question is recalled as correct or incorrect—that is, whether it is recalled and placed in the correct or incorrect box. These encompass both *accurate recalls*, where the evaluator correctly remembers the sign of the candidate's answer, and *false memories*, where the evaluator misattributes the sign. To decompose the source of recall bias, I construct two additional measures. *Accurate recall* measures how many of the truly correct (incorrect) questions the evaluator recalled as correct (incorrect). *False memories* are questions retrieved with the wrong sign: truly incorrect questions placed in the correct box, or truly correct questions placed in the incorrect box.

In total, 1,010 participants completed the main study. Table E2 confirms balance across treatment arms.

6.2 Results

Predicted Performance in Quiz 2. Figure E5 shows average predicted Quiz 2 performance by treatment group. Despite having observed equal performance by gender and by domain, when evaluators rely on memory, they predict male candidates will outperform female candidates by 6.7 percentage points in sports and underperform by 1.8 percentage points in pop culture. When evaluators are allowed to review Quiz 1 responses in the perfect memory benchmark, these disparities fall by around a half.

To quantify the effect, I estimate equation 5, where the unit of observation is evaluator j evaluating candidate i in domain d (sports or pop culture). μ_d denotes domain fixed effects and X_j is a vector of

[23] Evaluators in the *Memory* condition additionally receive financial incentives for the recall task. *Baseline* evaluators do not, as access to the original responses makes recall trivial. This is a conservative design choice: the absence of recall incentives in the *Baseline* condition may attenuate the estimated effect of memory constraints, so effects in the *Memory* condition are if anything a lower bound on the impact of removing memory constraints.

[24] Any uniquely identifiable reference to a question was coded as a recall: the question title, topic, the candidate's answer, or relevant keywords. For example, for the pop-culture question "What are the names of Kim Kardashian's children?" responses such as "Kim's children", or "Saint, North, Psalm" were all coded as recalls.

[25] Appendix E4 reports robustness checks using the alternative coding where a question is counted as recalled if at least one assistant coded it.

baseline controls including evaluator gender, age, education, employment status, political affiliation, and number of prior approvals on Prolific. Following Bordalo et al. (2019), I classify a domain as stereotype-consistent with a gender if that gender outperforms in it — confirmed in the reference population by Panel (A) of Figure E4. The indicator Stereotype-Inconsistent$_{id}$ equals one when candidate i's gender is inconsistent with domain d: a female candidate evaluated in sports, or a male candidate evaluated in pop culture. The coefficient of interest β_3 captures the penalty for stereotype-inconsistent candidates under memory constraints:

$$\begin{aligned} \mathrm{Y}_{idj} = \beta_0 + \beta_1 \mathrm{Memory}_j + \beta_2 \mathrm{Stereotype\text{-}Inconsistent}_{id} + \\ \beta_3 \mathrm{Memory}_j \cdot \mathrm{Stereotype\text{-}Inconsistent}_{id} + \gamma X_j + \mu_d + \varepsilon_{idj} \end{aligned} \tag{5}$$

Columns 1–3 of Table **5** report estimates for all evaluators and separately by evaluator gender. Despite observing identical Quiz 1 performance, evaluators predict significantly lower performance for stereotype-inconsistent candidates under memory constraints: the gender gap in the *Memory* group is approximately 2.5 times as large as in the perfect-memory benchmark (column 1). Strikingly, this effect is entirely driven by male evaluators, whose baseline gender gap nearly triples under memory constraints (column 2). Female evaluators show no gender gaps in the perfect memory benchmark, where they can review the signals, and a non-significant increase in discrimination under higher memory load (column 3).

Biased Recall of Quiz 1 Performance. Do evaluators recall candidates' Quiz 1 performance in a biased way? Panel (a) of Figure **8** displays the cumulative distribution of the share of Quiz 1 questions recalled as correct (placed in the correct answer box, out of total questions recalled), by candidate gender, domain, and treatment condition. In the *Baseline* group, more than 90 percent of evaluators place exactly 50 percent of questions in the correct box in both domains, as expected since they could review the candidate's answers. The share of questions recalled as correct displays more dispersion for evaluators in the *Memory* group, who also exhibit a gender gap in recall in the stereotypical direction: they retrieve a higher share of sports answers as correct if they observe male candidates and a higher share of pop-culture answers as correct for female candidates, despite having seen identical 50 percent performance.

Columns 4–6 of Table **5** report regression estimates at the evaluator-by-domain level, where the dependent variable is the share of recalled Quiz 1 questions in domain d that the evaluator places in the correct box, out of total questions recalled. Since candidates answered exactly 50 percent of questions correctly in each domain, a share above (below) 0.5 indicates that the evaluator recalls the candidate as having performed better (worse) than they actually did. A candidate is *stereotype-*

inconsistent in a domain if their gender is the minority performer in that domain: a female candidate in sports or a male candidate in pop-culture. In the *Baseline* condition, evaluators recall on average 49.6 percent correct answers, close to the true 50 percent. Under memory constraints, evaluators recall stereotype-inconsistent candidates as having answered a 5.33 percentage point lower share of questions correctly, and this recall bias is entirely driven by male evaluators (column 5). Appendix Table E3 decomposes this pattern using as dependent variables the count of questions placed in the correct and incorrect boxes: under memory constraints, stereotype-inconsistent candidates have 0.17 more questions placed in the incorrect box (columns 4–6), with no corresponding increase in questions placed in the correct box (columns 1–3), resulting in lower perceived performance.[26]

I now turn to the question-level analysis to study which specific signals evaluators misremember and in which direction. At this level, a signal is *stereotype-inconsistent* if it contradicts the gender stereotype for the domain: a correct answer in sports or a mistake in pop-culture by a female candidate, and a correct answer in pop-culture or a mistake in sports by a male candidate. Figure E4, Panel (b), displays gender gaps for each question from the reference population: correct pop-culture answers and sports mistakes are more common among women (female-typed), while correct sports answers and pop-culture mistakes are more common among men (male-typed). The figure also reveals substantial variation in how gender-typed individual signals are: answering the Devil Wears Prada question correctly carries a 20 percentage point gender gap, while mistakes on the Kim Kardashian question are strongly male-typed and mistakes on the Larry Bird question are strongly female-typed.

To study whether evaluators disproportionately retrieve stereotype-consistent signals under memory load, I follow a two-step procedure. First, I residualize the probability that a question is recalled as correct (incorrect) on evaluator-level controls and two question-level controls: whether the candidate answered the question correctly, and gender-specific question difficulty—defined as the share of correct answers for each question by gender in the reference population—interacted with the memory indicator. Following Bordalo et al. (2019), controlling for gender-specific question difficulty is important to isolate the effect of stereotypes from a generic difficulty effect: a correctly answered question is a male-typed signal if men answer it correctly *more often than women*, not simply if it is easy for everyone. Without this control, lower recall for stereotype-inconsistent questions could reflect the fact that male-typed questions tend to be harder for women, rather than the effect of gender stereotypes on retrieval. In the second step, I regress the residualized recall indicator on *Memory*, *Female Candidate*, and their interaction, where the coefficient on the interaction captures whether memory constraints disproportionately suppress recall of a question

[26] Similar to the teacher experiment, Appendix Table E5 shows evidence of misattribution errors at retrieval: under memory constraints, evaluators are more likely to misattribute questions as mistakes when evaluating a stereotype-inconsistent candidate.

when evaluating a female candidate.

Figure 8, Panel (b), shows the coefficient on *Female Candidate* × *Memory* from the residualized question-level regressions, separately for correct and incorrect answers. Questions are ordered from most male-typed to most female-typed based on the gender gap in the reference population. For correct answers, the coefficient is negative and largest in magnitude for the most male-typed questions—sports correct answers—indicating that evaluators under memory constraints are less likely to retrieve a female candidate's success in sports, where stereotype-inconsistency is greatest. For incorrect answers, the coefficient is negative for pop-culture mistakes and most negative for the most stereotype-inconsistent signals. The gradient in both panels reflects stereotype-consistent retrieval: the more stereotype-inconsistent the signal, the larger the recall penalty under memory constraints. The pattern is concentrated among male evaluators, who drive the effect in both panels, while female evaluators show a qualitatively similar but weaker pattern.

Table 6 provides question-level estimates of selective recall, separately for questions the candidate answered correctly (columns 1–3) and incorrectly (columns 4–6). The dependent variable in columns 1–3 is a dummy equal to one if a truly correct answer is retrieved and placed in the correct box; in columns 4–6, it equals one if a truly incorrect answer is retrieved and placed in the incorrect box. $\textit{Stereotype-Inconsistent}_{qid}$ is a dummy equal to one if the signal is stereotype-inconsistent, as defined above. The coefficient of interest, $\textit{Stereotype-Inconsistent}_{qid} \times \textit{Memory}_i$, captures whether stereotype-inconsistent signals are differentially retrieved under memory constraints.

Columns 1–3 show results for correct answers. Pooling all evaluators (column 1), a correct answer is 6 percentage points less likely to be retrieved under memory constraints when it is stereotype-inconsistent—for instance, a female candidate answering a sports question correctly is less likely to have that success retrieved. The effect is entirely driven by male evaluators (coefficient of -14.6 percentage points in column 2, significant at the 1 percent level), while female evaluators show no significant effect (column 3). Columns 4–6 show a symmetric pattern for incorrect answers: a mistake is 8.7 percentage points less likely to be retrieved under memory constraints when the signal is stereotype-inconsistent (column 4)—for instance, a male candidate making a sports mistake is less likely to have that error retrieved. This too is driven exclusively by male evaluators (column 5: -19.3 percentage points). The large negative coefficient on *Memory* confirms that memory constraints reduce total recall volume by around half, independent of stereotype-consistency. Together, these results show that memory constraints generate not only lower recall of signals but also selective retrieval that systematically disadvantages stereotype-inconsistent candidates, and that this mechanism operates more strongly for male evaluators.

Recall of Quiz 1 performance and Quiz 2 predicted performance. Similar to the teacher experiment, recall content predicts assessments, showing that evaluators base their decisions on their

biased memories: conditional on total recall volume and evaluator fixed effects, recalling a higher share of questions as correctly answered in a domain predicts higher Quiz 2 ability (Figure E6). This relationship persists despite financial incentives for accurate performance prediction. If evaluators were aware of their recall bias, incentives should lead them to correct for it; the fact that they do not suggests the bias operates unconsciously. Evaluators thus base their assessments on the biased subset of signals retrieved, incorporating both selective retrieval and misattribution errors into their predictions of candidates' future ability. Table E6 estimates the impact of accurate recalls and false memories on prediction of Quiz 2 performance, exploiting within-evaluator variation in recall across domain including evaluator fixed effects: both channels predict assessments, confirming that both selective retrieval and misattribution errors are consequential for evaluation outcomes and that evaluators "trust" their memories.[27]

Decomposition of Recall Channels. An accounting exercise analogous to the teacher experiment maps the recall biases to the Memory-induced assessment gap.[28] I estimate that the fall-back-on-prior channel—evaluators relying more heavily on stereotypes when recalling fewer signals—accounts for 73.1 percent of the gap, and the biased signal recall channel for a further 11.6 percent, together explaining around 85 percent of the Memory-induced assessment gap. Compared to the teacher experiment, both channels operate but the biased signal recall channel plays a smaller relative role.[29]

7 Conclusion

This paper shows that cognitive load amplifies discrimination through limited and biased recall. When Italian middle school teachers evaluate a large number of students, girls are significantly less likely than equally skilled boys to be recommended for the science track; the gap disappears when teachers consult student records. A teacher experiment identifies limited and biased recall under cognitive load as the mechanism, showing that bias starts at the recall stage: under memory constraints, teachers recall fewer objective signals and disproportionately retrieve stereotype-consistent characteristics, making girls 40 percent less likely than identical boys to receive a science

[27] Figure E7 also shows that recalling fewer signals in total, conditional on recall bias, is associated with larger gender gaps in predicted ability, consistent with lower total recall—holding recall bias constant—leading evaluators to fall back on priors.

[28] The decomposition uses bins for the difference between correctly and incorrectly recalled questions as the recall content measure, paralleling the teacher experiment. See Appendix D.2 for details.

[29] Two interpretations are consistent with this pattern. First, the teacher experiment involves a binary decision, so moderate differences in recalled content can discretely flip the outcome once a threshold is crossed. Second, recalling more female-typed versus male-typed characteristics strongly predicts track recommendations, whereas the composition of recalled quiz answers has a more limited effect on the continuous Prolific assessment.

recommendation. A large-scale online experiment replicates the findings in a different setting with a general population of financially incentivized evaluators, confirming that the mechanism extends beyond education, and showing that male evaluators display larger biases when assessing stereotype-inconsistent candidates, recalling more failures and fewer successes relative to identical stereotype-consistent candidates.

The findings point toward a coherent class of interventions that reduce cognitive burden or restore access to objective records at the decision stage, rather than targeting evaluators' attitudes directly. Providing structured summaries at teacher recommendation meetings re-anchors decisions to individual performance; reducing the number of simultaneous evaluations lowers memory load. A notable example already applied in some prominent MBA programs is the introduction of scribes—course assistants who record student participation in real time—which reduces the memory demands of evaluation and improves equity without requiring any change in evaluators' underlying beliefs. These remedies are broadly applicable wherever evaluators must form judgments from large, heterogeneous information sets under time pressure—in promotions, hiring, and performance reviews. More broadly, tools that provide evaluators with easy access to individual information at the time of decision represent a promising avenue for reducing memory-driven disparities.

This paper opens two directions for future research. First, how do past experiences shape recall distortions of new information? Teachers with larger historical gender gaps exhibit stronger stereotype-consistent recall under memory constraints, suggesting that past discriminatory behavior and biased memory retrieval are mutually reinforcing. Second, how do recall distortions vary with candidate ability? Stereotype-consistent retrieval may dominate for average students, while surprise effects could generate discrimination reversals for very counter-stereotypical candidates—such as exceptionally high-performing girls in math. Taking into account the factors that shape recall of individual-level information in theoretical and empirical analyses of discrimination could provide a deeper and more realistic account of its dynamics and persistence.

References

AlmaLaurea. 2023. "Focus Gender Gap, Rapporto 2023." Alma Laurea.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial bias in bail decisions." *The Quarterly Journal of Economics*, 133(4): 1885–1932.

Arrow, Kenneth J. 1998. "What has economics to say about racial discrimination?" *Journal of economic perspectives*, 12(2): 91–100.

Baltrunaite, Audinga, Alessandra Casarico, and Lucia Rizzica. 2024. "Women in economics: The role of gendered references at entry in the profession." *Bank of Italy Temi di Discussione (Working Paper) No*, 1438.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. "Attention discrimination: Theory and field experiments with monitoring information acquisition." *American Economic Review*, 106(6): 1437–75.

Becker, Gary S. 1957. "The economics of discrimination."

Benson, Alan, Danielle Li, and Kelly Shue. 2021. "Potential" and the gender promotion gap." Working paper.

Bertrand, Marianne, and Esther Duflo. 2017. "Field experiments on discrimination." *Handbook of economic field experiments*, 1: 309–393.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review*, 94(4): 991–1013.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit discrimination." *American Economic Review*, 95(2): 94–98.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The dynamics of discrimination: Theory and evidence." *American economic review*, 109(10): 3395–3436.

Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope. 2019. "Inaccurate statistical discrimination: An identification problem." National Bureau of Economic Research.

Bordalo, Pedro, Giovanni Burro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2025. "Imagining the future: memory, simulation, and beliefs." *Review of Economic Studies*, 92(3): 1532–1563.

Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer. 2023. "Memory and probability." *The Quarterly Journal of Economics*, 138(1): 265–311.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about gender." *American Economic Review*, 109(3): 739–73.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, competitiveness, and career choices." *The quarterly journal of economics*, 129(3): 1409–1447.

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

Charles, Constantin. 2022. "Memory Moves Markets." *Available at SSRN 4019728*.

Dube, Oeindrila, Sandy Jo MacArthur, and Anuj K Shah. 2025. "A cognitive view of policing." *The Quarterly Journal of Economics*, 140(1): 745–791.

Eberhardt, Markus, Giovanni Facchini, and Valeria Rueda. 2023. "Gender differences in reference letters: Evidence from the economics job market." *The Economic Journal*, 133(655): 2676–2708.

Egan, Mark, Gregor Matvos, and Amit Seru. 2022. "When Harry fired Sally: The double standard in punishing misconduct." *Journal of Political Economy*, 130(5): 1184–1248.

Esponda, Ignacio, Ryan Oprea, and Sevgi Yuksel. 2023. "Seeing what is representative." *The Quarterly Journal of Economics*, qjad020.

Fyock, Jack, and Charles Stangor. 1994. "The role of memory biases in stereotype maintenance." *British Journal of Social Psychology*, 33(3): 331–343.

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.

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating impartiality: The impact of" blind" auditions on female musicians." *American economic review*, 90(4): 715–741.

Greenwald, Anthony G, Debbie E McGhee, and Jordan LK Schwartz. 1998. "Measuring individual differences in implicit cognition: the implicit association test." *Journal of personality and social psychology*, 74(6): 1464.

Hastie, Reid, and Purohit A Kumar. 1979. "Person memory: Personality traits as organizing principles in memory for behaviors." *Journal of Personality and Social Psychology*, 37(1): 25.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge university press.

Jiang, Zhengyang, Hongqi Liu, Cameron Peng, and Hongjun Yan. 2022. "Investor Memory and Biased Beliefs: Evidence from the Field." *Available at SSRN*.

Kahana, Michael Jacob. 2012. *Foundations of human memory.* OUP USA.

Kessler, Judd B, Corinne Low, and Xiaoyue Shan. 2022. "Lowering the Playing Field: Discrimination through Contrast Effects."

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. "Systemic discrimination among large US employers." *The Quarterly Journal of Economics*, 137(4): 1963–2036.

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.

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland. 2015. "Expectations of brilliance underlie gender distributions across academic disciplines." *Science*, 347(6219): 262–265.

List, John A. 2004. "The nature and extent of discrimination in the marketplace: Evidence from the field." *The Quarterly Journal of Economics*, 119(1): 49–89.

Lowes, Sara, Nathan Nunn, James A Robinson, and Jonathan Weigel. 2015. "Understanding ethnic identity in africa: Evidence from the implicit association test (iat)." *American Economic Review*, 105(5): 340–345.

Macrae, C Neil, Miles Hewstone, and Riana J Griffiths. 1993. "Processing load and memory for stereotype-based information." *European Journal of Social Psychology*, 23(1): 77–87.

Mullainathan, Sendhil. 2002. "A memory-based model of bounded rationality." *The Quarterly Journal of Economics*, 117(3): 735–774.

Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt. 2019. "Poorly measured confounders are more useful on the left than on the right." *Journal of Business & Economic Statistics*, 37(2): 205–216.

Phelps, Edmund S. 1972. "The statistical theory of racism and sexism." *The american economic review*, 62(4): 659–661.

Radbruch, Jonas, and Amelie Schiprowski. 2022. "Interview sequences and the formation of subjective assessments."

Sarsons, Heather. 2017. "Interpreting signals in the labor market: evidence from medical referrals." *Job Market Paper*.

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. 2021. "Gender differences in recognition for group work." *Journal of Political Economy*, 129(1): 101–147.

Schmader, Toni, Jessica Whitehead, and Vicki H Wysocki. 2007. "A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants." *Sex roles*, 57(7): 509–514.

Sherman, Jeffrey W, and Leigh A Frost. 2000. "On the encoding of stereotype-relevant information under cognitive load." *Personality and Social Psychology Bulletin*, 26(1): 26–34.

Singh, Manasvini, and Atheendar Venkataramani. 2022. "Capacity strain and racial disparities in hospital mortality." National Bureau of Economic Research.

Stangor, Charles, and David McMillan. 1992. "Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures." *Psychological Bulletin*, 111(1): 42–61.

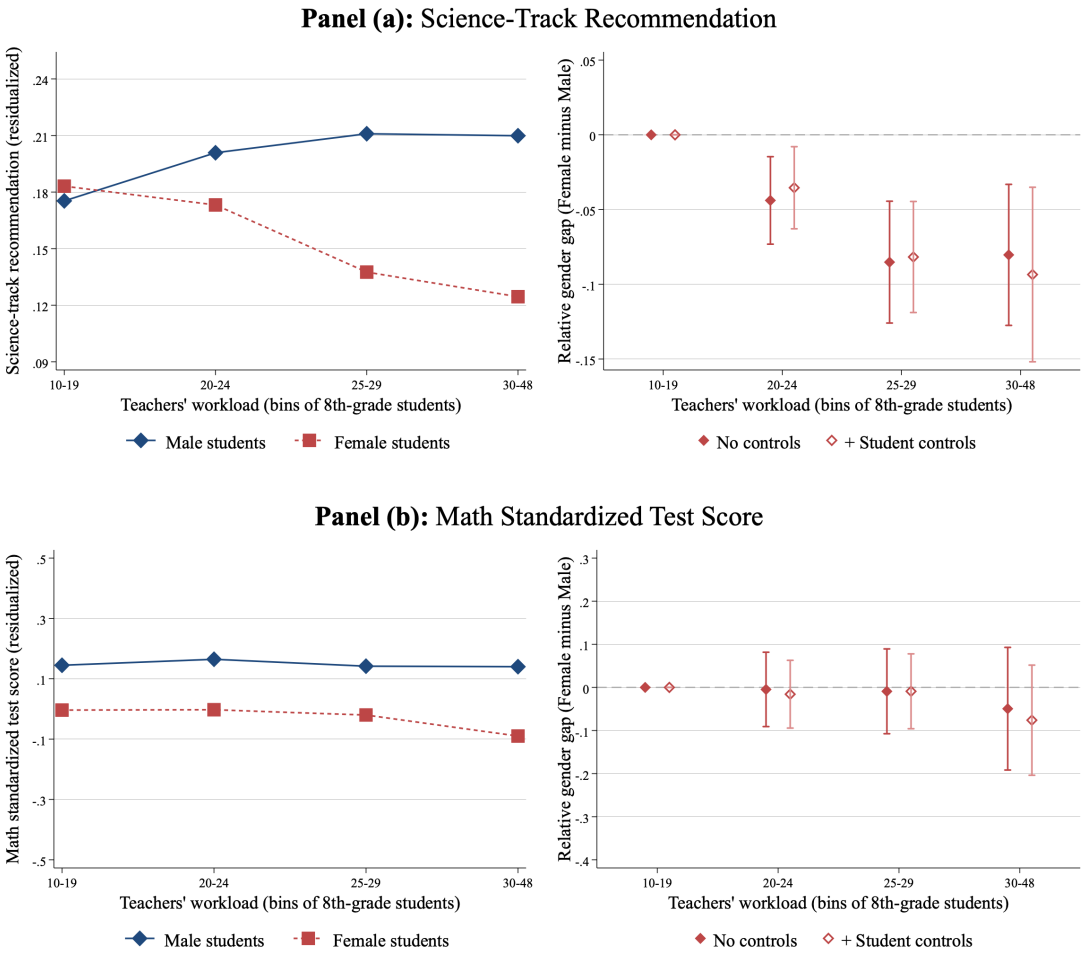
Trix, Frances, and Carolyn Psenka. 2003. "Exploring the color of glass: Letters of recommendation for female and male medical faculty." *Discourse & Society*, 14(2): 191–220.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty." *science*, 185(4157): 1124–1131.

Wu, Alice. 2018. "Gendered language on the economics job market rumors forum." *AEA Papers and Proceedings*, 108: 175–179.

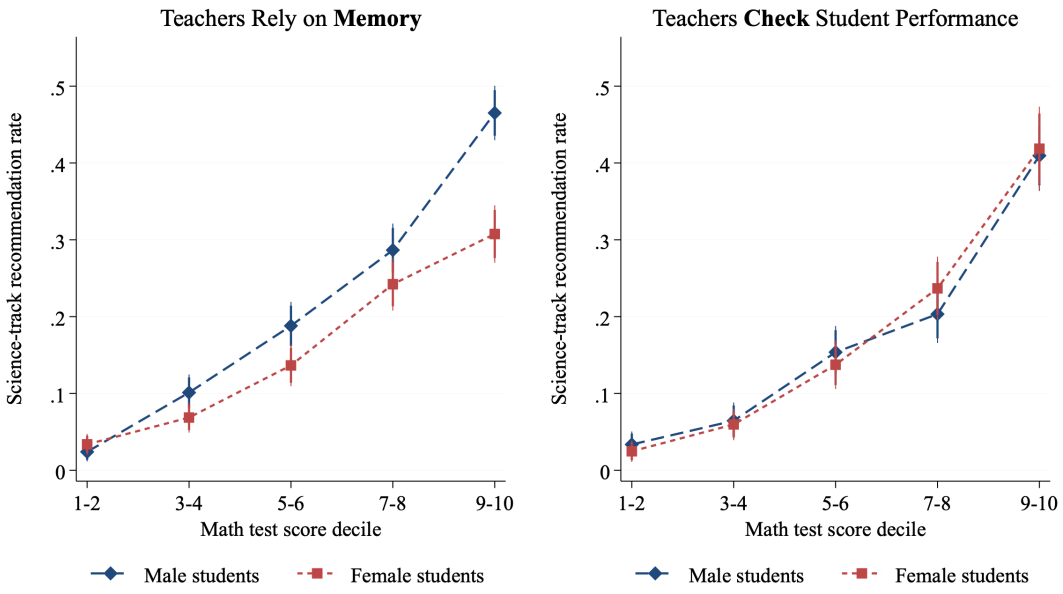
8 Figures and Tables

Figure 1. Gender Gaps and Teachers' Workload (*Administrative data*)



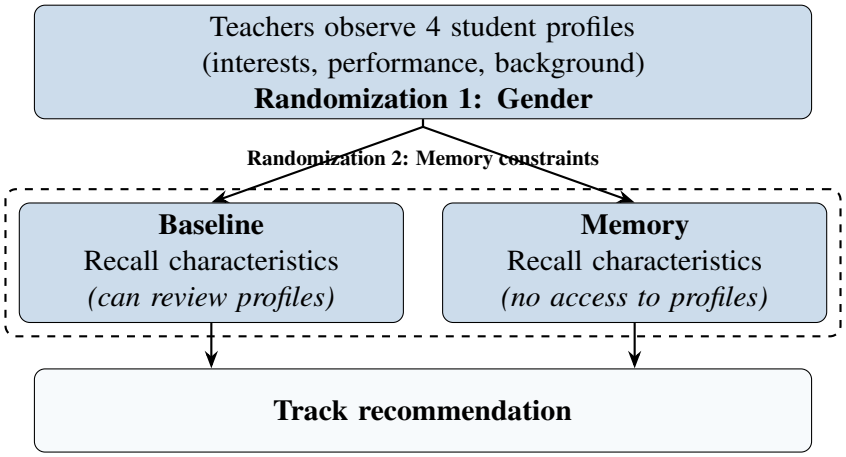
Notes: The left panels show the residualized probability of being recommended for the scientific track (Panel a) and standardized math test scores (Panel b) for boys and girls as a function of their math teacher's number of students to recommend. Each figure is constructed by regressing the outcome on student controls, teacher and year fixed effects, and plotting the residuals by gender after adding back the mean of the dependent variable. The right panels show the γ_s coefficients from specification 1, without controls (dark blue) and with the full set of controls described in equation 1. Caps show 95 percent confidence intervals with clustering by teacher.

Figure 2. Gender Gaps for Teachers Who Rely on Memory vs. Check Student Performance (*Administrative data*)



Notes: The figure displays the share of male and female students recommended for the scientific track by grade 5 standardized math test score decile, separately for teachers who report relying on memory rather than consulting records (*Memory*, left panel) and teachers who report checking students' academic performance in class registers (*Baseline*, right panel). Shaded bands show 90 percent and 95 percent confidence intervals. Sample: core-subject teachers who completed the survey.

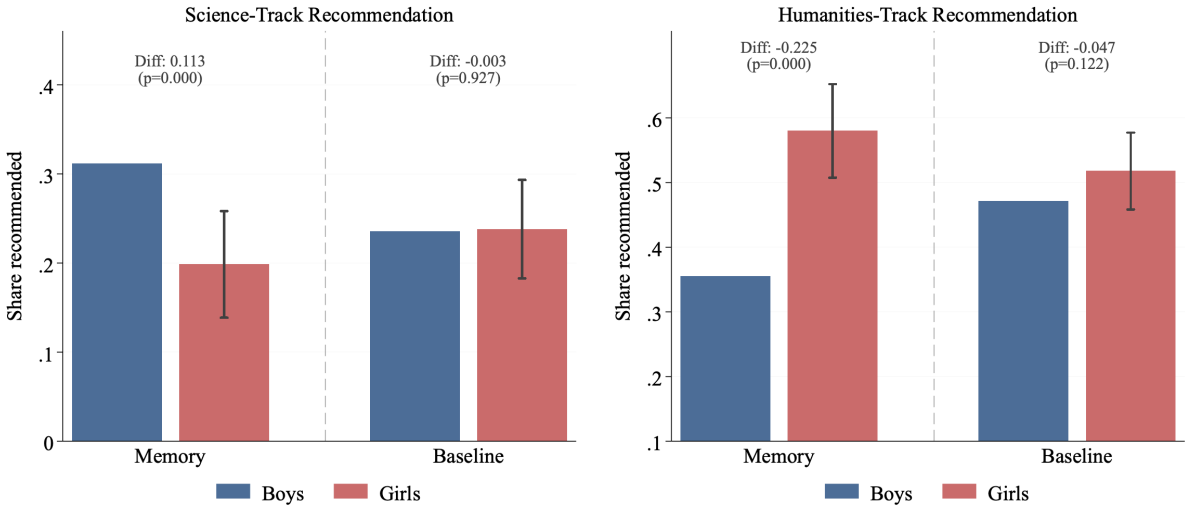
Figure 3. Experimental Design



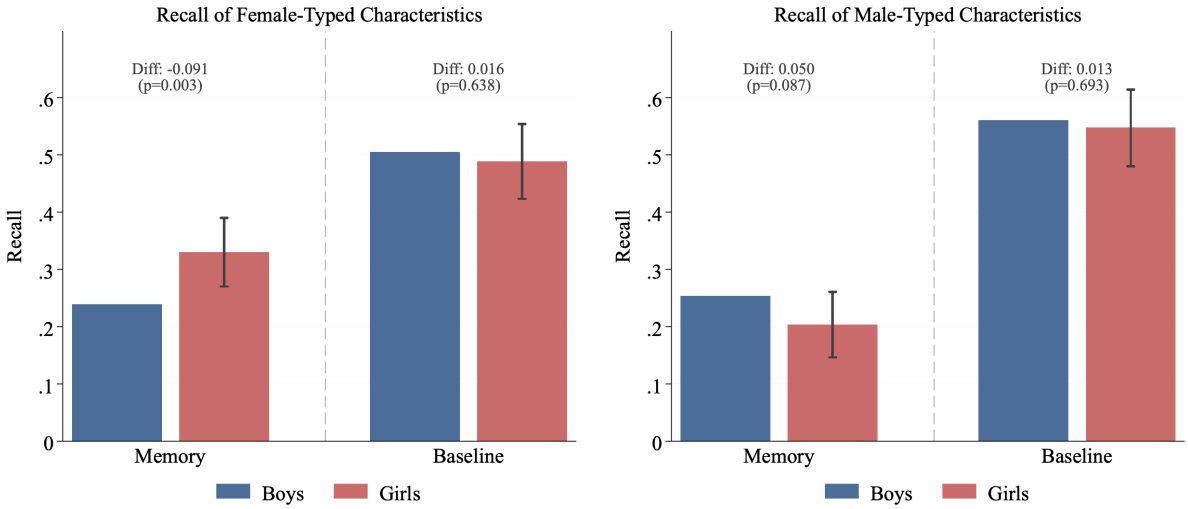
Notes: This figure shows the teacher experiment design. I randomized whether teachers must retrieve student information from memory or can review it before deciding (*Memory* vs. *Baseline* conditions) across teachers, while student gender is randomized at the student profile level. Each teacher observes 4 student profiles. After observing the profiles, teachers recall each student's characteristics and provide a track recommendation (decision screens shown in Figure C2).

Figure 4. High School Track Recommendations and Recall of Student Characteristics (*Teacher Experiment*)

Panel (a): Track recommendations

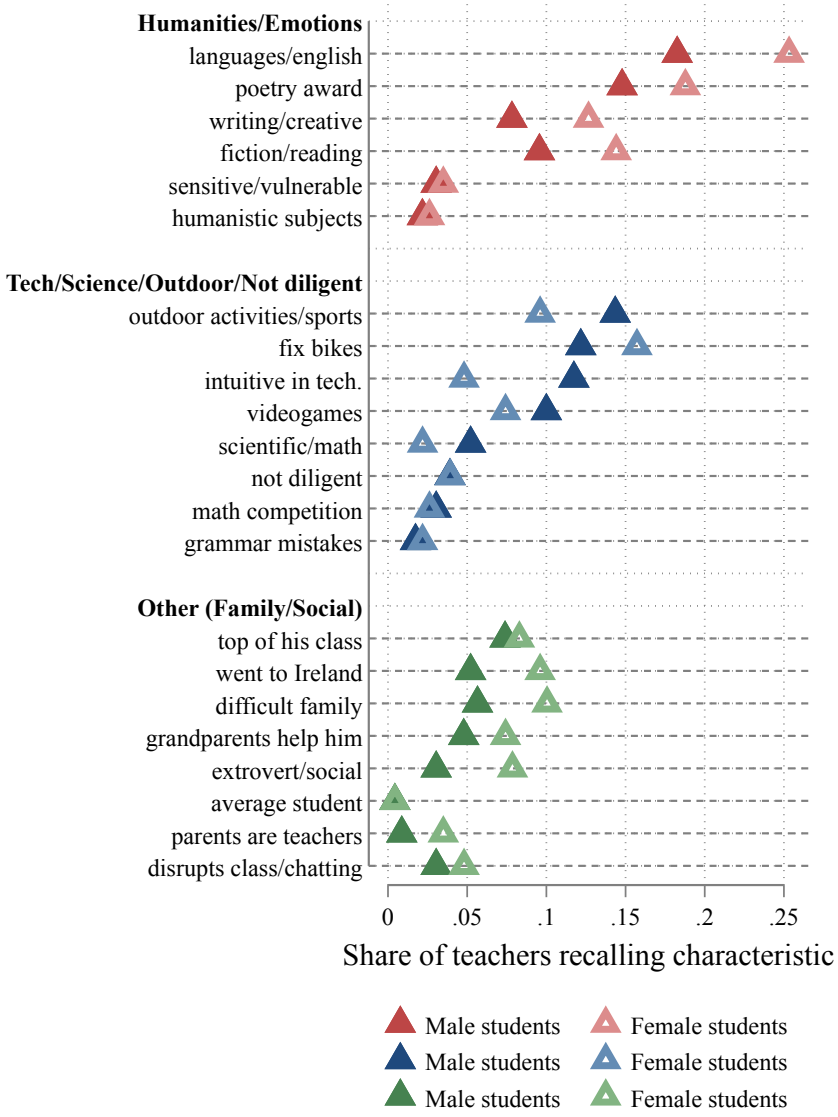


Panel (b): Recall of student characteristics



Notes: Panel (a) shows the share of students recommended to the science track (left) and humanities tracks (right) by student gender and memory condition. Humanities tracks correspond to female-typed tracks, as described in Section 5.2. Panel (b) shows the probability that teachers recall female-typed (left) and male-typed (right) characteristics by student gender and memory condition. In the *Memory* condition, teachers retrieve student characteristics from memory; in the *Baseline* condition, teachers may review profiles before deciding. Characteristics are classified as female- or male-typed as described in Section 5. Student profiles are identical across gender and memory conditions. Caps show 95 percent confidence intervals.

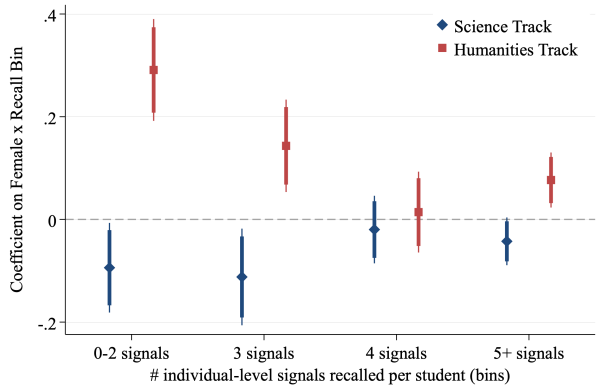
Figure 5. Recalled Characteristics by Student Gender, Teachers in *Memory* group (*Teacher Experiment*)



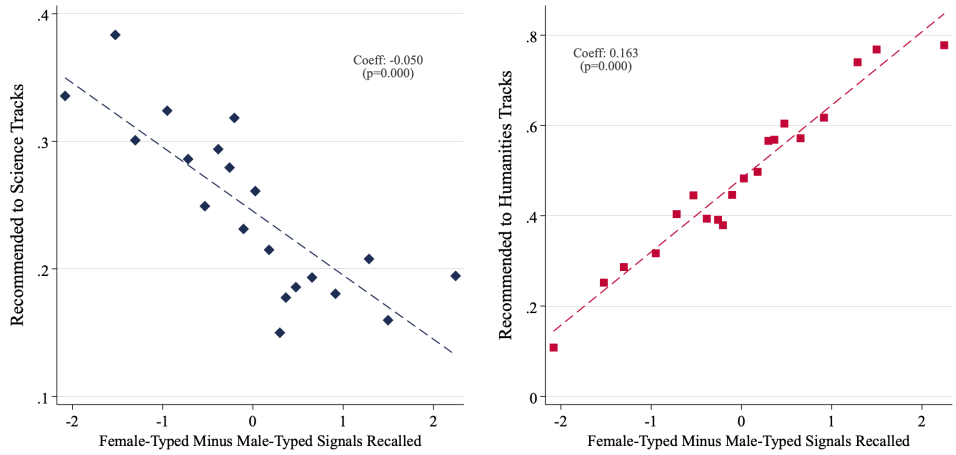
Notes: Share of teachers recalling each student characteristic when evaluating male and female students with identical profiles, in the *Memory* condition. Filled markers indicate male students; hollow markers indicate female students. Characteristics are grouped into Humanities/Emotions (red), Tech/Science/Outdoor/Not Diligent (blue), and Other/Family/Social (green). A characteristic is coded as recalled whether it is remembered as belonging to a student.

Figure 6. Recall of Individual Information and Biases in Track Recommendations (*Teacher Experiment*)

Panel (a): Limited Recall and Track Recommendations



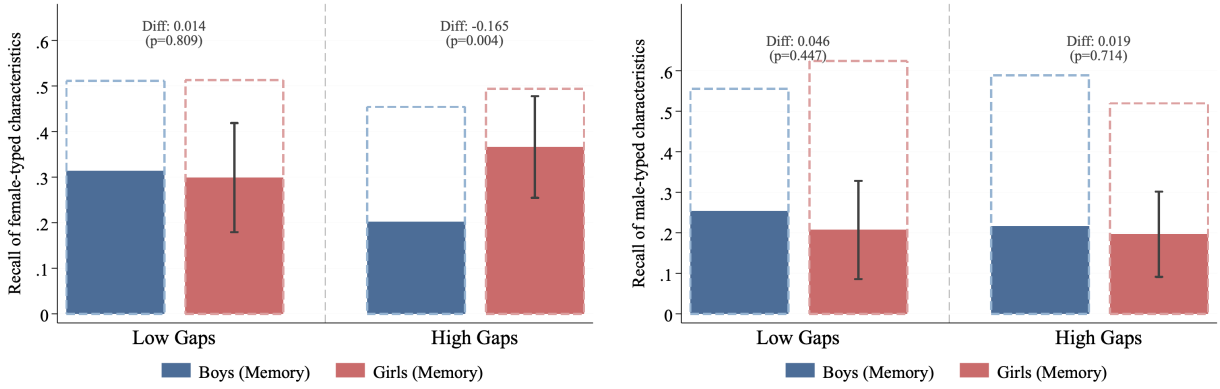
Panel (b): Biased Recall and Track Recommendations



Notes: Panel (a) shows the coefficient on Female × Recall Bin from a regression of track recommendations on indicators for the number of individual-level signals recalled, their interaction with an indicator for female students, the gender gap in recalled characteristics (difference between female and male-typed signals recalled), and teacher and profile fixed effects. Dark blue diamonds: science-track recommendations; red squares: humanities-track recommendations. Caps show 95 percent and 90 percent confidence intervals, clustered at the teacher level. Panel (b) shows binned scatterplots of the relationship between track recommendations and the gender gap in recalled characteristics (female-typed minus male-typed signals recalled), controlling for total signals recalled, and teacher and profile fixed effects.

Figure 7. Recall Biases and Teachers' Historical Gender Gaps in Scientific Track Recommendations (*Teacher Experiment and Administrative Data*)

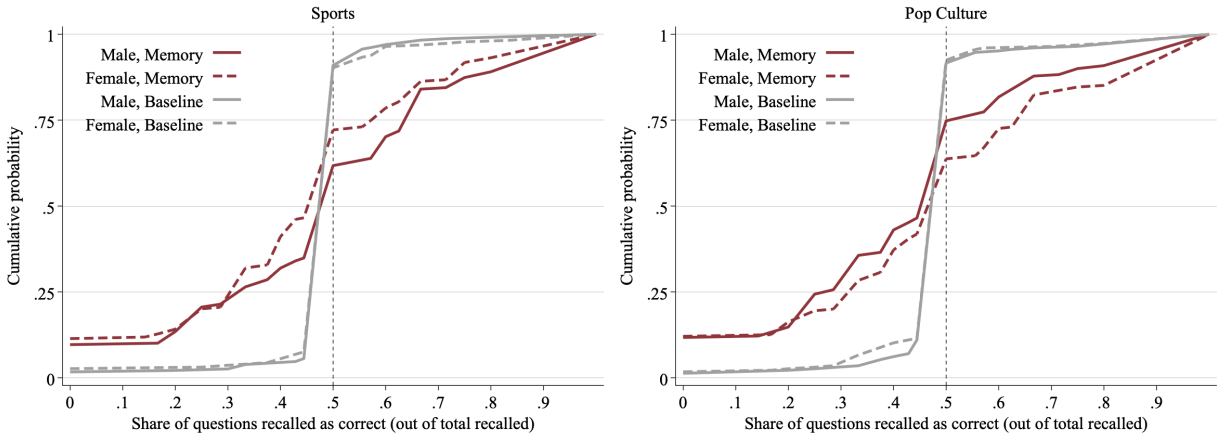
Panel (a): Recall any Female-Typed Characteristics **Panel (b):** Recall any Male-Typed Characteristics



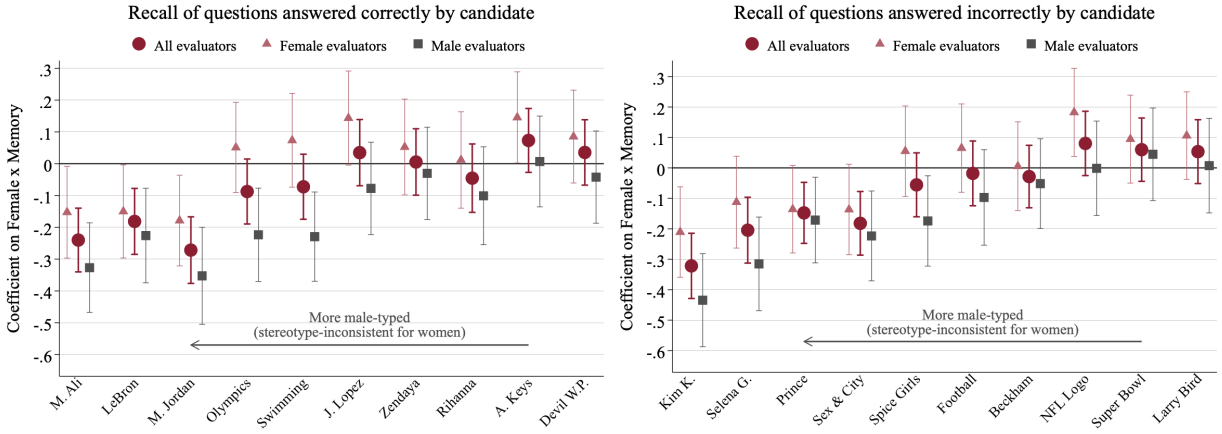
Notes: Each panel shows the probability of recalling female-typed and male-typed characteristics for teachers with above- and below-average historical gender gaps in actual science-track recommendations, separately by memory condition. Filled bars indicate the *Memory* condition; dashed bars indicate the *Baseline* condition. Caps show 95 percent confidence intervals. Teachers are classified as high- or low-gap based on whether their historical gender gap in scientific track recommendations exceeds the sample mean. Sample: core-subject teachers who completed the experiment.

Figure 8. Biased Recall of Quiz 1 Performance (*Prolific Experiment*)

Panel (a): Cumulative distribution of the share of Quiz 1 questions recalled as correct, among all questions recalled.



Panel (b): Question-level recall estimates, by evaluator gender. Questions ordered from more male-typed (left) to more female-typed (right).



Notes: Panel (a) shows the cumulative distribution of recalled performance share – the fraction of recalled Quiz 1 answers placed in the correct box, out of total questions recalled – by candidate gender, treatment condition, and domain. Maroon lines correspond to the *Memory* condition; gray lines to the *Baseline* condition. Solid lines correspond to male candidates; dashed lines to female candidates. The dashed vertical line at 0.5 indicates unbiased recall. Panel (b) shows question-level estimates of the coefficient on Female $\times$ Memory from regressions of residualized recall on candidate gender, memory condition, and their interaction, separately for correct answers (left) and incorrect answers (right). Residuals partial out gender specific question difficulty (alone and interacted by memory treatment) and evaluator demographics. Questions are ordered from more male-typed (left) to more female-typed (right) based on the gender gap in correct answer rates in the reference population (see Figure E4). Maroon circles: all evaluators; rose triangles: female evaluators; gray squares: male evaluators. Vertical lines are 95 percent confidence intervals clustered at the evaluator level.

Table 1. Teachers' Workload and Science-Track Recommendations (*Administrative data*)

	All students			High-Achievers		
	(1)	(2)	(3)	(4)	(5)	(6)
DV: Recommendation for Science Track						
Female	-0.0237***	-0.00117	-0.00358	-0.0790***	-0.0276	-0.0276
	(0.00581)	(0.00752)	(0.00859)	(0.0128)	(0.0215)	(0.0217)
High Workload	-0.00328	0.0169	0.0180	-0.0168	0.0190	0.0260
	(0.00969)	(0.0110)	(0.0146)	(0.0238)	(0.0268)	(0.0348)
High Workload × Female		-0.0420***	-0.0477***		-0.0834***	-0.0823***
		(0.0113)	(0.0121)		(0.0267)	(0.0264)
Mean DV Males	0.200	0.200	0.200	0.465	0.465	0.465
Math Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Std.test scores	Yes	Yes	Yes	Yes	Yes	Yes
Stud. Controls	No	No	Yes	No	No	Yes
Classmates Controls	No	No	Yes	No	No	Yes
R-squared	0.307	0.307	0.228	0.294	0.295	0.308
N. Teachers	313	313	313	312	312	312
Observations	17821	17821	17821	5018	5018	5018

Notes: This table shows coefficients α_2, β, and γ from estimation of model 2. One observation is a student assigned to a math teacher in a given year. High Workload is a dummy equal to one if the number of eighth-grade students the math teacher must recommend in a given year exceeds the median (one standard deviation corresponds to approximately 7 students). The sample includes student cohorts matched with their math teachers, from Carlana (2019). Student controls include standardized math and Italian test scores, mother's education dummies, father's occupation dummies, immigrant status, class size, the total number of students assigned to the teacher in all grades (grades 6 to 8), and the number of years spent with the teacher (1, 2, or 3). Classmate controls include average standardized test scores in math and Italian, the fraction of females, immigrants, and high-socioeconomic-status students in the class (excluding the student). Standard errors are clustered at the teacher level. *p<0.1; **p<0.05; ***p<0.01.

Table 2. Gender Gap in Scientific Track Recommendations: Teachers Relying on Memory vs. Checking Student Performance (*Administrative data*)

	DV: Scientific Track Recommendation				
	(1)	(2)	(3)	(4)	(5)
Female	-0.044***	-0.020	-0.020	-0.020	-0.020
	(0.009)	(0.012)	(0.012)	(0.012)	(0.012)
Memory		0.039***	0.026**	0.021*	0.019
		(0.014)	(0.012)	(0.012)	(0.012)
Memory × Female		-0.039**	-0.038**	-0.038**	-0.038**
		(0.016)	(0.016)	(0.016)	(0.016)
Mean control	0.187	0.187	0.187	0.187	0.187
Year FE	Yes	Yes	Yes	Yes	Yes
Std. Test Scores & Stud. Controls	Yes	Yes	Yes	Yes	Yes
School FE	No	No	Yes	Yes	Yes
Teacher Controls	No	No	No	Yes	Yes
Controls for Teachers' Attitudes	No	No	No	No	Yes
R^2	0.131	0.133	0.150	0.150	0.151
Observations	15080	15080	15080	15080	15080

Notes: This table reports coefficients β_1, β_2, β_3 from the estimation of model 3. Memory is a dummy equal to 1 if a teacher reports relying on recollection rather than checking the class register when assigning recommendations. Teacher controls include subject studied, subject taught, age, gender, father's education, type of contract, an indicator for school in the North, and an indicator for born in the North. Student-level controls include math and Italian standardized test scores, immigrant status, and mother's college education. Teacher attitudes controls include IAT scores and teachers' reported beliefs about the importance of grades, student attitudes, and parental background in shaping track recommendations. The sample includes eighth-grade students matched with their core-subject teachers (math and Italian). Standard errors are clustered at the teacher level. *p<0.1; **p<0.05; ***p<0.01.

Table 3. Teachers' Experiment: Balance

Variable	(1) Baseline	(2) SD	(3) Memory	(4) SD	(5) P-val.	(6) Std. Diff.
Panel A: Teachers Background						
Female	0.838	(0.369)	0.816	(0.388)	(0.544)	-0.041
School in Northern Italy	0.730	(0.445)	0.715	(0.453)	(0.719)	-0.024
Age	50.129	(9.574)	50.340	(9.928)	(0.820)	0.015
Highest level of education: PhD	0.110	(0.313)	0.100	(0.301)	(0.753)	-0.022
Highest level of education: Master	0.840	(0.367)	0.858	(0.350)	(0.619)	0.035
Highest level of education: BA or less	0.050	(0.219)	0.042	(0.201)	(0.698)	-0.027
Low education mother	0.243	(0.430)	0.204	(0.404)	(0.348)	-0.066
Years in teaching	16.758	(10.337)	17.020	(10.697)	(0.794)	0.018
Permanent contract	0.808	(0.395)	0.811	(0.392)	(0.930)	0.006
Humanistic subject	0.477	(0.501)	0.498	(0.501)	(0.667)	0.029
Scientific subject	0.290	(0.455)	0.300	(0.459)	(0.834)	0.014
Other subject	0.228	(0.421)	0.193	(0.396)	(0.368)	-0.061
Viceprincipal as role in the school	0.033	(0.180)	0.029	(0.168)	(0.799)	-0.017
Responsible for orientation as role in the school	0.066	(0.249)	0.077	(0.268)	(0.656)	0.030
Panel B: Attitudes						
Ability to motivate difficult students (0-5)	3.702	(0.839)	3.653	(0.818)	(0.547)	-0.042
Importance of Grades	46.217	(16.502)	45.680	(17.728)	(0.750)	-0.022
Importance of Attitudes	34.167	(12.864)	35.634	(14.759)	(0.280)	0.075
Importance of Parents	19.615	(19.723)	18.686	(18.977)	(0.626)	-0.034
Remember students' performance	0.412	(0.493)	0.430	(0.496)	(0.708)	0.026
Check grades during the teaching meeting	0.344	(0.476)	0.295	(0.457)	(0.292)	-0.074
Do not check grades but think more holistically	0.656	(0.476)	0.632	(0.483)	(0.612)	-0.035
IAT	0.376	(0.726)	0.331	(0.678)	(0.555)	-0.045
Past Students: Gender Gap Science-Track Rec.	3.591	(10.946)	4.191	(11.297)	(0.607)	0.038
Past Students (high GPA): Gender Gap Science-Track Rec.	18.900	(24.263)	19.770	(26.410)	(0.743)	0.024
Past Students: Gender Gap Science-Track Choice	3.742	(12.934)	4.138	(12.157)	(0.764)	0.022
Past Students (high GPA): Gender Gap Science-Track Choice	15.984	(24.098)	16.503	(27.260)	(0.847)	0.014
Observations	241		207			
Joint F-test				0.965		

Notes: Observable characteristics for teachers who completed the experiment, separately for the *Baseline* and *Memory* groups. Normalized differences follow Imbens and Rubin (2015).

Table 4. Track Recommendations and Memory Constraints (*Teacher Experiment*)

	DV: Science-Tracks			DV: Humanities Tracks		
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.0562***	0.000845	0.00117	0.126***	0.0463**	0.0496***
	(0.0160)	(0.0159)	(0.0164)	(0.0194)	(0.0186)	(0.0186)
Memory	0.0181	0.0796***	0.0757***	-0.0287	-0.115***	-0.111***
	(0.0168)	(0.0240)	(0.0246)	(0.0175)	(0.0260)	(0.0259)
Memory × Female		-0.125***	-0.122***		0.175***	0.169***
		(0.0329)	(0.0330)		(0.0397)	(0.0396)
Mean Baseline Males	0.235	0.235	0.235	0.471	0.471	0.471
Mean Baseline Females	0.238	0.238	0.238	0.518	0.518	0.518
Student FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	No	No	Yes
R^2	0.375	0.380	0.384	0.346	0.353	0.359
N. teachers	448	448	448	448	448	448
Observations	1761	1761	1761	1761	1761	1761

Notes: This table reports coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ from estimation of equation (4). One observation is a student profile evaluated by a teacher (each teacher evaluates 4 profiles). Columns 1–3 report results for the scientific track; columns 4–6 for female-typed humanities tracks (classical, linguistic, artistic, and human sciences). Columns 1 and 4 report average gender gaps controlling for student fixed effects; columns 2 and 5 add the memory interaction; columns 3 and 6 add teacher controls. Teacher controls include birth year, gender, subject taught, father's education, contract type, and indicators for school and birth region in Northern Italy. Standard errors are clustered at the teacher level. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 5. Predicted Quiz 2 Performance and Biased Recall of Quiz 1 Performance (*Prolific Experiment*)

	DV: Predicted performance in Quiz 2			DV: Share of Quiz 1 questions recalled as correct		
	(1)	(2)	(3)	(4)	(5)	(6)
	All Evaluators	Male Evaluators	Female Evaluators	All Evaluators	Male Evaluators	Female Evaluators
Stereotype-Inconsistent	-0.0168***	-0.0221***	-0.0122	0.00592	0.0229*	-0.0108
	(0.00540)	(0.00810)	(0.00765)	(0.00932)	(0.0137)	(0.0134)
Memory	0.0191***	0.0274**	0.0123	0.00496	0.0153	-0.00584
	(0.00708)	(0.0107)	(0.00969)	(0.0140)	(0.0223)	(0.0183)
Stereotype-Inconsistent × Memory	-0.0256***	-0.0392***	-0.0132	-0.0533**	-0.0826**	-0.0380
	(0.00901)	(0.0138)	(0.0122)	(0.0221)	(0.0353)	(0.0286)
DV Mean (St.-Consistent, Baseline)	0.535	0.537	0.534	0.496	0.491	0.501
R-squared	0.0705	0.102	0.0843	0.0247	0.0437	0.0394
Demographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Domain FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2020	998	988	1812	846	933

Notes: The table reports estimates of equation 5. One observation is an evaluator by domain pair (two observations per evaluator). The dependent variable in columns 1–3 is predicted Quiz 2 performance (predicted share correct answers in Quiz 2); in columns 4–6 it is the share of recalled Quiz 1 answers placed in the correct box. *Memory* equals one for evaluators in the Memory condition. *Stereotype-Inconsistent* equals one if the candidate's gender is inconsistent with the domain (female in sports or male in pop culture). Columns 1 and 4 pool all evaluators; columns 2 and 5 restrict to male evaluators; columns 3 and 6 restrict to female evaluators. Controls include evaluator gender, age, education, employment status, political affiliation, Prolific approval history. Standard errors are clustered at the evaluator level. *$p<0.1$; **$p<0.05$; ***$p<0.01$.

Table 6. Biased Recall of Quiz 1 Questions: Question-Level Analysis (*Prolific Experiment*)

	DV: Recall as correct (correct box) Sample: questions answered correctly			DV: Recall as incorrect (incorrect box) Sample: questions answered incorrectly		
	(1)	(2)	(3)	(4)	(5)	(6)
	All evaluators	Male evaluators	Female evaluators	All evaluators	Male evaluators	Female evaluators
Stereotype-Inconsistent	0.0207	0.0516	0.00115	0.0298	0.0704*	0.00132
	(0.0267)	(0.0409)	(0.0341)	(0.0264)	(0.0412)	(0.0331)
Memory	-0.445***	-0.277***	-0.594***	-0.470***	-0.474***	-0.467***
	(0.0465)	(0.0663)	(0.0642)	(0.0239)	(0.0361)	(0.0320)
Stereotype-Inconsistent × Memory	-0.0598**	-0.146***	0.0157	-0.0873***	-0.193***	-0.00420
	(0.0287)	(0.0433)	(0.0378)	(0.0281)	(0.0417)	(0.0371)
Mean Baseline	0.811	0.765	0.856	0.815	0.768	0.861
Domain FE	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Gender-specific ability	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.283	0.304	0.264	0.250	0.263	0.235
N. Obs	10090	4990	4930	10090	4990	4930

Notes: The table reports the estimated probability that an evaluator accurately retrieves a given question, separately for questions the candidate answered correctly (columns 1–3) and incorrectly (columns 4–6). One observation is a question–evaluator pair. *Stereotype-Inconsistent* equals one if the observed signal is stereotype-inconsistent with respect to the candidate's gender: a correct answer in sports or an incorrect answer in pop-culture for a female candidate, and a correct answer in pop-culture or an incorrect answer in sports for a male candidate. *Memory* equals one for evaluators in the Memory condition. Demographic controls include evaluator gender, age, education, employment status, political affiliation, and number of prior approvals on Prolific. Question-level controls include gender-specific question difficulty—measured as the share of correct answers for each question by gender in the reference population—and its interaction with the *Memory* indicator. Domain fixed effects are included. Standard errors are clustered at the evaluator level. *$p<0.1$; **$p<0.05$; ***$p<0.01$.

Online Appendix

Contents

A Observational Evidence (Section 3), Additional Figures and Tables **47**

A.1 Additional Tables and Figures 47

A.2 Sensitivity and Robustness 59

B Observational Evidence (Section 4), Additional Figures and Tables **64**

B.1 Additional Tables and Figures 64

C Teacher Experiment (Section 5) **73**

C.1 Experimental Design and Procedures 73

C.2 Attrition 77

C.3 Additional Tables and Figures 79

C.4 Accurate recalls and false memories 84

C.5 Hypothetical Profiles and Actual Students 86

C.6 Recommendations for All Tracks across *Memory* and *Baseline* condition 88

D Back-of-the-Envelope Decomposition of the Recall Channels **90**

D.1 Teacher Experiment 91

D.2 Prolific Experiment 92

E Prolific Experiment **94**

E.1 Experimental Design and Procedures 94

E.2 Additional Tables and Figures 98

F Conceptual Framework **105**

A Observational Evidence (Section 3), Additional Figures and Tables

A.1 Additional Tables and Figures

Table A1. Summary statistics of students and their math teachers

Variable	Mean	SD	Min	Max	N
Track Recommendation and Choice:					
Top-tier scientific recommendation	0.183	0.386	0.000	1.000	17,821
Vocational track recommendation	0.389	0.488	0.000	1.000	17,821
Top-scientific choice	0.270	0.444	0.000	1.000	15,089
Top-classical choice	0.052	0.222	0.000	1.000	15,089
Technical choice	0.311	0.463	0.000	1.000	15,089
Medium humanities choice	0.212	0.409	0.000	1.000	15,089
Vocational choice	0.156	0.363	0.000	1.000	15,089
Other Characteristics:					
Follow teacher recommendation: top-tier scientific	0.730	0.444	0.000	1.000	17,821
Std. test score Reading	0.060	0.988	-3.862	2.076	16,221
Std. test score Math	0.082	1.004	-2.987	2.746	16,231
High Education Mother (university)	0.548	0.498	0.000	1.000	13,521
High Occupation Mother	0.236	0.424	0.000	1.000	13,093
Immigrant	0.192	0.394	0.000	1.000	17,821
Female	0.482	0.500	0.000	1.000	17,821
Math grade 8	7.024	1.293	1.000	10.000	16,600
Italian grade 8	7.076	1.137	1.000	10.000	16,604
Number of Years with Math Teacher	2.556	0.748	1.000	3.000	17,821
Class size	22.252	2.664	12.000	29.000	17,821
Characteristics of Students' Math Teachers:					
Teacher: female	0.793	0.405	0.000	1.000	17,821
Teacher: age	47.261	16.801	0.000	66.000	17,821
Teacher: graduated with honors	0.154	0.361	0.000	1.000	17,821
Teacher: born in North	0.600	0.490	0.000	1.000	17,821
Teacher: permenent contract	0.847	0.360	0.000	1.000	17,821
Teacher: has children	1.260	1.109	0.000	5.000	17,821
Teacher: number of 8th grade students	23.463	5.461	12.000	48.000	17,821
Teacher: number of students	53.256	15.000	15.000	144.000	17,821

Notes: This table reports summary statistics about 8th-grade students and their assigned math teachers in the main administrative sample.

Table A2. Characteristics of students and students' teachers by student gender

Variable	(1) Male	(2) SD	(3) Female	(4) SD	(5) Diff.	(6) P-val.
Students' Characteristics						
Top-tier scientific recommendation	0.200	(0.400)	0.164	(0.370)	-0.037	(0.000)***
Vocational track recommendation	0.425	(0.494)	0.350	(0.477)	-0.075	(0.000)***
Top-scientific choice	0.317	(0.465)	0.221	(0.415)	-0.096	(0.000)***
Top-classical choice	0.034	(0.182)	0.071	(0.256)	0.036	(0.000)***
Technical choice	0.407	(0.491)	0.210	(0.407)	-0.197	(0.000)***
Medium humanities choice	0.092	(0.289)	0.336	(0.472)	0.244	(0.000)***
Vocational choice	0.149	(0.356)	0.163	(0.369)	0.014	(0.021)**
Follow teacher recommendation: top-tier scientific	0.703	(0.457)	0.760	(0.427)	0.056	(0.000)***
Std. test score Reading	-0.028	(0.946)	0.144	(0.931)	0.172	(0.000)***
Std. test score Math	0.159	(0.962)	-0.015	(0.947)	-0.174	(0.000)***
High Education Mother (university)	0.412	(0.492)	0.420	(0.494)	0.007	(0.312)
High Occupation Mother	0.170	(0.376)	0.177	(0.381)	0.007	(0.240)
Immigrant	0.193	(0.394)	0.190	(0.393)	-0.002	(0.720)
Number of Years with Math Teacher	2.542	(0.755)	2.571	(0.741)	0.029	(0.009)***
Mean math score classmates	0.054	(0.388)	0.053	(0.391)	-0.002	(0.790)
Mean reading score classmates	0.038	(0.370)	0.037	(0.376)	-0.001	(0.844)
Share immigrant classmates	0.196	(0.141)	0.194	(0.142)	-0.002	(0.311)
Share classmates with highly educated mother	0.412	(0.228)	0.414	(0.229)	0.002	(0.508)
Share classmates with high occupation father	0.170	(0.160)	0.173	(0.161)	0.003	(0.230)
Students' Teachers' Characteristics						
Teacher: female	0.792	(0.406)	0.794	(0.405)	0.002	(0.783)
Teacher: age	47.242	(16.827)	47.282	(16.773)	0.040	(0.874)
Teacher: graduated with honors	0.157	(0.364)	0.150	(0.357)	-0.007	(0.177)
Teacher: born in North	0.602	(0.490)	0.598	(0.490)	-0.003	(0.635)
Teacher: permenent contract	0.846	(0.361)	0.848	(0.359)	0.002	(0.655)
Teacher: has children	1.255	(1.106)	1.265	(1.112)	0.010	(0.542)
Teacher: number of 8th grade students	23.443	(5.404)	23.485	(5.521)	0.042	(0.608)
Teacher: number of students	53.303	(14.970)	53.206	(15.034)	-0.097	(0.667)
Observations	9,232		8,589		17,821	

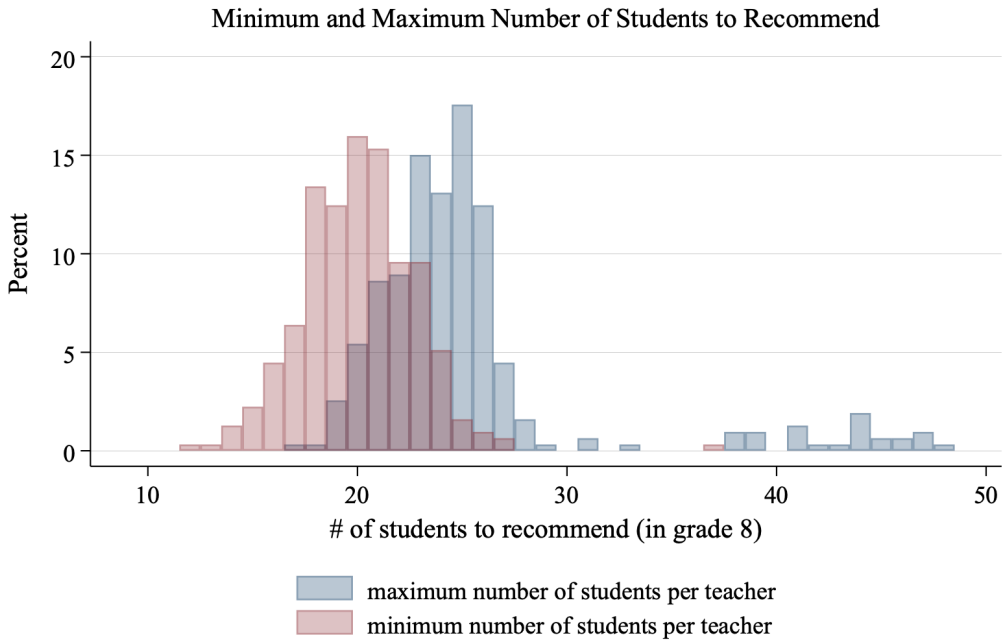
Notes: This table reports summary statistics about 8th-grade students and their assigned math teachers in the main administrative sample by student gender.

Table A3. Scientific track recommendation and choice

	DV: scientific track choice				
	(1)	(2)	(3)	(4)	(5)
1(Scientific track recommendation)	0.688***	0.666***	0.567***	0.562***	0.556***
	(0.0125)	(0.0136)	(0.0158)	(0.0159)	(0.0160)
1(Scientific track recommendation) × Female	-0.00827	0.00000170	0.00292	0.00111	-0.00186
	(0.0163)	(0.0162)	(0.0160)	(0.0159)	(0.0162)
Female	-0.0604***	-0.0609***	-0.0505***	-0.0489***	-0.0575***
	(0.00701)	(0.00706)	(0.00697)	(0.00712)	(0.00702)
R-squared	0.407	0.440	0.467	0.471	0.484
N. Obs	15089	15089	15089	15089	15088
Teacher FE	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes
Std. test scores	No	No	Yes	Yes	Yes
Teacher assigned grades	No	No	Yes	Yes	Yes
Students' controls	No	No	No	Yes	Yes
Classmates Characteristics	No	No	No	Yes	Yes
School by Year FE	No	No	No	No	Yes
Mean DV	0.270	0.270	0.270	0.270	0.270
Sd Dependent Variable	0.444	0.444	0.444	0.444	0.444

Notes: Standard errors clustered at the teacher by year level. Sample: students in grade 8 in years from 2011-16 in 98 schools, matched with their math and literature teachers (from Carlana 2019). Students' controls include: immigrant, second generation immigrant, education mother, occupation father.

Figure A1. Minimum and maximum number of 8th grade students assigned to math teachers across years



Notes: This figure shows the maximum and minimum numbers of 8th-grade students assigned to math teachers in the main observational sample.

Table A4. Student and Classmate Characteristics and Teacher Workload

	(1) Math Score	(2) Ita. Score	(3) Immigrant	(4) High Edu. Mother	(5) High Occ. Father
Panel A: Student Characteristics					
Female	-0.161***	0.216***	0.00713	0.00625	0.000988
	(0.0232)	(0.0244)	(0.00968)	(0.0123)	(0.0103)
High Workload	0.0163	0.0284	-0.00336	0.0154	-0.00288
	(0.0317)	(0.0331)	(0.0115)	(0.0177)	(0.0140)
High Workload × Female	-0.0472	-0.0443	-0.0157	-0.0230	-0.00201
	(0.0316)	(0.0309)	(0.0125)	(0.0165)	(0.0152)
Mean DV Males	0.177	-0.0310	0.193	0.551	0.235
R-squared	0.150	0.130	0.0914	0.136	0.145
# Students	16231	16221	17821	13518	13092
# Teachers	313	313	313	304	304
	(1) Math Score	(2) Ita. Score	(3) Immigrant	(4) High Edu. Mother	(5) High Occ. Father
Panel B: Classmate Characteristics					
Female	0.00349	-0.00985**	0.000480	-0.000132	-0.000619
	(0.00421)	(0.00438)	(0.00188)	(0.00202)	(0.00165)
High Workload	-0.00540	0.00638	-0.0106	0.0138	-0.000852
	(0.0258)	(0.0260)	(0.0104)	(0.0140)	(0.00911)
High Workload × Female	-0.00165	0.00767	-0.00264	0.000419	0.00322
	(0.00684)	(0.00673)	(0.00239)	(0.00321)	(0.00291)
Mean DV Males	0.0542	0.0379	0.196	0.412	0.170
R-squared	0.712	0.662	0.712	0.682	0.687
# Students	17821	17821	17821	17821	17821
# Teachers	313	313	313	313	313
Math Teacher FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports coefficients on *Female*, *High Workload*, and *High Workload* × *Female* from equation (1), with student and classmate characteristics as dependent variables. Panel A reports results for individual student characteristics; Panel B for leave-out means of classmate characteristics. All specifications include math teacher and year fixed effects. Standard errors are clustered at the teacher level. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A5. Teacher-Assigned Grades by Teacher Workload

	(1) Math Grade	(2) Ita. Grade
Female	0.249***	0.466***
	(0.0323)	(0.0297)
High Workload	0.0471	0.0350
	(0.0369)	(0.0408)
High Workload × Female	-0.0360	-0.0364
	(0.0427)	(0.0389)
Mean DV Males	6.921	6.864
R-squared	0.110	0.134
# Students	16600	16604
# Teachers	313	313
Math Teacher FE	Yes	Yes
Year FE	Yes	Yes

Notes: This table reports coefficients on *Female*, *High Workload*, and *High Workload* × *Female* from equation (1), with teacher-assigned grades in grade 8 as dependent variables. All specifications include math teacher and year fixed effects. Standard errors are clustered at the teacher level. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A6. Teachers' Workload and Recommendation for Scientific Track

	DV: Scientific Track Rec.			DV: Math Test Score		
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.0113	0.00692	0.0176	-0.168***	-0.158***	-0.255***
	(0.0126)	(0.0120)	(0.0114)	(0.0400)	(0.0361)	(0.0287)
WorkloadBin(20–24) × Female	-0.0439***	-0.0354**	-0.0337**	-0.00470	-0.0158	-0.0377
	(0.0149)	(0.0140)	(0.0135)	(0.0438)	(0.0400)	(0.0314)
WorkloadBin(25–29) × Female	-0.0852***	-0.0817***	-0.0799***	-0.00913	-0.00906	-0.0157
	(0.0207)	(0.0189)	(0.0180)	(0.0500)	(0.0442)	(0.0339)
WorkloadBin(30+) × Female	-0.0803***	-0.0935***	-0.0833***	-0.0495	-0.0761	-0.0520
	(0.0240)	(0.0297)	(0.0294)	(0.0723)	(0.0649)	(0.0517)
Mean DV Males	0.200	0.200	0.200	0.159	0.159	0.159
# Students	17821	17821	17821	17821	17821	17821
# Teachers	313	313	313	313	313	313
Math Teacher FE	No	Yes	Yes	No	Yes	Yes
Year FE	No	Yes	Yes	No	Yes	Yes
Std. Test Scores	No	No	Yes	No	No	Yes
Stud. Controls	No	Yes	Yes	No	Yes	Yes

Notes: This table shows coefficients from estimation of model 1. One observation is a student assigned to a math teacher in a given year. WorkloadBin categories correspond to four roughly equal groups of the number of eighth-grade students the math teacher must recommend in a given year: fewer than 19, 20–24, 25–29, and more than 30. Coefficients reported are the γ interaction terms (Female × WorkloadBin) from equation 1 for columns 1 and 3, and γ coefficients from the same specification for the math test score in columns 4 and 6. The sample includes student cohorts matched with their math teachers, from Carlana (2019). Student controls include standardized math and Italian test scores, parents occupations, immigrant status, the total number of students assigned to the teacher in grades 6 and 7, the number of years spent with the teacher and average classmates' characteristics. Standard errors are clustered at the teacher level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table A7. Teachers' Workload and Recommendation for Top-Humanities Track

	All students			High-Achievers		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Recommendation for Top-Humanities Track						
Female	0.0146***	0.00508*	0.00707**	0.0370***	0.0149	0.0139
	(0.00241)	(0.00299)	(0.00327)	(0.00690)	(0.00912)	(0.00935)
High Workload	0.00269	-0.00586*	-0.00807*	-0.00241	-0.0178**	-0.0249**
	(0.00349)	(0.00338)	(0.00413)	(0.00820)	(0.00875)	(0.0119)
High Workload × Female		0.0177***	0.0170***		0.0359***	0.0359***
		(0.00459)	(0.00468)		(0.0135)	(0.0135)
Mean DV Males	0.0142	0.0142	0.0142	0.0317	0.0317	0.0317
R-squared	0.0722	0.0731	0.0668	0.134	0.135	0.140
# Students	17821	17821	17821	5018	5018	5018
# Teachers	313	313	313	312	312	312
Math Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Std.test scores	Yes	Yes	Yes	Yes	Yes	Yes
Stud. Controls	No	No	Yes	No	No	Yes
Quality Classmates	No	No	Yes	No	No	Yes

Notes: This table shows coefficients α_2, β, and γ from estimation of model 2. One observation is a student assigned to a math teacher in a given year. The sample includes the students' cohorts matched with their math teachers, from Carlana (2019). High Workload is a dummy equal to 1 if the number of eighth-grade students the math teacher must recommend in a given year exceeds the median (one standard deviation corresponds to around 7 students). Students' controls include students' standardized math and Italian test scores, students' mother education dummies, students' father occupation dummies, immigrant status, the total number of students assigned to the teacher (in 6th and 7th grade as well), the number of years spent with the teacher (1,2,3 years). Quality of classmates controls include average standardized test scores of classmates in math and Italian, the fraction of females, immigrants, and high-socioeconomic status students in the class (excluding the student). Standard errors are clustered at the teacher level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table A8. Measuring teacher's workload as having more than one 8th grade class

	DV: top-scientific track recommendation		
	(1)	(2)	(3)
Female × Workload:#8th graders in other classes	-0.003**	-0.003**	-0.003**
	(0.001)	(0.001)	(0.001)
Female × Workload:#8th graders in same class	-0.011***	-0.009***	-0.009***
	(0.002)	(0.002)	(0.002)
Mean DV Males	0.183	0.183	0.183
R-squared	0.152	0.314	0.315
# Students	17821	17821	17821
# Teachers	313	313	313
Math Teacher FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Std.test scores	No	Yes	Yes
Stud. Controls	No	Yes	Yes
Quality Classmates	No	No	Yes

Notes: This table reports coefficients on the *Female* × *Workload* interaction terms from estimation of equation (2), using alternative measures of teacher workload that separately identify the contribution of teaching multiple eighth-grade classes and of class size. One observation is a student assigned to a math teacher in a given year. *Workload: # 8th graders in other classes* measures the number of eighth-grade students assigned to the teacher in classes other than the student's own class; *Workload: own class size* measures the number of students in the student's own class. Student controls include standardized math and Italian test scores, mother's education, father's occupation, immigrant status, the number of years spent with the same math teacher, and the total number of students assigned to the teacher across all grades. Classmate controls include leave-out means of standardized math and Italian test scores, immigrant status, mother's education, and father's occupation. All specifications include math teacher and year fixed effects. Standard errors are clustered at the teacher level. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A9. Measuring teacher's workload as total number of students in academic year

	DV: top-scientific track recommendation		
	(1)	(2)	(3)
Workload:#6th abd 7th graders × Female	-0.000	-0.000	-0.000
	(0.001)	(0.000)	(0.000)
Workload:#8th graders × Female		-0.005***	-0.004***
		(0.001)	(0.001)
Mean DV Males	0.183	0.183	0.183
R-squared	0.151	0.152	0.315
# Students	17821	17821	17821
# Teachers	313	313	313
Math Teacher FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Std.test scores	No	No	Yes
Stud. controls	No	No	Yes
Quality classmates	No	No	Yes

Notes: This table shows coefficients on the Female × Workload interaction terms from estimation of model 2, using the number of 6th- and 7th-grade students and the number of eighth-grade students separately as workload measures. One observation is a student assigned to a math teacher in a given year. The coefficient on Female × Workload: 6th and 7th graders tests whether total teaching load predicts gender gaps. The coefficient on Female × Workload: 8th graders replicates the main result using a continuous measure of recommendation-specific load. The sample includes student cohorts matched with their math teachers, from Carlana (2019). Student controls include standardized math and Italian test scores, mother's education dummies, father's occupation dummies, immigrant status and the number of years spent with the teacher. Standard errors are clustered at the teacher level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table A10. Teachers' Workload and Recommendation for Science Track, Heterogeneity by Teacher's Implicit Stereotype

	Female Students		Male Students	
	(1)	(2)	(3)	(4)
High Workload	-0.020	-0.014	0.005	-0.006
	(0.014)	(0.018)	(0.013)	(0.015)
High Workload × Math Teacher Stereotype (IAT)	-0.033*	-0.038**	-0.008	-0.011
	(0.017)	(0.016)	(0.013)	(0.012)
Mean DV	0.160	0.160	0.195	0.195
R-squared	0.170	0.304	0.169	0.354
# Students	8025	8025	8641	8641
# Teachers	291	291	291	291
Math Teacher FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Std.test scores	No	Yes	No	Yes
Stud. Controls	No	Yes	No	Yes
Quality Classmates	No	Yes	No	Yes

Notes: This table reports heterogeneity in the effect of high recommendation workload on science-track recommendations by teachers' implicit gender stereotypes. The dependent variable is a dummy equal to one if the student is recommended for the scientific track. One observation is a student assigned to a math teacher in a given year. High Workload is a dummy equal to one if the number of eighth-grade students the teacher must recommend in a given year exceeds the median. Math Teacher Stereotype is the teacher's score on an Implicit Association Test linking boys to science and girls to humanities (higher values indicate stronger implicit math–male associations). Columns 1 and 2 restrict to female students and columns 3 and 4 to male students. The sample includes student cohorts matched with their math teachers, from Carlana (2019). Student controls include standardized math and Italian test scores, mother's education dummies, father's occupation dummies, immigrant status and the number of years spent with the teacher. Standard errors are clustered at the teacher level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table A11. Teachers' Workload and Recommendation for Science Track, Heterogeneity by Italian Ability

	High Achievers only in Math (extreme signals)		High Achievers Math and Italian (mixed signals)	
	(1)	(2)	(3)	(4)
Female	-0.014	-0.020	-0.079***	-0.046*
	(0.037)	(0.035)	(0.027)	(0.027)
High Workload	0.066*	0.051	-0.025	0.013
	(0.035)	(0.045)	(0.036)	(0.046)
High Workload × Female	-0.017	-0.011	-0.091***	-0.100***
	(0.050)	(0.048)	(0.034)	(0.033)
Mean DV Males	0.315	0.315	0.572	0.572
R-squared	0.277	0.350	0.293	0.334
# Students	1714	1714	3263	3263
# Teachers	273	273	302	302
Math Teacher FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Std.test scores	No	Yes	No	Yes
Stud. Controls	No	Yes	No	Yes
Quality Classmates	No	Yes	No	Yes

Notes: This table reports heterogeneity in the effect of high recommendation workload on science-track recommendations, separately for students who are high achievers only in math (extreme signals: high math score, lower Italian score) and students who are high achievers in both math and Italian (mixed signals: high performance in both subjects). High Workload is a dummy equal to one if the number of eighth-grade students the math teacher must recommend in a given year exceeds the median. One observation is a student assigned to a math teacher in a given year. The sample includes student cohorts matched with their math teachers, from Carlana (2019). Student controls include standardized math and Italian test scores, mother's education dummies, father's occupation dummies, immigrant statusn and the number of years spent with the teacher. Standard errors are clustered at the teacher level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

A.2 Sensitivity and Robustness

Table A12. Teacher workload and gender gaps: continuous and quintile specification

	(1)	(2)	(3)	(4)	(5)	(6)
DV: Recommendation for Science Track						
Female	-0.0226***	-0.0219***	-0.0231***	0.0132	0.0141	0.0126
	(0.00561)	(0.00564)	(0.00587)	(0.00947)	(0.00956)	(0.00970)
Workload (1SD)	0.0107*	0.0106*	0.00743			
	(0.00615)	(0.00627)	(0.00635)			
Female × Workload (1SD)	-0.0252***	-0.0244***	-0.0244***			
	(0.00733)	(0.00736)	(0.00733)			
Workload Q2				0.0222	0.0248	0.0212
				(0.0136)	(0.0155)	(0.0153)
Workload Q3				0.0305*	0.0335	0.0283
				(0.0173)	(0.0216)	(0.0205)
Workload Q4				0.0191	0.0203	0.0160
				(0.0166)	(0.0238)	(0.0221)
Workload Q5				0.0379**	0.0365	0.0278
				(0.0176)	(0.0243)	(0.0233)
Female × Workload Q2				-0.0292*	-0.0294*	-0.0290*
				(0.0150)	(0.0150)	(0.0149)
Female × Workload Q3				-0.0309*	-0.0325*	-0.0316*
				(0.0170)	(0.0171)	(0.0169)
Female × Workload Q4				-0.0593***	-0.0595***	-0.0592***
				(0.0156)	(0.0156)	(0.0156)
Female × Workload Q5				-0.0753***	-0.0740***	-0.0740***
				(0.0179)	(0.0180)	(0.0179)
Mean DV (Males)	0.200	0.200	0.200	0.200	0.200	0.200
R-squared	0.308	0.314	0.315	0.308	0.315	0.316
# Students	17821	17821	17821	17821	17821	17821
# Teachers	313	313	313	313	313	313
Math Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Std. test scores	Yes	Yes	Yes	Yes	Yes	Yes
Student controls	No	Yes	Yes	No	Yes	Yes
Classmate quality	No	No	Yes	No	No	Yes

Notes: This table reports coefficients from two alternative specifications of model 1. Columns 1 to 3 use a continuous workload measure (standardized). Columns 4 to 6 use quintile-based workload bins (Q1 is the reference category), and report the Female × Workload *Qk* interaction terms. One observation is a student assigned to a math teacher in a given year. Student controls include standardized math and Italian test scores, mother's education dummies, father's occupation dummies, immigrant status, the total number of students assigned to the teacher in grades 6 and 7, and the number of years spent with the teacher (1, 2, or 3) and classmates controls. All specifications include math teacher and year fixed effects. Standard errors are clustered at the teacher level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table A13. Counterfactual Decision Rules for Teachers

	True Recommendation	Very Good in Math	Top 30% students in class	Top X% students in class (Teacher specific fraction)
	(1)	(2)	(3)	(4)
High Workload	0.0172	-0.00816	-0.0114*	0.0115
	(0.0145)	(0.0112)	(0.00607)	(0.00734)
Female	-0.000914	-0.00306	-0.0108	-0.00372
	(0.00842)	(0.00675)	(0.00691)	(0.00675)
High Workload × Female	-0.0458***	0.00198	0.0162*	-0.0124
	(0.0121)	(0.00953)	(0.00934)	(0.00978)
R-squared	0.315	0.527	0.440	0.522
N. Obs	16231	16231	16231	16231
Teacher FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Std. test scores	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Mean DV				
Sd Dependent Variable	0.197	0.212	0.177	0.228
ysd	0.398	0.408	0.382	0.419

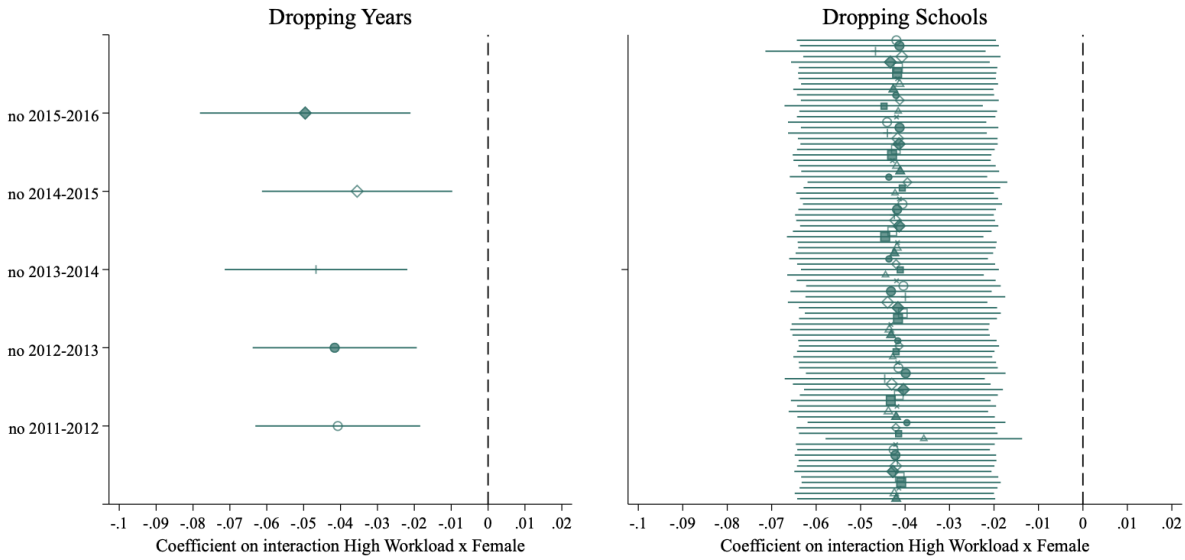
Notes: Column 1 replicates the main specification with the true science-track recommendation as outcome. Columns 2–4 replace the dependent variable with counterfactual decision rules based on students' standardized math test scores (INVALSI): whether the student scores in the top 20 percent nationally in math standardized test scores (column 2), whether the student ranks in the top 20 percent of their class (column 3), and whether the student ranks in the top X percent of their class, where X is the leave-out mean share of students recommended by the same teacher in other years (column 4). All specifications include teacher and year fixed effects, standardized test scores, and student controls. Standard errors are clustered at the teacher level. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

Table A14. Teacher workload and gender gaps: controlling for class percentile rank

	DV: scientific track recommendation		
	(1)	(2)	(3)
High Workload	0.019	0.021*	0.019
	(0.012)	(0.012)	(0.015)
Female	-0.008	-0.046***	-0.038***
	(0.008)	(0.008)	(0.009)
High workload × Female	-0.048***	-0.051***	-0.051***
	(0.012)	(0.012)	(0.012)
Class percentile INVALSI Math (100=top)	0.004***	0.001***	-0.000
	(0.000)	(0.000)	(0.000)
Class percentile INVALSI Italian (100=top)	0.002***	0.000***	0.001***
	(0.000)	(0.000)	(0.000)
Class percentile Grade Math (100=top)		0.004***	0.001
		(0.000)	(0.000)
Class percentile Grade Italian (100=top)		0.001***	0.003***
		(0.000)	(0.000)
Mean DV Males	0.197	0.194	0.194
R-squared	0.301	0.359	0.375
# Students	16213	15383	15383
# Teachers	313	313	313
Math Teacher FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Std.test scores	No	No	Yes
Grades controls	Yes	No	Yes
Student controls	Yes	No	Yes
Classmate quality	Yes	No	Yes

Notes: In this table, High Workload × Female captures the additional gender gap in high-workload years while directly controlling for students' within-class percentile rank in both INVALSI standardized tests and teacher-assigned grades. Class percentile controls are constructed as students' percentile rank (0–100, where 100 is the top) within their class in the indicated score. One observation is a student assigned to a math teacher in a given year. All specifications include math teacher and year fixed effects. Standard errors are clustered at the teacher level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Figure A2. Dropping years and schools



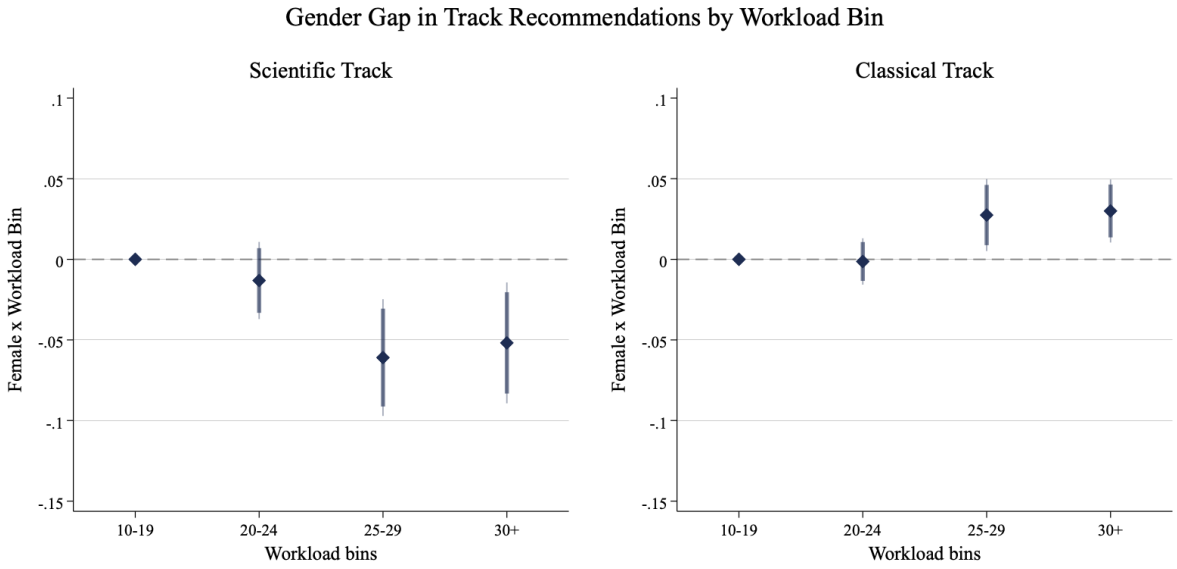
Notes: This figure shows the robustness of the main estimate (the High Workload × Female coefficient) to sequentially dropping each academic year and each school from the sample. Each point represents a separate regression estimate obtained by excluding one year or one school. Caps show 95 percent confidence intervals, clustered at the teacher level.

Table A15. Teacher Workload and Gender Gap in Track Recommendations (*Alternative Sample: survey teachers and their past students*)

	Science	Classical	Med. Humanities	Technical	Vocational
	(1)	(2)	(3)	(4)	(5)
High Workload × Female	-0.0355***	0.0209***	-0.00211	0.00114	0.0179
	(0.0122)	(0.00617)	(0.0119)	(0.0130)	(0.0120)
Mean DV Males	0.193	0.0221	0.0915	0.339	0.345
Observations	23387	23387	23387	23387	23387
# Teachers	285	285	285	285	285
Teacher FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports coefficients on Female, High Workload, and High Workload × Female from equation (2), estimated separately for each track type as the dependent variable. High Workload is a dummy equal to one if the number of 8th-grade students the math teacher must recommend in a given year exceeds the median. One observation is a student assigned to a math teacher in a given year. All specifications include teacher and year fixed effects. Student controls include standardized math and Italian test scores, mother's education, father's occupation, immigrant status, the total number of students assigned to the teacher in grades 6 and 7, and the number of years spent with the teacher. Standard errors are clustered at the teacher level. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Figure A3. Gender Gap in Track Recommendations by Teacher Workload (*Alternative Sample: survey teachers and their past students*)

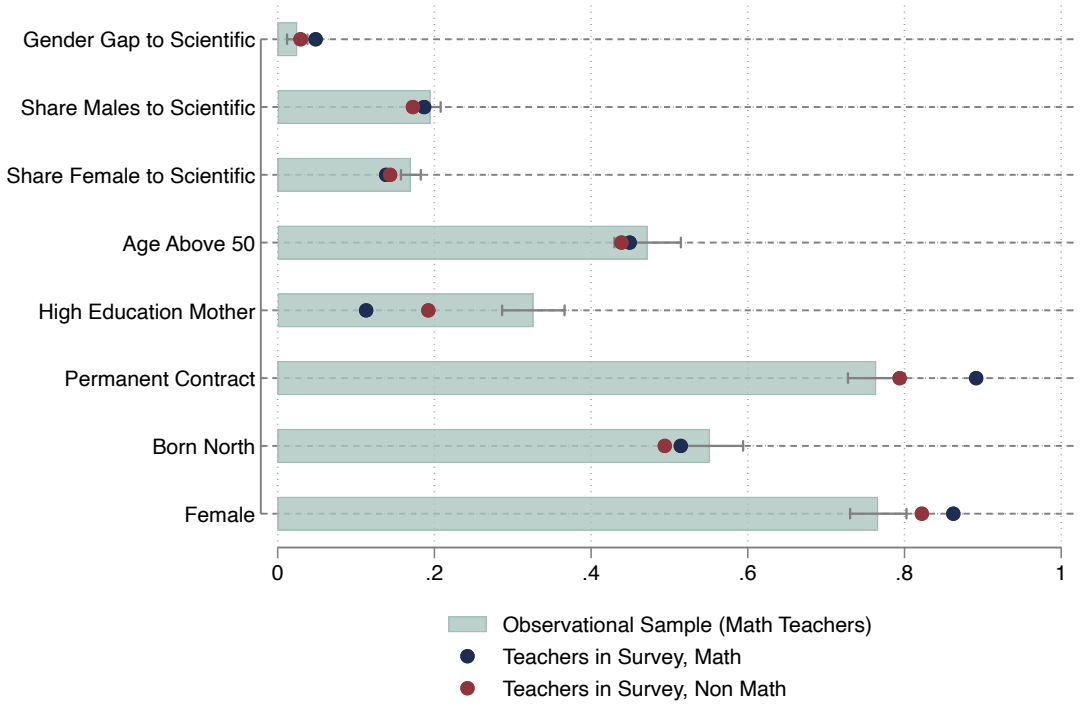


Notes: Each panel shows coefficients $\hat{\gamma}_s$ on Female $\times$ Workload Bin from equation (1), estimated separately for each track type as the dependent variable. The reference bin is the lowest workload group (fewer than 19 students to recommend). Coefficients capture the additional gender gap in each workload bin relative to the lowest bin. Student controls, teacher fixed effects, and year fixed effects are included in all specifications. The sample excludes the 2020–21 cohort due to COVID-19 disruptions. Caps show 95 percent and 90 percent confidence intervals, clustered at the teacher level.

B Observational Evidence (Section 4), Additional Figures and Tables

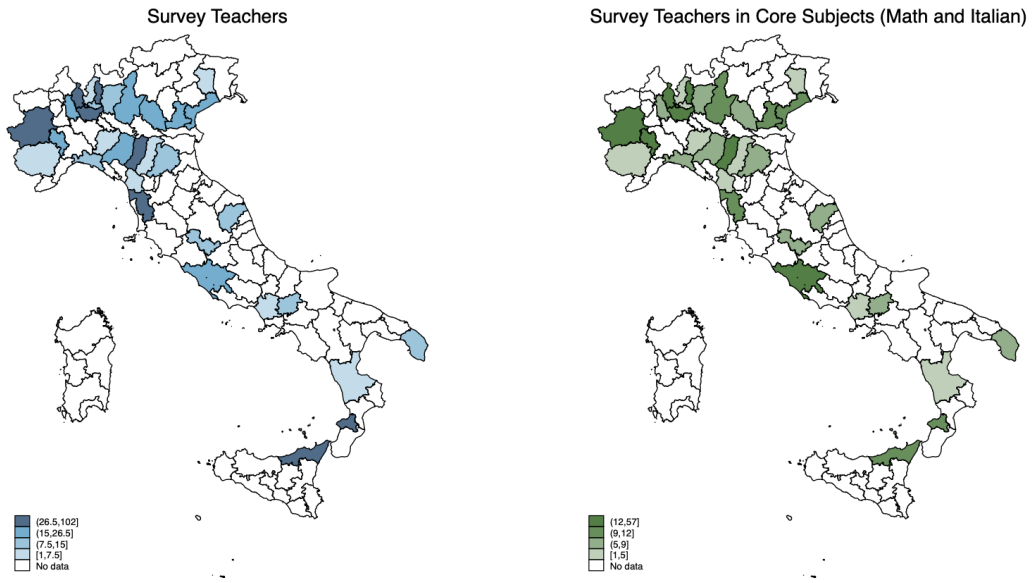
B.1 Additional Tables and Figures

Figure B1. Teachers in Survey and Observational Sample



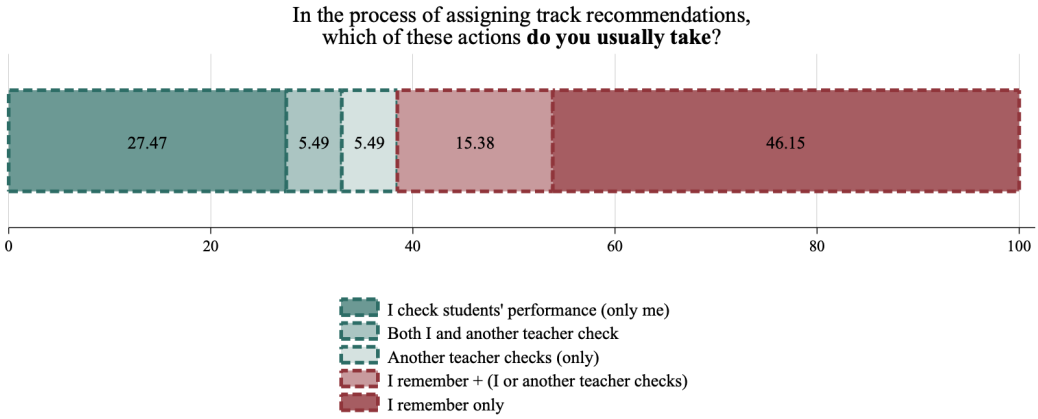
Notes: This figure compares the characteristics of teachers in the main observational sample with math teachers and non-math teachers who participated in the survey. Each row reports the share of teachers with a given characteristic: female, born in the North of Italy, permanent contract, high maternal education, age above 50, share of female students recommended to the science track, share of male students recommended to the science track, and the gender gap in science-track recommendations. The figure assesses the representativeness of the survey sample relative to the broader population of middle school math teachers.

Figure B2. Survey Teachers (all teachers and teachers in math and Italian)



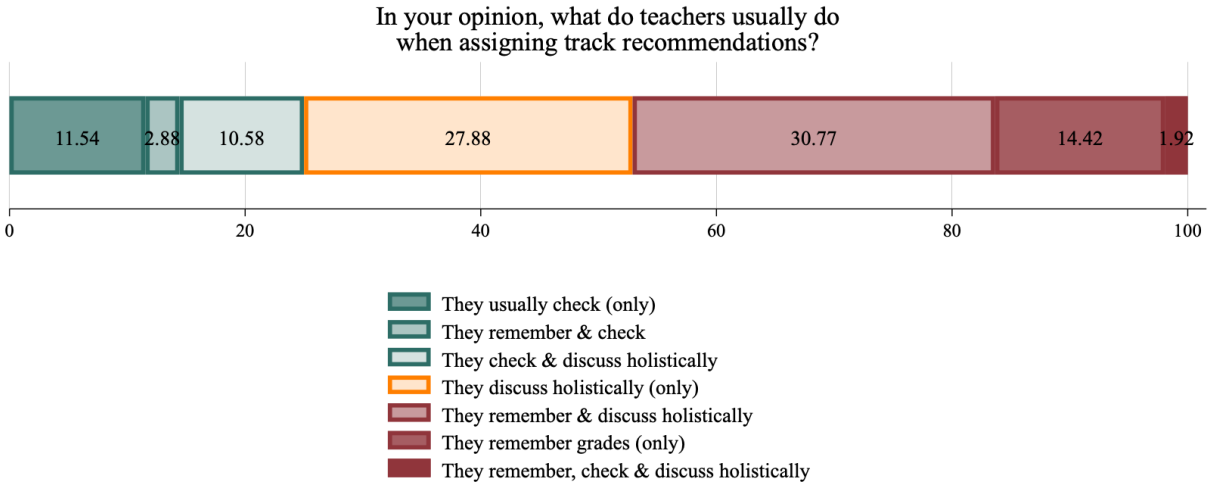
Notes: This figure shows the geographical distribution of schools attended by teachers who participated in the survey. The left panel includes all survey teachers; the right panel restricts to teachers of core subjects (math and Italian).

Figure B3. Teacher Survey: Actions taken when assigning track recommendations



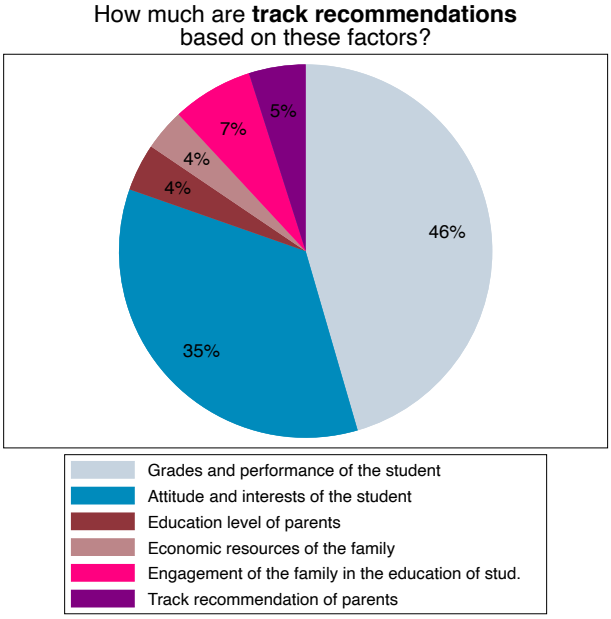
Notes: The graph shows teachers' answers to the following survey question: "In the process of assigning track recommendations, which of these actions do you usually take? (A) I check my students' performance in the class register, (B) Another teacher checks students' performance in the class register, (C) I remember my students' performance without checking it in the class register". The sample includes survey teachers in core subjects (math and Italian).

Figure B4. Survey to teachers: what do you think other teachers usually do?



Notes: The graph shows teachers' answers to the following survey question: "In the process of assigning track recommendations: (A) usually, teachers remember the academic performance of their students without checking the registry in the teaching meeting, (B) usually, teachers keep the registry open and check students' grades, (C) teachers discuss during the teaching meeting without explicitly checking the registry, but thinking more broadly about students' attitudes, and interests." The sample includes all survey teachers.

Figure B5. Survey Question: Reported Determinants of Track Recommendations



Notes: This figure reports the distribution of teachers' responses to the survey question: "How important are the following factors when you assign track recommendations?" Teachers rated each factor on a scale from 1 to 100.

Table B1. Correlates of Relying on Memory versus Consulting Records

	DV: Relies on memory (=1) vs. consults records (=0)										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Female	0.128										0.0606
	(0.124)										(0.135)
Age		0.00767*									0.00766
		(0.00438)									(0.00469)
Born in Northern Italy			-0.173**								-0.174**
			(0.0711)								(0.0744)
Permanent contract				-0.0101							-0.151
				(0.176)							(0.202)
IAT					0.0337						0.0383
					(0.0547)						(0.0520)
Importance for track rec.: student ability						-0.00128					0.00344
						(0.00240)					(0.00367)
Importance for track rec.: attitudes and interests							0.00452				0.00650*
							(0.00304)				(0.00384)
Importance for track rec.: parents SES (education and wealth)								-0.00277			0.000630
								(0.00407)			(0.00571)
Importance for track rec.: parental engagement									-0.00193		0.00427
									(0.00493)		(0.00611)
Importance for track rec.: parents' recommendation										-0.00240	0.00477
										(0.00574)	(0.00797)
Mean DV	0.615	0.615	0.615	0.615	0.615	0.615	0.615	0.615	0.615	0.615	0.615
N. Teachers	182	182	182	182	182	182	182	182	182	182	182
R^2	0.00617	0.0197	0.0313	0.0000180	0.00219	0.00719	0.0191	0.00826	0.00646	0.00658	0.0750

Notes: This table reports OLS estimates where the dependent variable is a dummy equal to one if the teacher reports relying on memory rather than consulting student records when assigning track recommendations. The sample includes core-subject teachers (math and Italian) who report either relying on memory or consulting records. Columns (1)–(10) report univariate regressions of the memory dummy on each teacher characteristic separately; column (11) includes all characteristics simultaneously. Missing values for each covariate are imputed to zero and a missing indicator is included in all specifications. Robust standard errors in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table B2. Student Characteristics by Teacher Recall Behavior

	(1) Math Score	(2) Ita. Score	(3) Math Grade	(4) Ita. Grade	(5) Immigrant	(6) High Edu. Mother	(7) High Occ. Father
Female	-5.653***	7.529***	0.217***	0.486***	-0.0117	0.0230**	-0.00176
	(1.218)	(1.293)	(0.0350)	(0.0348)	(0.00827)	(0.0111)	(0.0140)
Memory	0.177	1.809	-0.0567	-0.0392	-0.00320	0.0126	0.0149
	(1.361)	(1.254)	(0.0556)	(0.0507)	(0.00873)	(0.0127)	(0.0128)
Female × Memory	0.0543	-1.297	0.00562	-0.0308	0.00352	-0.00322	-0.0183
	(1.544)	(1.576)	(0.0425)	(0.0416)	(0.00991)	(0.0143)	(0.0176)
R-squared	0.0804	0.0716	0.0534	0.114	0.0563	0.0965	0.0650
# Students	10764	10602	15496	15498	15793	12434	12377
# Teachers	180	180	182	182	182	179	180
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports coefficients from regressing student characteristics on *Female*, *Memory*, and *Female* × *Memory*, including school and cohort fixed effects. *Memory* is a dummy equal to one if the teacher reports relying on recollection rather than consulting the class register when assigning track recommendations. The dependent variables are: standardized math and Italian test scores measured in fifth grade; teacher-assigned math and Italian grades in seventh grade; and indicators for immigrant status, mother having a high level of education (college or higher), and father having a high-skilled occupation. The coefficient of interest is *Female* × *Memory*. The sample includes students assigned to core-subject teachers who report either relying on memory or consulting records. Standard errors are clustered at the teacher level. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table B3. Gender gap in scientific track recommendations for teachers who report checking students' performance vs. relying on memory. Sample restricted to math teachers only.

	DV: Scientific Track Recommendation				
	(1)	(2)	(3)	(4)	(5)
Female	-0.049***	-0.016	-0.014	-0.014	-0.013
	(0.013)	(0.017)	(0.017)	(0.017)	(0.017)
Memory		0.057***	0.043**	0.046***	0.023
		(0.019)	(0.018)	(0.017)	(0.019)
Memory × Female		-0.055**	-0.057**	-0.058**	-0.059**
		(0.023)	(0.024)	(0.024)	(0.024)
Mean control	0.200	0.200	0.200	0.200	0.200
Year FE	Yes	Yes	Yes	Yes	Yes
Std. Test Scores & Stud. Controls	Yes	Yes	Yes	Yes	Yes
School FE	No	No	Yes	Yes	Yes
Teacher Controls	No	No	No	Yes	Yes
Controls for Teachers' Attitudes	No	No	No	No	Yes
R^2	0.138	0.141	0.158	0.159	0.162
Observations	7364	7364	7364	7364	7364

Notes: This figure reports coefficients β_1, β_2, β_3 from the estimation of model 3. Memory is a dummy equal to 1 if a teacher reports that when assigning track recommendations, she usually remembers her students' performance without checking it in the class register. Teacher controls include subject studied group, subject taught group, age, gender, education father, teacher's contract group, school north, born in the north. Student-level controls include math and Italian standardized test scores in grade 5, a dummy equal to 1 if the student is an immigrant, and a dummy equal to 1 if the student's mother has completed college. The sample includes 8th-grade students matched with their math teachers. Standard errors are clustered at the teacher level. *p<0.1; **p<0.05, ***p<0.01.

Table B4. Gender gap in scientific-track recommendations: alternative specification separating teachers who rely only on memory from those who rely on memory while someone checks records

	DV: Scientific Track Recommendation			
	(1)	(2)	(3)	(4)
Female	-0.020	-0.020	-0.020	-0.020
	(0.012)	(0.012)	(0.012)	(0.013)
Memory only	0.044***	0.032**	0.027*	0.024*
	(0.015)	(0.014)	(0.014)	(0.013)
Memory + check	0.027	0.012	0.007	0.005
	(0.018)	(0.014)	(0.015)	(0.015)
Memory only × Female	-0.043**	-0.042**	-0.042**	-0.041**
	(0.018)	(0.018)	(0.018)	(0.018)
Memory + check × Female	-0.029	-0.028	-0.028	-0.029
	(0.021)	(0.021)	(0.021)	(0.021)
Mean control	0.187	0.187	0.187	0.187
Observations	15080	15080	15080	14853
R^2	0.133	0.150	0.151	0.152
Year FE	Yes	Yes	Yes	Yes
Std. Test Scores & Stud. Controls	Yes	Yes	Yes	Yes
School FE	No	Yes	Yes	Yes
Teacher Controls	No	No	Yes	Yes
Controls for Teachers' Attitudes	No	No	No	Yes

Notes: This table reports coefficients β_1, β_2, and β_3 from the estimation of model 3. *Memory only* is a dummy equal to one if a teacher reports relying on memory when assigning track recommendations and reports that neither she nor another teacher checks student records. *Memory + check* is a dummy equal to one if a teacher reports relying on memory and also reports that either she or another teacher checks student records. Teacher controls include subject studied, subject taught, age, gender, father's education, contract type, an indicator for teaching in the North, and an indicator for being born in the North. Student-level controls include standardized math and Italian test scores, an indicator for immigrant status, and an indicator for whether the student's mother completed college. The sample includes 8th-grade students matched with their core-subject teachers (math and Italian). Standard errors are clustered at the teacher level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B5. Gender gap in scientific track recommendations for teachers who report checking students' performance vs. relying on memory. Interacting Controls by Female

	DV: Scientific Track Recommendation			
	(1)	(2)	(3)	(4)
Memory	0.019	0.016	0.015	0.014
	(0.012)	(0.011)	(0.012)	(0.012)
Memory × Female	-0.038**	-0.033**	-0.030**	-0.027*
	(0.016)	(0.015)	(0.014)	(0.014)
Mean control	0.187	0.187	0.187	0.187
Observations	15080	15080	15080	15080
R^2	0.151	0.157	0.159	0.160
Year FE	Yes	Yes	Yes	Yes
Std. Test Scores & Stud. Controls	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes
Teacher Controls	Yes	Yes	Yes	Yes
Controls for Teachers' Attitudes	Yes	Yes	Yes	Yes
Student Controls x Female	No	Yes	Yes	Yes
Teacher Controls x Female	No	No	Yes	Yes
Controls for Teachers' Attitudes x Female	No	No	No	Yes

Notes: This figure reports coefficients β_1, β_2, β_3 from the estimation of model 3. Memory is a dummy equal to 1 if a teacher reports that when assigning track recommendations, she usually remembers her students' performance without checking it in the class register. Teacher controls include subject studied group, subject taught group, age, gender, education father, teacher's contract group, school north, born in the north. Student-level controls include math and Italian standardized test scores, a dummy equal to 1 if the student is an immigrant, and a dummy equal to 1 if the student's mother has completed college. The sample includes 8th-grade students matched with their teachers in core subjects (math and Italian). Standard errors are clustered at the teacher level. *p<0.1; **p<0.05, ***p<0.01.

C Teacher Experiment (Section 5)

C.1 Experimental Design and Procedures

Table C1. Pre-registration and Study Materials

Registry/Material	Link
AEA RCT Registry. RCT ID: AEARCTR-0010756	https://www.socialscienceregistry.org/trials/10756
Teacher Survey and Experiment Protocol	https://www.dropbox.com/scl/fi/dfd1q4mij1h49u1kgul23/Teacher_experimental_protocol.pdf?rlkey=nfyewctbr9sz7cvy9749udfii&dl=0

C.1.1 Hypothetical Student Profiles

Student: [Roberto/Roberta] (excellent in both math and humanities)
Roberto/Roberta is among the best students in his/her class in both humanities and science subjects. Last semester, he/she got a 9 in Italian and an 8 in math, and his/her GPA is 8.5. Roberto/Roberta was selected to participate in a math competition at the regional level, and he/she reached the final rounds.

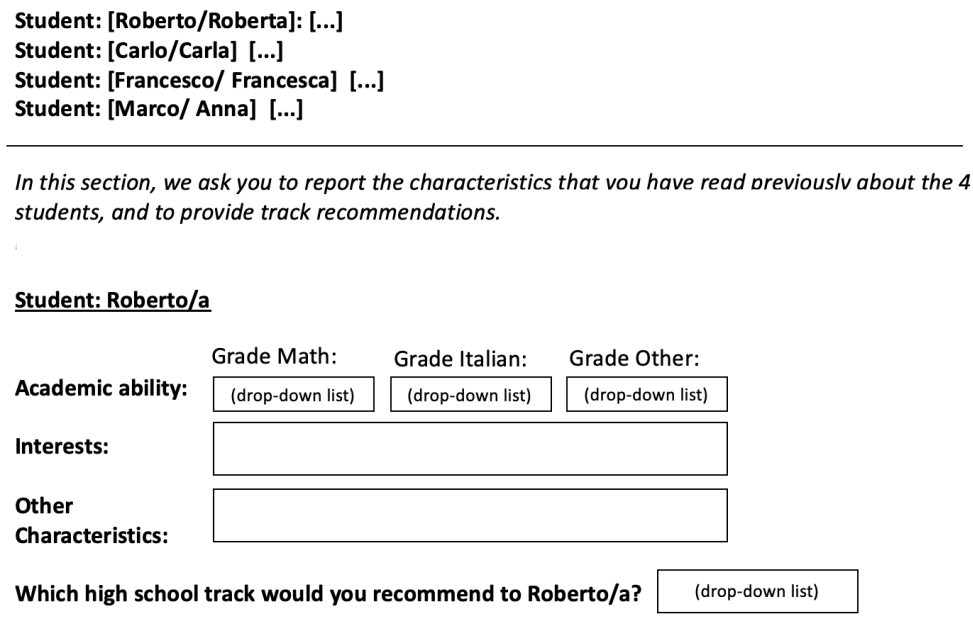
Student: [Carlo/Carla] (good in humanities and poor in math)
Carlo/Carla comes from a disadvantaged family background. His/Her father left when he/she was 5, his/her mother had some health issues and he/she mainly lives with his/her grandparents. However, his/her grandparents support him/her a great deal in his/her education, and he/she manages to do quite well at school. He/She got a 6 in math and an 8 in Italian, and he/she got a GPA of 8. He/She loves reading fiction and poetry. He/She is very creative in his/her essays although he/she often makes grammar mistakes. He/She also participated in a poetry competition, where he/she received an award for his/her poem called "My teenage years as a digital native".

Student: [Francesco/Francesca] (average student, good in languages)
Francesco/a is an above-average student in his/her class, though not exceptional. He/She received a 7 in math and Italian, and has a GPA of 8. He/She is also very passionate about languages, and received an 8 in English. He/She spent 3 weeks in Ireland during the summer, where he/she improved his/her spoken English. He/She loves outdoor activities and spending his/her free time with his/her group of friends. His/Her parents are both high school teachers.

Student: [Marco/Anna] (good in math and technical subjects and poor in humanities)
Marco/Anna is a very extroverted and social boy/girl. He/She is not very diligent at

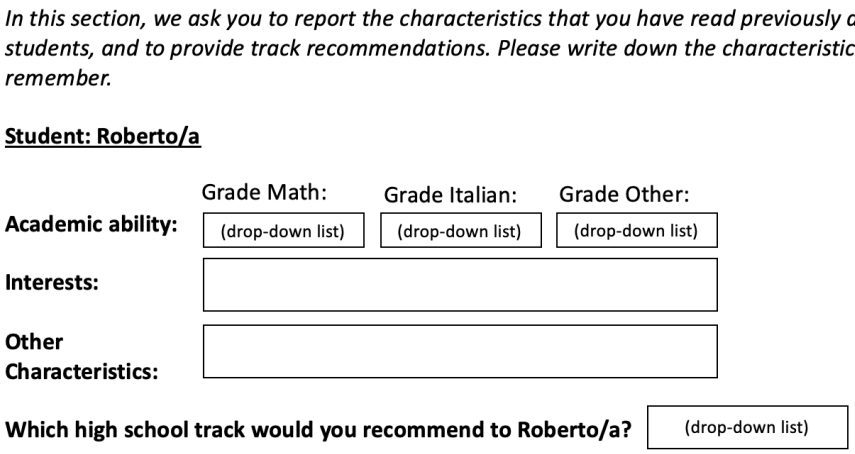
school and often forgets to do homework. He/She is frequently reprimanded by teachers in class for talking with his/her friends and disrupting lessons. He/She is very intuitive and naturally talented in math and technical subjects, in which he/she received an 8, while receiving a 6 in Italian. In his/her free time, he/she loves playing computer games with his/her friends and helping his/her older brother fix bikes in the garage at home.

Figure C1. Teachers' Survey Experiment: *Baseline* condition



Notes: This figure shows the decision screen presented to teachers in the *Baseline* condition of the teacher experiment. After observing all four student profiles, teachers in the *Baseline* condition had access to the profiles and were asked to recall each student's academic ability, interests, and other characteristics using open-ended free-text fields, and to provide a track recommendation via a drop-down list.

Figure C2. Teachers' Survey Experiment: *Memory* Treatment



Notes: This figure shows the decision screen presented to teachers in the *Memory* condition of the teacher experiment. The screen is identical in structure to the *Baseline* condition, but teachers could not review student profiles and had to retrieve characteristics from memory. Teachers reported recalled academic ability (math, Italian, and other grades via drop-down list), interests, and other characteristics in open-ended fields, and provided a track recommendation via a drop-down list.

Table C2. Student profile characteristics and stereotype-typing

Student	Characteristic	Type
Carlo/a — *Extreme signals (humanities-oriented)*		
Carlo/a	Loves reading fiction	Female-typed
Carlo/a	Poetry	Female-typed
Carlo/a	Sensitive/emotional	Female-typed
Carlo/a	Creative in essays	Female-typed
Carlo/a	Grammar mistakes	Male-typed
Carlo/a	Difficult family situation	Neutral
Carlo/a	Grandparents help him/her	Neutral
Carlo/a	Math grade: 6	Grades
Carlo/a	Italian grade: 8	Grades
Carlo/a	GPA around 8 across subjects	Grades

Student	Characteristic	Typed as
Roberto/a — *Mixed signals*		
Roberto/a	Strong in humanities subjects	Female-typed
Roberto/a	Math competition/finalist	Male-typed
Roberto/a	Strong in scientific subjects	Male-typed
Roberto/a	Top student of his/her class	Neutral
Roberto/a	Italian grade: 9	Grades
Roberto/a	Math grade: 8	Grades
Roberto/a	GPA 8.5	Grades
Francesco/a — *Mixed signals*		
Francesco/a	Interested in languages / English	Female-typed
Francesco/a	Outdoor activities / friends / sports	Male-typed
Francesco/a	Above average but not excellent	Neutral
Francesco/a	Spent time in Ireland / study trip	Neutral
Francesco/a	Parents are teachers	Neutral
Francesco/a	Math grade: 7	Grades
Francesco/a	Italian grade: 7	Grades
Francesco/a	GPA around 8	Grades
Francesco/a	English grade 8	Grades
Marco/Anna — *Extreme signals (technology-oriented)*		
Marco/Anna	Not diligent / forgets homework	Male-typed
Marco/Anna	Intuitive in math/tech/logic	Male-typed
Marco/Anna	Plays computer / videogames	Male-typed
Marco/Anna	Repairs bicycles in her/his free time	Male-typed
Marco/Anna	Disturbs in class / distractible / talkative	Neutral
Marco/Anna	Extrovert / social	Neutral
Marco/Anna	Technology grade: 8	Grades
Marco/Anna	Math grade: 8	Grades
Marco/Anna	Italian grade: 6	Grades

Notes: This table lists the characteristics included in each experimental student profile and their stereotype-typing used in the analysis. Pink: female-typed characteristics. Blue: male-typed characteristics. Neutral characteristics are shown in black. Profiles with "extreme signals" have characteristics more distinctive of one gender; profiles with "mixed signals" have some male-typed and some female-typed characteristics.

Table C3. Track recommendations by gender

	Science	Technical	Vocational	Classical	Med. Humanities
	(1)	(2)	(3)	(4)	(5)
Female	-0.051***	-0.167***	-0.106***	0.038***	0.281***
	(0.006)	(0.006)	(0.007)	(0.003)	(0.006)
Mean DV Males	0.193	0.343	0.343	0.023	0.090
Observations	18258	18258	18258	18258	18258

Notes: This table reports the coefficient on *Female* from separate OLS regressions of track recommendation dummies on student gender. The dependent variable in each column is a dummy equal to one if the student was recommended for the indicated track. The sample includes eighth-grade students matched with survey teachers who participated in the survey between 2016 and 2022. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

C.2 Attrition

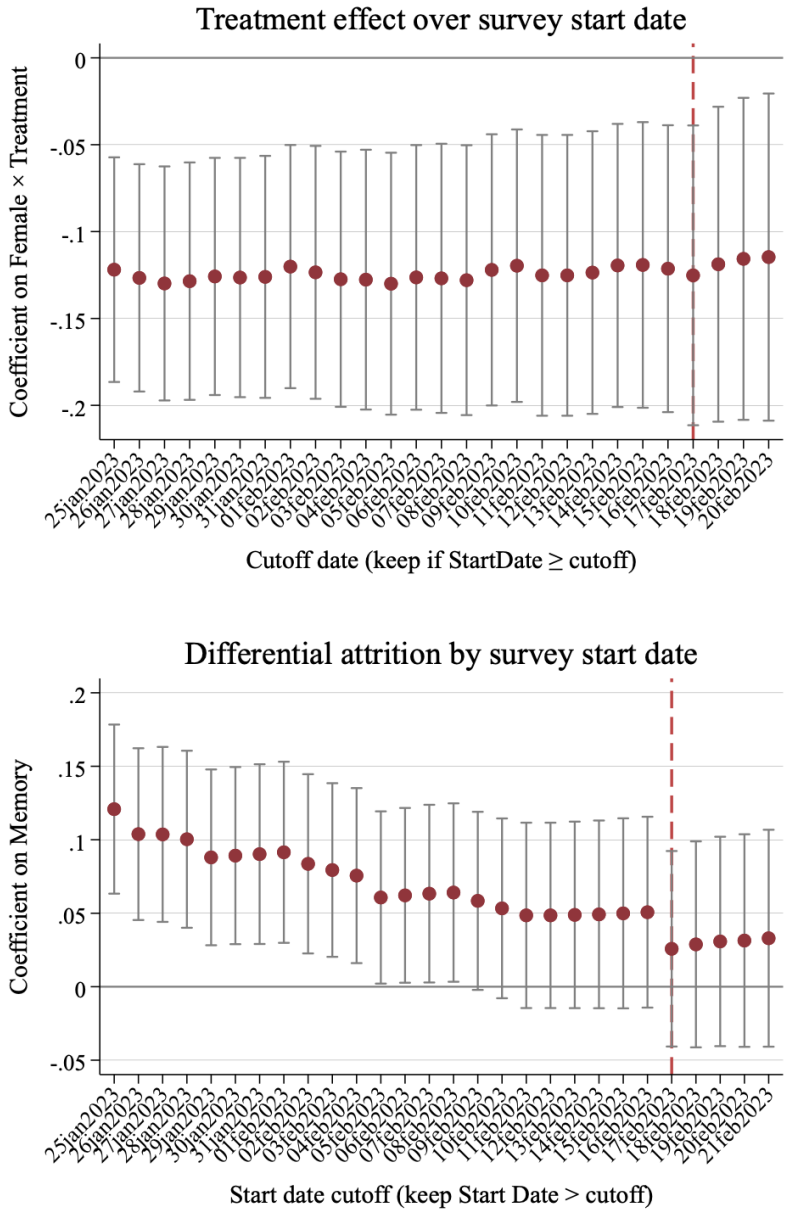
In this section, I examine differential attrition between the *Memory* and *Baseline* conditions and show that it does not bias the main results.

The survey was administered in two phases. In the first phase, the survey did not include a force-response option, which led to differential attrition: teachers in the *Memory* condition were more likely to drop out before providing track recommendations, plausibly because the recall task was more demanding for them. On February 17, a force-response option was introduced, eliminating differential attrition.

Panel (b) of Figure C3 shows how differential attrition in the *Memory* condition evolves with the survey start date. For each date on the horizontal axis, the figure reports the rate of differential attrition restricting the sample to respondents who started the survey on or after that date. As expected, differential attrition decreases monotonically as the sample is restricted to later start dates. Among respondents starting after February 10, differential attrition is no longer statistically detectable at the 5 percent level; among those starting after February 17, it drops close to zero.

Panel (a) of Figure C3 shows the main treatment effect across survey start dates. The estimate is stable throughout and is virtually identical in the period when differential attrition is absent, indicating that attrition does not bias the results. This is further confirmed by the balance table (Table **3**), which shows that treated and control respondents are similar along a broad set of observable characteristics.

Figure C3. Treatment Effects and Attrition over Survey Start Date



Notes: Panel (a) plots the estimated Memory × Female coefficient from equation (4) over time, using each teacher's survey start date as the running variable. Panel (b) plots the coefficient from a regression of an attrition indicator—equal to one if the teacher did not complete the recommendation task—on the Memory treatment indicator, estimated on the subsample of teachers who started the survey after each date on the horizontal axis. The vertical line marks February 17, when the forced-response feature was introduced and differential attrition disappeared.

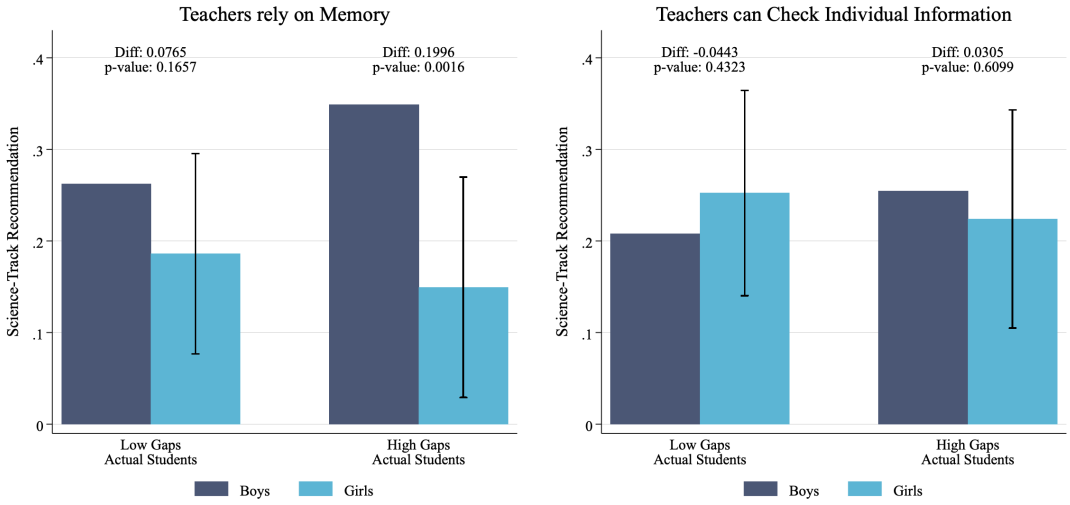
C.3 Additional Tables and Figures

Table C4. Teacher Experiment: Biased recall of students' characteristics

	Recall Female-Typed Characteristics			Recall Male-Typed Characteristics		
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.0156	-0.0166	-0.0189	-0.0134	-0.0130	-0.0171
	(0.0332)	(0.0239)	(0.0235)	(0.0339)	(0.0290)	(0.0283)
Memory	-0.265***	-0.263***	-0.261***	-0.307***	-0.311***	-0.309***
	(0.0292)	(0.0268)	(0.0263)	(0.0327)	(0.0306)	(0.0299)
Memory × Female	0.106**	0.102***	0.0985***	-0.0366	-0.0269	-0.0303
	(0.0449)	(0.0372)	(0.0367)	(0.0447)	(0.0391)	(0.0386)
Mean control (Baseline-Male)	0.504	0.504	0.504	0.560	0.560	0.560
Memory × Female as %	21.107%	20.216%	19.529%	-6.526%	-4.804%	-5.405%
R^2	0.0511	0.357	0.361	0.109	0.278	0.290
Observations	1761	1761	1761	1761	1761	1761
N. teachers	448	448	448	448	448	448
Student FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

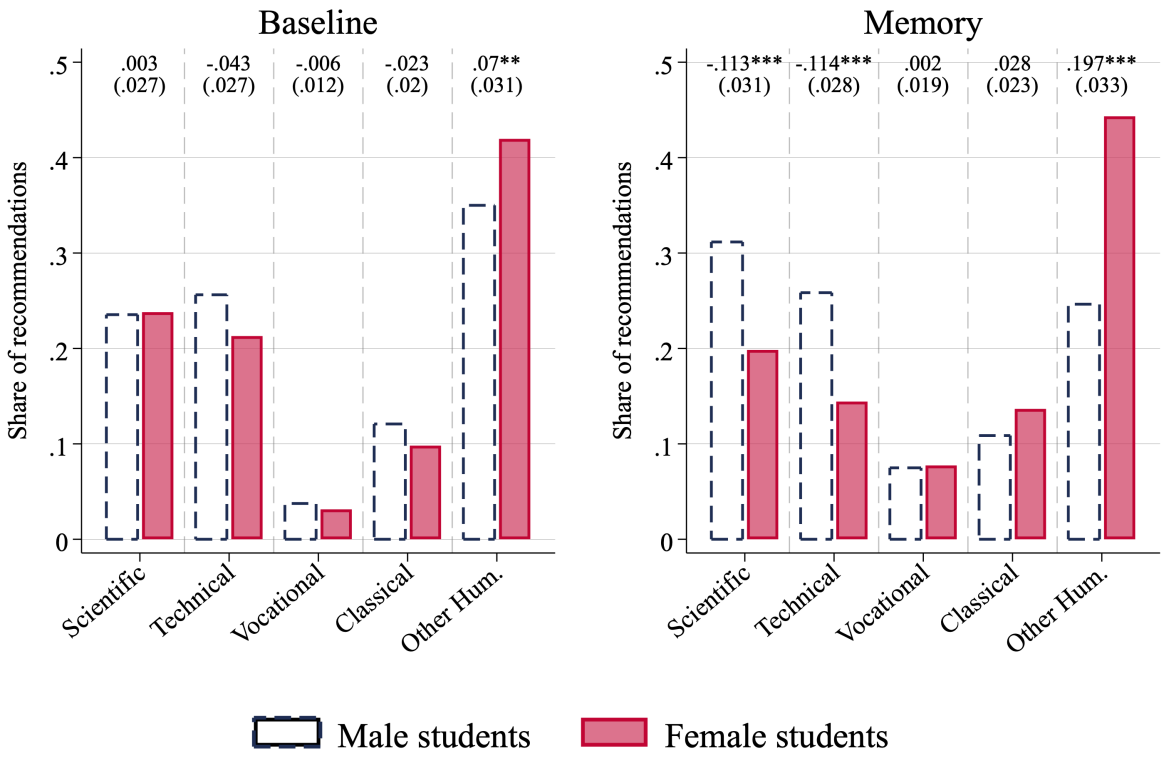
Notes: This table shows coefficients β_2 and β_3 from the estimation of equation 4 where the dependent variables are dummies indicating if the teacher recalls female-typed characteristics (columns 1–3), and male-typed characteristics (columns 4–6). One observation is a student profile assigned to a teacher (each teacher observes 4 student profiles). Controls include teacher birth year, gender, subject taught (humanistic, scientific, other), father education, type of contract (permanent/fixed term/other), whether the school is in the North, and whether the teacher was born in Northern Italy. Standard errors are clustered at the teacher level. *p<0.1; **p<0.05, ***p<0.01.

Figure C4. Gender Gaps with Actual Students and Behavior in Experiment



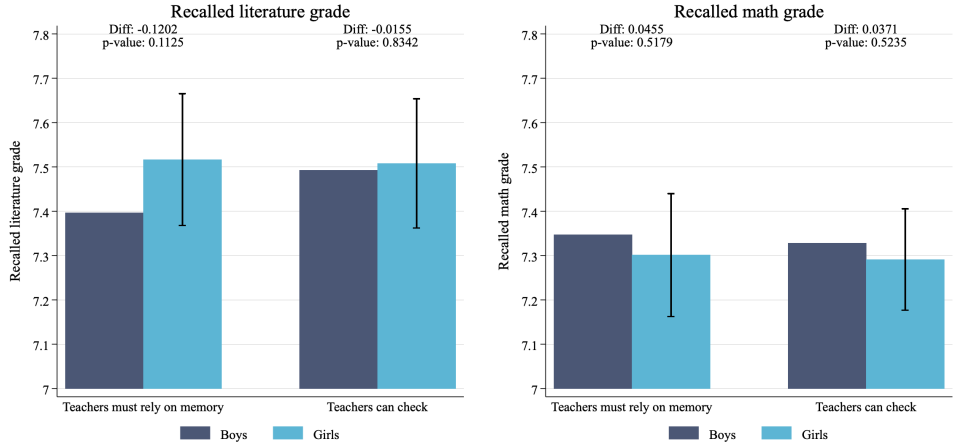
Notes: Panel (a) shows the share of male and female hypothetical students recommended for the science track, categorized by teachers' historical gender gaps in science-track recommendations for past students. The left graph presents data for teachers in the *Memory* condition, while the right graph shows data for teachers in the *Baseline* condition. Panel (b) depicts the probability of recalling female-typed characteristics based on students' gender, distinguishing between teachers with high and low historical gaps with actual students. The colored bars represent teachers in the *Memory* condition, and the dashed bars represent teachers in the *Baseline* condition. Panel (c) provides the same statistics as Panel (b), but focuses on the recall of male-typed characteristics.

Figure C5. Track Recommendations by Gender and Memory Condition



Notes: Each panel shows the share of male (outline, dark blue) and female (filled, red) students recommended to each track, by memory condition. Science, technical, and vocational tracks are classified as male-typed; classical and other humanities tracks (artistic, linguistic, and human sciences) are classified as female-typed, as defined in Section 5.2. The *Baseline* condition allows teachers to review student profiles before deciding; the *Memory* condition requires teachers to rely on memory. Numbers above bars show the female minus male difference with standard errors in parentheses; confidence intervals on female bars reflect the 95 percent confidence interval of the female–male difference from a regression of each track recommendation on an indicator for female student. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Figure C6. Recall of Students' Math and Literature Grades



Notes: This figure shows the share of teachers who recall each grade value for the student's literature grade (left panel) and math grade (right panel), separately by student gender and memory condition. Bars show group means. Caps show 95 percent confidence intervals.

Table C5. Selective recall of grades

	Recalled literature grade			Recalled math grade			Recalled gap lit-math		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.023	0.016	0.021	-0.022	-0.019	-0.011	0.042	0.034	0.034
	(0.032)	(0.032)	(0.038)	(0.025)	(0.026)	(0.030)	(0.050)	(0.050)	(0.059)
Female × Memory	0.105	0.108	0.106	-0.000	-0.009	-0.093	0.088	0.101	0.181
	(0.076)	(0.076)	(0.088)	(0.061)	(0.061)	(0.078)	(0.102)	(0.103)	(0.123)
Mean control	7.500	7.500	7.500	7.310	7.310	7.310	7.310	7.310	7.310
Observations	1534	1534	1511	1544	1544	1522	1523	1523	1501
N. teachers	414	414	391	416	416	394	413	413	391
R^2	0.536	0.540	0.630	0.453	0.455	0.550	0.482	0.485	0.574
Student FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Teacher FE	No	No	Yes	No	No	Yes	No	No	Yes

Notes: This table shows coefficients β_2 and β_3 from the estimation of equations 4 where the dependent variables are the recalled literature grades (columns 1-3), the recalled math grades (columns 4-6), and the gap between the recalled literature-math grades (columns 7-9). Columns 3, 6 and 9 include teachers fixed effects. Controls include: teacher birth year, gender, subject taught (humanistic, scientific, other), father education, type of contract (permanent/fixed term/other), whether the school is in the North, and whether the teacher is born in Northern Italy. Standard errors are clustered at the teacher level.

Table C6. Heterogeneity by profiles with mixed vs. extreme signals

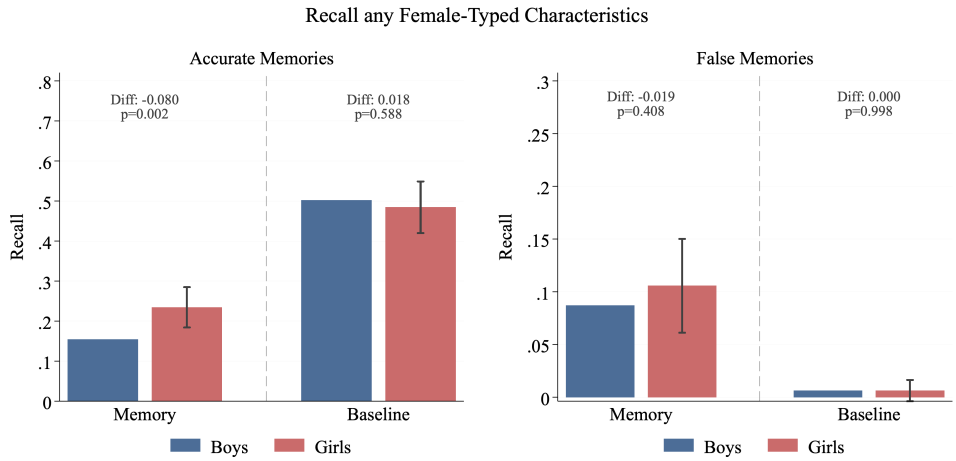
	Extreme Profiles		Mixed Profiles	
	(1)	(2)	(3)	(4)
Panel A: Recall of Female-Typed Characteristics				
Female	0.00134	-0.00211	-0.0254	-0.0302
	(0.0450)	(0.0252)	(0.0461)	(0.0384)
Memory	-0.189***	-0.179***	-0.334***	-0.339***
	(0.0391)	(0.0335)	(0.0385)	(0.0368)
Memory × Female	0.0599	0.0343	0.140**	0.154***
	(0.0653)	(0.0512)	(0.0627)	(0.0582)
Mean control (Baseline-Male)	0.440	0.440	0.556	0.556
R^2	0.0291	0.457	0.0783	0.266
Panel B: Recall of Male-Typed Characteristics				
Female	-0.0314	-0.0277	0.00319	-0.00984
	(0.0457)	(0.0311)	(0.0449)	(0.0447)
Memory	-0.221***	-0.223***	-0.398***	-0.400***
	(0.0410)	(0.0360)	(0.0439)	(0.0435)
Memory × Female	-0.00359	0.00690	-0.0573	-0.0552
	(0.0639)	(0.0514)	(0.0590)	(0.0589)
Mean control (Baseline-Male)	0.506	0.506	0.615	0.615
R^2	0.0527	0.405	0.186	0.202
Observations	878	878	883	883
N. Teachers	444	444	446	446
Student FE	No	Yes	No	Yes
Controls	No	Yes	No	Yes

Notes: This table reports heterogeneity in biased recall by whether the student profiles provide mixed or extreme signals. Extreme profiles are those where the student excels clearly in one subject and struggles in the other. Mixed profiles are those where the student performs well in both. Panel A reports the Memory × Female coefficient on female-typed recall and Panel B reports it for male-typed recall. Columns 1 and 3 report estimates without controls, and columns 2 and 4 include student fixed effects and teacher controls. Standard errors are clustered at the teacher level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

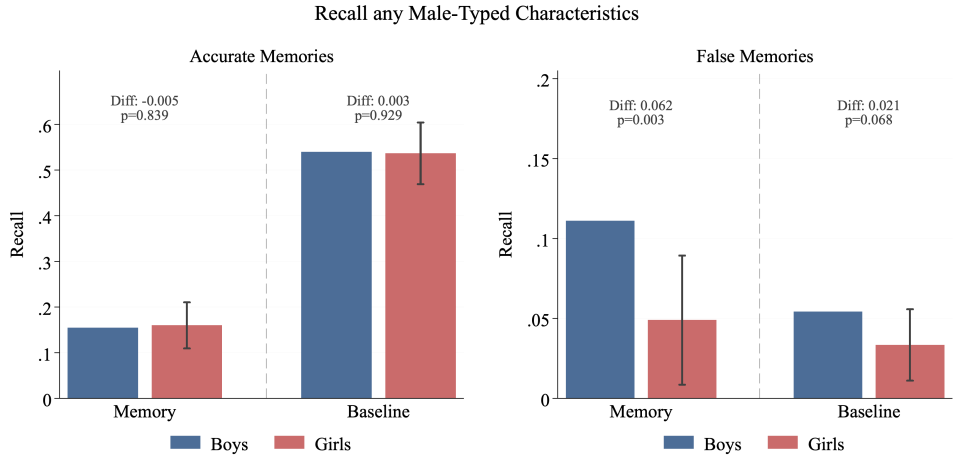
C.4 Accurate recalls and false memories

Figure C7. Accurate Recall and False Memories of Stereotype-Typed Characteristics

Panel (a): Recall any Female-typed characteristics

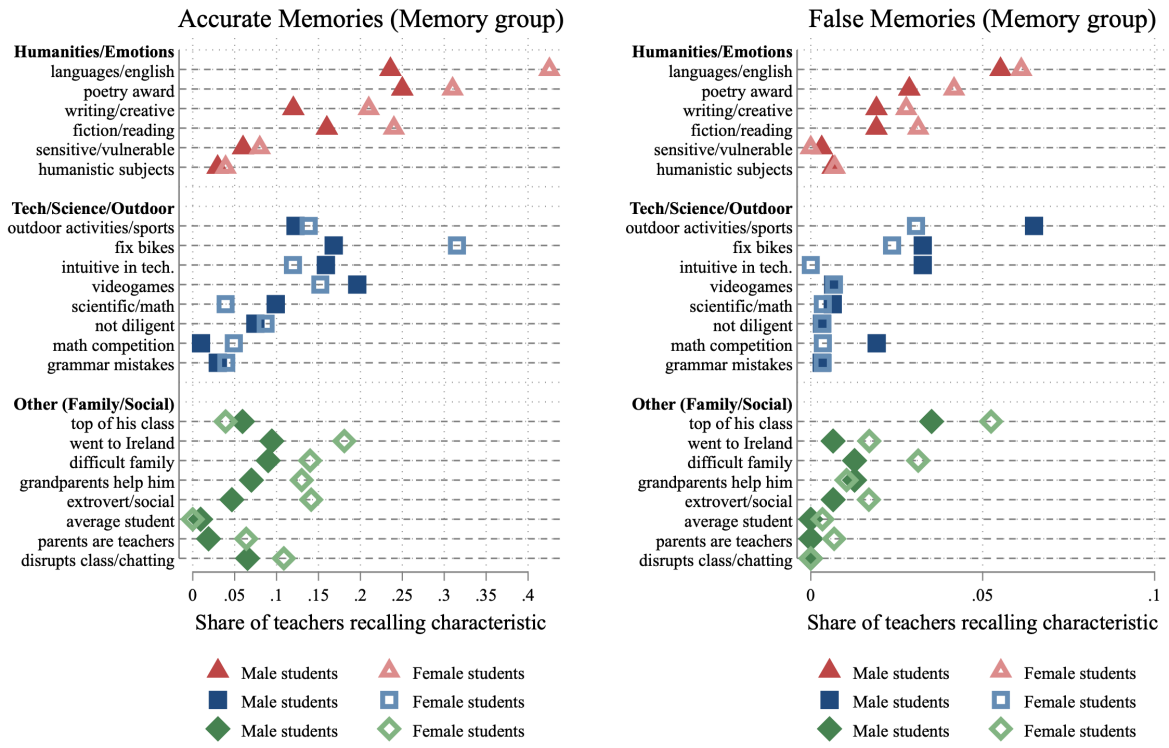


Panel (b): Recall any Male-typed characteristics



Notes: Each panel decomposes recall of female and male-typed characteristics into accurate recalls—characteristics truly belonging to the evaluated student—and false memories—characteristics belonging to a different student but attributed to the evaluated student. Panel (a) reports results for female-typed characteristics; Panel (b) for male-typed characteristics. Both groups observed the same student profiles. Teachers in the *Memory* condition must retrieve student characteristics from memory; teachers in the *Baseline* condition can review profiles before deciding. Sample: 448 teachers from 68 middle schools.

Figure C8. Accurate and False Memories by Student Characteristic



Notes: Each panel shows the share of teachers recalling each student characteristic, decomposed into accurate recalls (characteristic correctly attributed to the student it belongs to) and false memories (characteristic attributed to a different student). The left panel shows accurate recalls of each characteristics for teacher assigned to male and female students; the right panel shows false memories of each characteristics, for teachers assigned to female and male students. Filled markers indicate male students; hollow markers indicate female students. Sample: teachers in the *Memory* condition of the experiment.

C.5 Hypothetical Profiles and Actual Students

Figure C9. Past Students with 8/10 in math and 9/10 in Italian, GPA around 8.5/10 (same grades as *Roberto/Roberta*. Additional characteristics of student profile with respect to past students with the same grades: *participated in a math competition.*

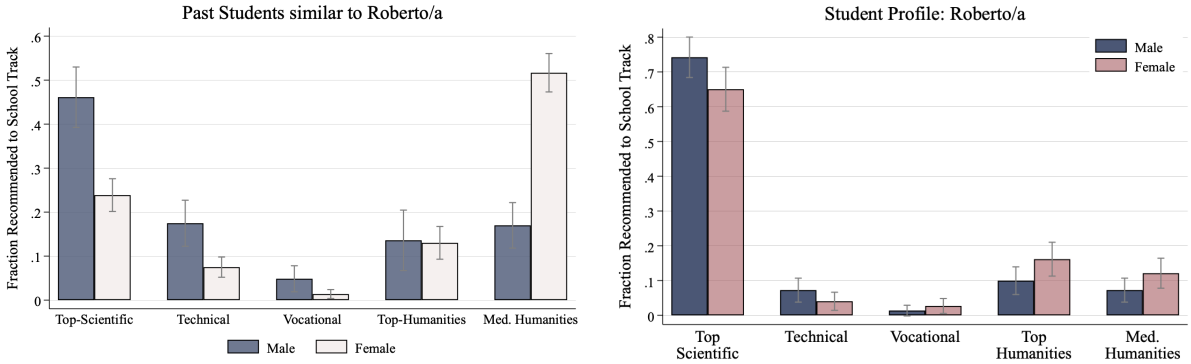


Figure C10. Past Students with 7/10 in math and Italian, GPA around 8, English grade 8 (same grades as *Francesco/Francesca*). Additional characteristics of student profile with respect to past students with the same grades: *very passionate about languages and spent summer in Ireland to learn English.*

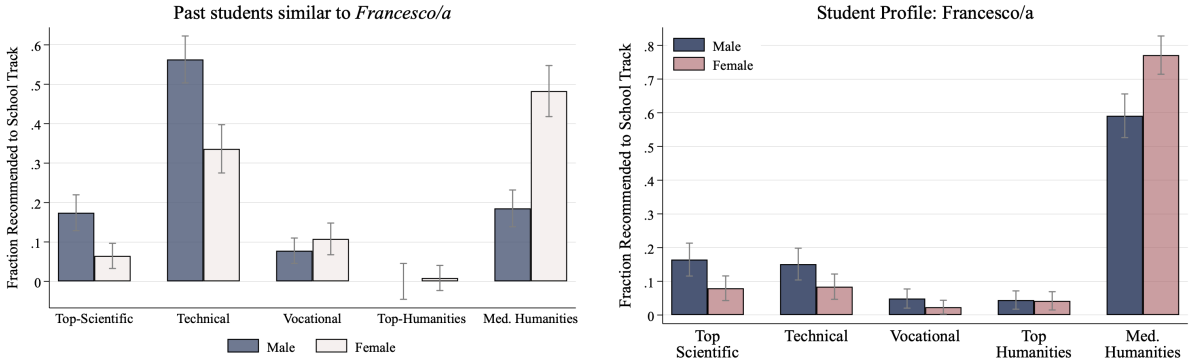


Figure C11. Past Students with 8/10 in math and 6/10 in Italian (same grades as *Marco/Anna*). Additional characteristics of student profile with respect to past students with the same grades: very good in technology, passionate about playing video games and fixing bikes.

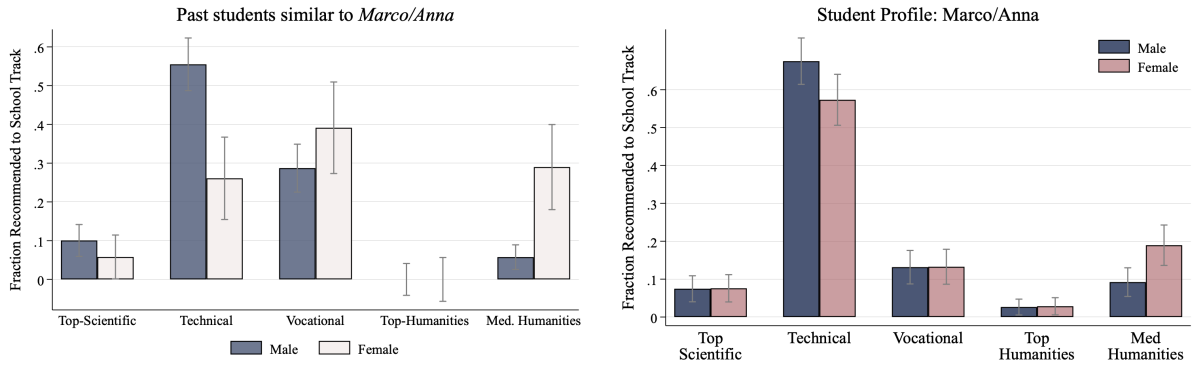
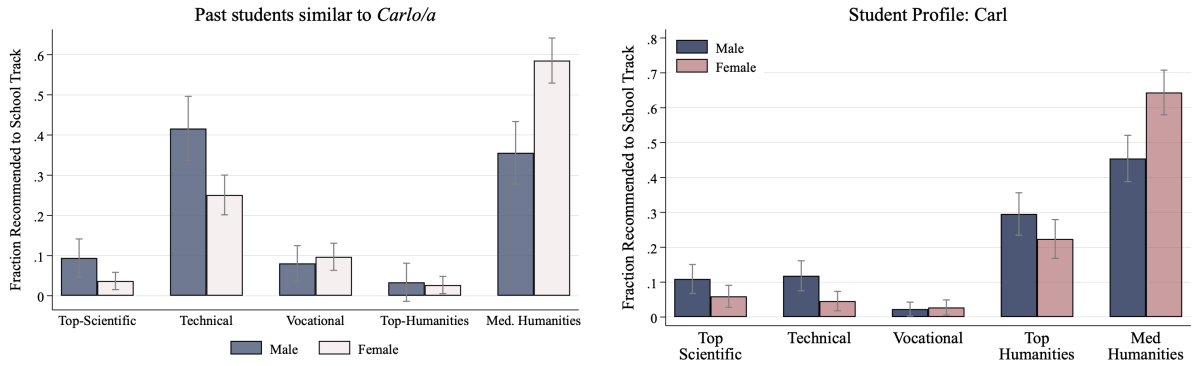


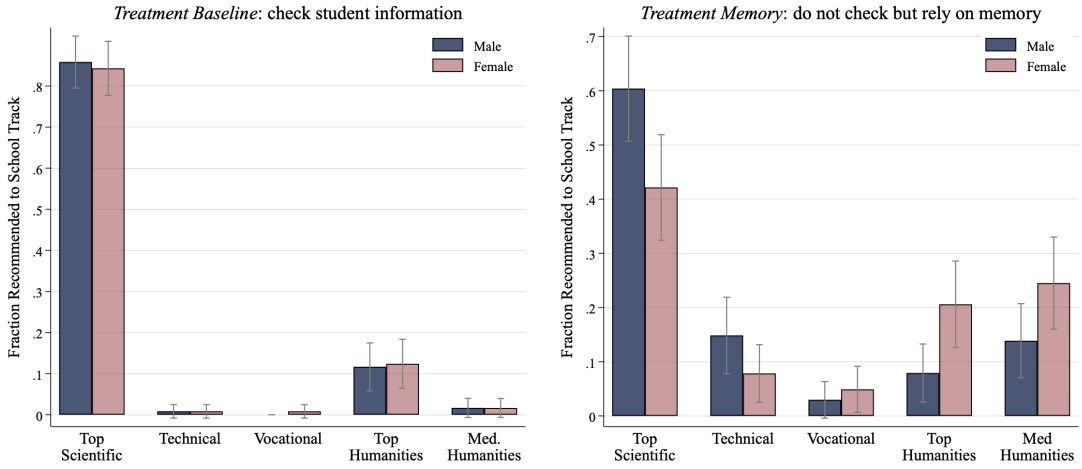
Figure C12. Past Students with 6/10 in math and 8/10 Italian, GPA around 8 (same grades as *Carlo/Carla*). Additional characteristics of student profile with respect to past students with the same grades: loves reading fiction and poetry, creative in essays.



Notes: The histograms above shows the distribution of track recommendations assigned by teachers in the administrative sample to past students with the same grades as the hypothetical profile described above. The modal recommendation in the administrative data is indicated. The comparison assesses whether experimental recommendations for the hypothetical profiles are consistent with teachers' real-world recommendation practices for students with similar academic performance.

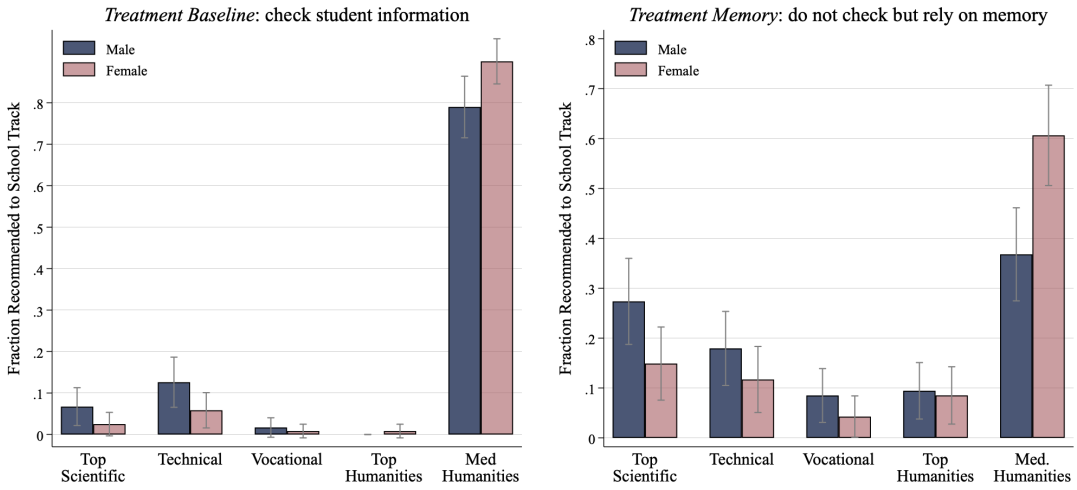
C.6 Recommendations for All Tracks across *Memory* and *Baseline* condition

Figure C13. *Roberto/a:* Excellent Student in Math and Italian



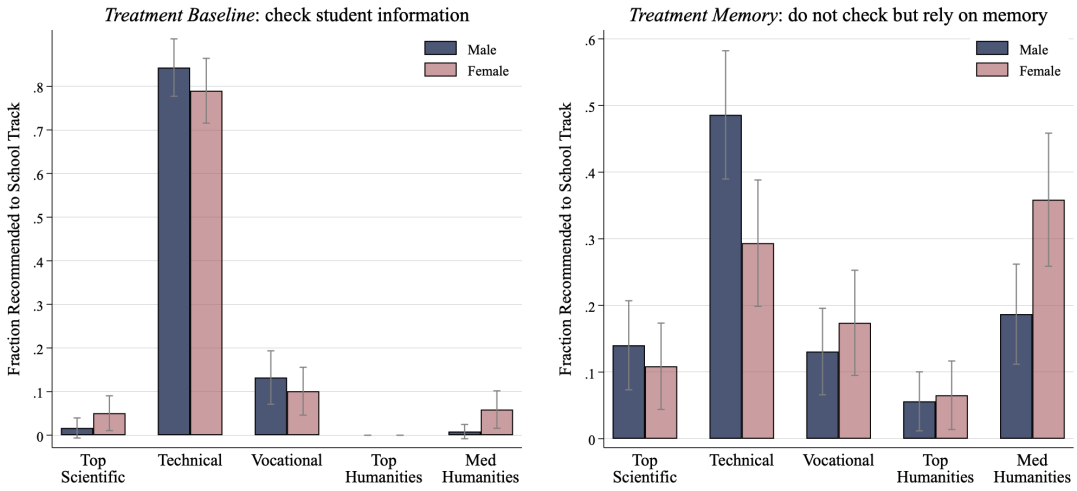
Vignette: *Roberto/a is among the best students in his class both in humanistic and scientific subjects. Last semester, he/she got a 9 in Italian and an 8 in math, and his/her GPA is 8.5. Roberto/a was selected to participate in a math competition at the regional level and he/she reached the final rounds.*

Figure C14. *Francesco/a:* Average student passionate about languages



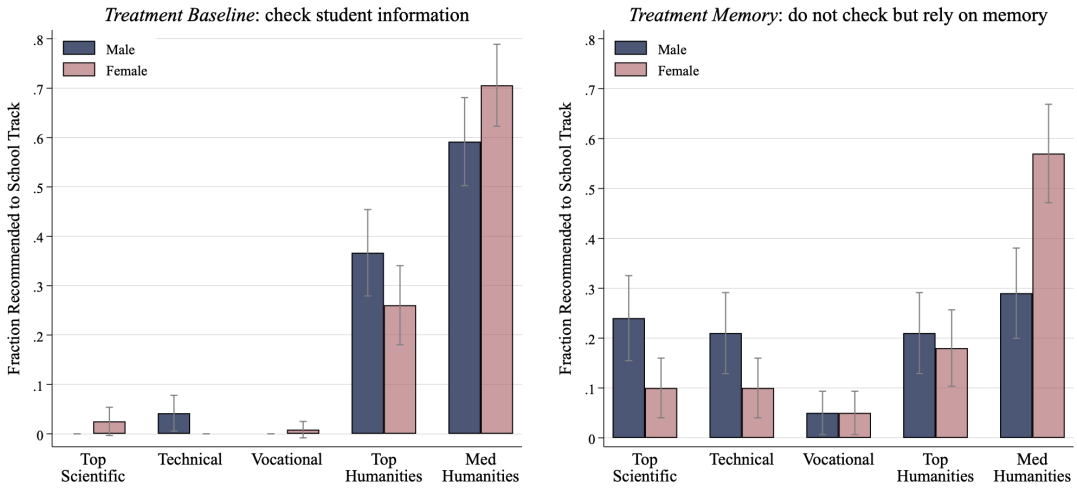
Vignette: *Francesco/a is a good student but not excellent. He got 7 in math and Italian, and his GPA is around 8. He is also very passionate about languages, and he got an 8 in English. He spent 3 weeks in Ireland in the summer where he substantially improved his ability to speak English, which is considerably above average. He cares a lot about his group of friends and both his parents are high school teachers.*

Figure C15. *Marco/Anna:* Good student in math and technical subjects passionate about fixing bikes/videogames



Vignette: *Marco/Anna is a very extroverted and social boy. He is not very diligent at school and he often forgets to do his homework. He often disrupts lectures by chatting with his friends. He is very intuitive and talented in math, where he got 8, while he got 6 in Italian. He is passionate about fixing bikes with his older brother and he loves playing video games.*

Figure C16. *Carlo/a:* Good Student in Literature and Passionate about Poetry



Vignette: *Carlo/a comes from a disadvantaged family background. His father left when he was 5, his mother had some health issues and he mainly lives with his grandparents. However, his grandparents support him a lot in his education, and he manages to do quite well at school. He got a 6 in math and an 8 in Italian, and he got a GPA of 8. He loves reading fiction and poetry. He is very creative in his essays although he often makes grammar mistakes. He also participated in a poetry competition, where he received an award for his poem called "My teenage years as a digital native".*

Notes: Each of the above figures shows the distribution of track recommendations for one of the four hypothetical student profiles, separately by student gender and treatment (left panel is Baseline and right panel is Memory). The y-axis shows the fraction of teachers recommending each track. Bars show means. Caps show 95 percent confidence intervals.

D Back-of-the-Envelope Decomposition of the Recall Channels

I decompose the memory-induced gap Δ into two channels. C_1 captures the *fall-back-on-priors* channel: memory constraints reduce total recall, increasing reliance on prior beliefs when less individual-level information is retrieved. C_2 captures the *biased-recall* channel: conditional on total recall volume, memory constraints shift the composition of recalled signals in a stereotype-consistent direction. The decomposition is an accounting exercise mapping experimentally observed changes in recall into these two channels, rather than a causal mediation analysis.

D.1 Teacher Experiment

The total effect Δ is the coefficient on *Female* $\times$ *Memory* from the baseline specification in Section 5, using the probability of a female-typed track recommendation as outcome. Female-typed recommendations correspond to humanities tracks (classical, linguistic, artistic, human sciences); memory constraints generate a reallocation of girls toward female-typed tracks and away from male-typed tracks (science, technical, and vocational), so this outcome captures the full extent of the Memory-induced gap more completely than the science-track outcome alone (Table 4).[30]

C_1: Fall-back on priors. I partition the total recall volume into four bins: $s = 1$ if recall is ≤ 2 characteristics, $s = 2$ if recall is exactly 3, $s = 3$ if recall is exactly 4, and $s = 4$ if recall is ≥ 5. Bin $s = 4$ is the omitted category. I estimate the following equation (coefficients shown in Figure 6, panel (a)):

$$\begin{aligned} Y_{ij} = \alpha + \sum_{s=1}^{3} \gamma_s \, 1[RecallBin_{ij} = s] \cdot Female_{ij} + \sum_{s=1}^{3} \delta_s \, 1[RecallBin_{ij} = s] \\ + \beta \, Female_{ij} + \rho \, DiffRecallBin_{ij} + \mathbf{X}_j \theta + \mu_i + \varepsilon_{ij}, \end{aligned} \tag{6}$$

where μ_i are student fixed effects, $\mathbf{X}_j$ denotes teacher controls, and $DiffRecallBin_{ij}$ is the recall-content measure defined below. γ_s measures the increase in the gender gap associated with lower recall volumes (conditional on recall content). The estimated contribution of this channel is:

$$C_1 = \sum_{s=1}^{3} \Delta_s^R \cdot \gamma_s, \qquad \Delta_s^R = P(RecallBin = s \mid Memory) - P(RecallBin = s \mid Baseline).$$

C_2: Biased recall. I measure recall content as the difference between the number of female-typed and male-typed characteristics recalled:

$$DiffRecall_{ij} = \#(\text{female-typed recalled}) - \#(\text{male-typed recalled}),$$

grouped into bins $DiffRecallBin_{ij} \in \{\leq -4, -3, -2, -1, 0, 1, 2, 3, \geq 4\}$, with $k = 0$ (balanced recall) as the omitted category. I estimate:

$$\begin{aligned} Y_{ij} = \alpha + \sum_{k \neq 0} \phi_k \, 1[DiffRecallBin_{ij} = k] + \sum_{s=1}^{3} \delta_s \, 1[RecallBin_{ij} = s] \\ + \beta \, Female_{ij} + \mathbf{X}_j \theta + \mu_i + \varepsilon_{ij}. \end{aligned} \tag{7}$$

The coefficients ϕ_k measure how recommendations vary across recall-content bins relative to balanced recall, conditional on total recall volume. The contribution of this channel is: $C_2 = \sum_{k \neq 0} \Delta_k^D \cdot \phi_k$, where Δ_k^D is the

[30] Section 5.2 discusses the definition of female-typed and male-typed recommendations.

gender-differential shift in the recall-content distribution induced by memory constraints:

$$\Delta_k^D = [P(DiffRecallBin = k \mid Memory, Female) - P(DiffRecallBin = k \mid Baseline, Female)] - [P(DiffRecallBin = k \mid Memory, Male) - P(DiffRecallBin = k \mid Baseline, Male)]. \quad (8)$$

The overall effect is decomposed into $\Delta = C_1 + C_2 + R$, where R is the residual not captured by the two recall channels. The two channels together account for 70.7 percent of the total Memory-induced gap (Table D1).

Table D1. Decomposition of recall channels, teacher experiment

	Coefficient	Share (%)
Observed gap increase (Memory × Female)	0.1693	
C1 – Fall-back-on-prior (recall bins)	0.0739	43.6%
C2 – Biased signal recall (diff bins)	0.0481	28.4%
Total explained (C1 + C2)	0.1220	72.0%
Residual	0.0474	28.0%

D.2 Prolific Experiment

The total effect Δ is the coefficient on *Stereotype-Inconsistent* × *Memory* from equation (5), with predicted Quiz 2 performance as the outcome. The structure of the decomposition mirrors Section D.1, replacing student characteristics with remembered trivia questions and track recommendations with performance predictions.

C_1: Fall-back on priors. I partition total recall volume (number of Quiz 1 questions recalled) into quartiles ($s = 1, \ldots, 4$; bin 4 as reference). I estimate how the stereotype-inconsistency gap in assessments varies across recall bins, controlling for recall content and evaluator characteristics:

$$Y_{ijd} = \alpha + \sum_{s=1}^{3} \gamma_s \cdot 1[RecallBin_{ijd} = s] \cdot Incongr_{id} + \sum_{s=1}^{3} \delta_s \cdot 1[RecallBin_{ijd} = s] + \beta \cdot Incongr_{id} + \sum_{k \neq 0} \rho_k \cdot 1[DiffRecallBin_{ijd} = k] + \mathbf{X}_j\theta + \mu_d + \varepsilon_{ijd}. \quad (9)$$

where μ_d are domain fixed effects and $DiffRecallBin_{ij}$ is the recall-content bin variable defined below. The contribution of this channel is:

$$C_1 = \sum_{s=1}^{3} \Delta_s^R \cdot \gamma_s, \qquad \Delta_s^R = P(RecallBin = s \mid Memory) - P(RecallBin = s \mid Baseline).$$

C_2: Biased recall. I measure recall content as $DiffRecall_{ij} = CorrectRecalled_{ij} - IncorrectRecalled_{ij}$—the difference between the number of Quiz 1 questions the evaluator places in the right box and those

she placed in the incorrect answer box. Values higher than zero indicate that the evaluator recalls the candidate as having performed better than she actually did. I bin this variable into nine categories ($k \in \{\leq -4, -3, -2, -1, 0, 1, 2, 3, \geq 4\}$), using $k = 0$ as the reference, and estimate:

$$Y_{ij} = \alpha + \sum_{k \neq 0} \phi_k \cdot 1[DiffRecallBin_{ij} = k] + \sum_{s=1}^{3} \delta_s \cdot 1[RecallBin_{ij} = s] + \beta \cdot Incongr_i + \mathbf{X}_j \theta + \mu_d + \varepsilon_{ij}. \tag{10}$$

The coefficients ϕ_k measure how performance predictions vary with the recalled net-correct score, conditional on total recall volume. The contribution of this channel is $C_2 = \sum_{k \neq 0} \Delta_k^D \cdot \phi_k$, where Δ_k^D is the differential shift in the recall-content distribution between stereotype-inconsistent and stereotype-consistent candidates induced by memory constraints:

$$\Delta_k^D = \Big[P(DiffRecallBin = k \mid Memory, Incongr) - P(DiffRecallBin = k \mid Baseline, Incongr)\Big] - \Big[P(DiffRecallBin = k \mid Memory, Congr) - P(DiffRecallBin = k \mid Baseline, Congr)\Big]. \tag{11}$$

Overall decomposition.

$$\Delta = C_1 + C_2 + R,$$

where R is the residual not captured by the two channels. Table D2 reports the estimated contributions.

Table D2. Decomposition of recall channels, Prolific experiment

	Estimate	% of gap
Total observed gap		
Stereotype-Inconsistent × Memory	-0.0256	100.0%
Decomposition		
C1: Fall-back on priors (recall volume)	-0.0187	73.1%
C2: Biased signal recall (recall composition)	-0.0030	11.6%
Total explained (C1 + C2)	-0.0217	84.7%
Residual	-0.0039	15.3%

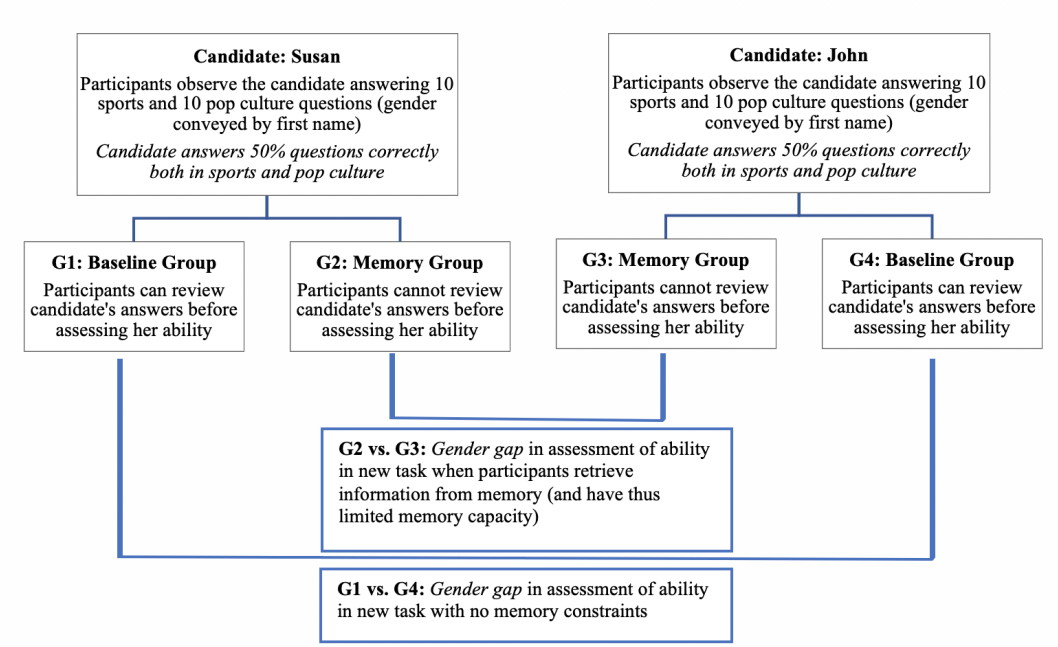
E Prolific Experiment

E.1 Experimental Design and Procedures

Table E1. Pre-registration and Study Materials

Study Material	**Link**
AEA RCT Registry. RCT ID: AEARCTR-0008921	https://www.socialscienceregistry.org/trials/8921
Prolific Experiment Protocol	https://www.dropbox.com/scl/fi/g5f0ez9z0uvczzspqzst8/Prolific_experimental_protocol.pdf?rlkey=ru0dusih3g505c0dndvpzinmi&st=5rhpaci1&dl=0

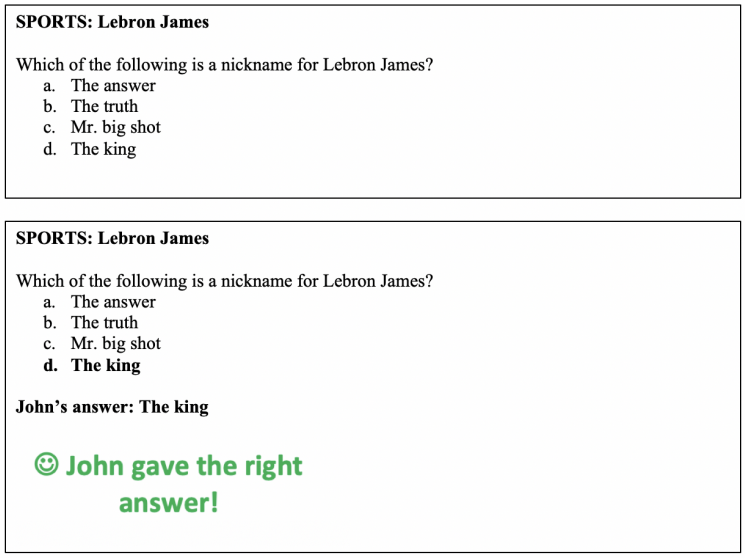
Figure E1. Experiment flow



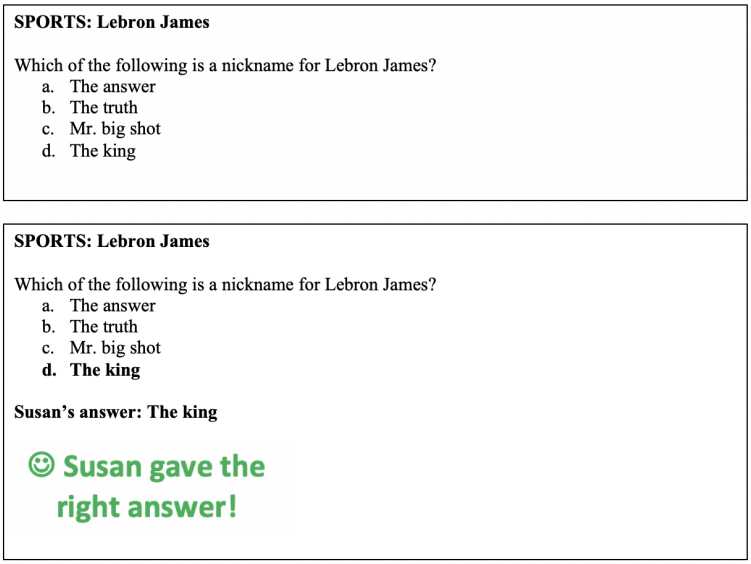
Notes: This figure shows the Prolific experiment design. The groups G1, G2, G3, G4 denote the experimental groups. I randomized the candidate's gender (conveyed through her first name) and whether participants can review their candidate answers before assessing their ability in the new task or whether they need to rely on memory (*Baseline* vs. *Memory* conditions). Participants observe one candidate answering 10 questions in sports and 10 questions in pop culture. The candidate answers correctly 50% of questions in both domains. After observing the candidate, participants are required to recall all correct and incorrect questions answered by the candidate (*free recall* task) and assess her ability in a new similar task. Questions are shown in Figures **??** and **??**. The decision screen for the recall task is shown in Figure E3.

Figure E2. example of a question

Panel (a): Male candidate



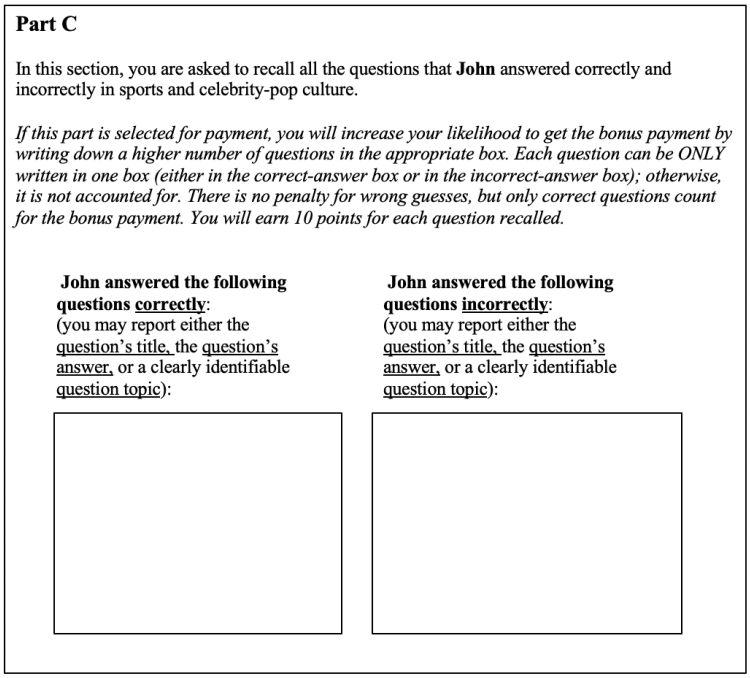
Panel (b): Female candidate



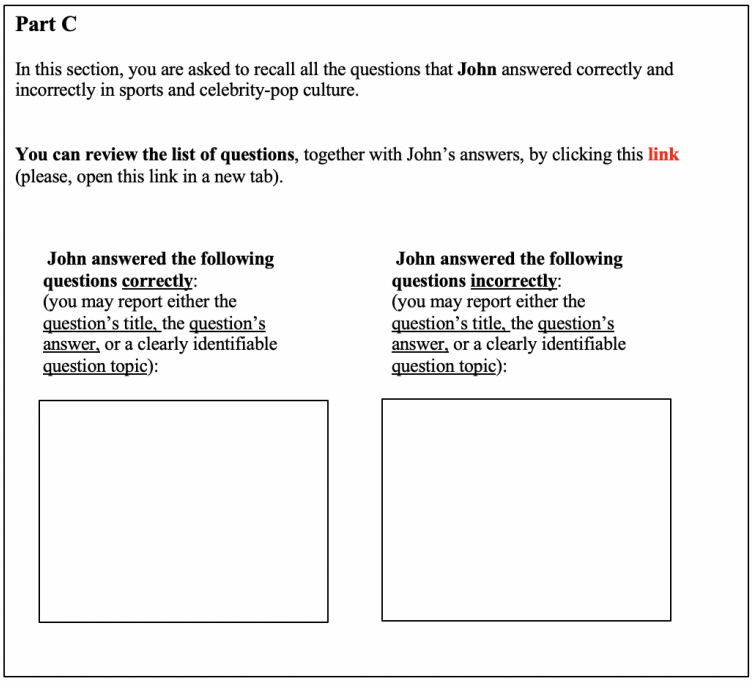
Notes: This figure shows an example of a trivia question as it appeared to evaluators during the observation phase of the Prolific experiment. Each evaluator observed their assigned candidate answering 10 sports questions and 10 pop-culture questions, presented one at a time. For each question, evaluators first saw the question and the four possible answers without the candidate's response, then saw the same question with the correct answer in bold, the candidate's selected answer, and a green smiley-face confirmation when the answer was correct.

Figure E3. Decision screen, free recall task

Panel (a): *Memory* condition



Panel (b): *Baseline* condition

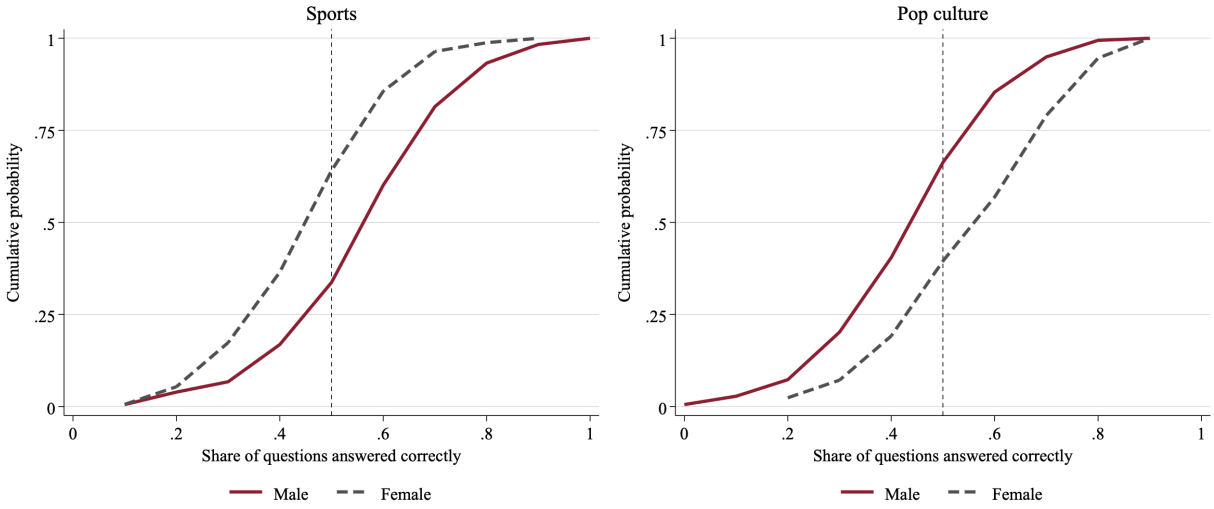


Notes: This figure shows the free recall task screen presented to evaluators in the Prolific experiment after observing the candidate's Quiz 1 performance. Evaluators were asked to list all questions they recalled the candidate answering correctly in the left box and all questions recalled as answered incorrectly in the right box.

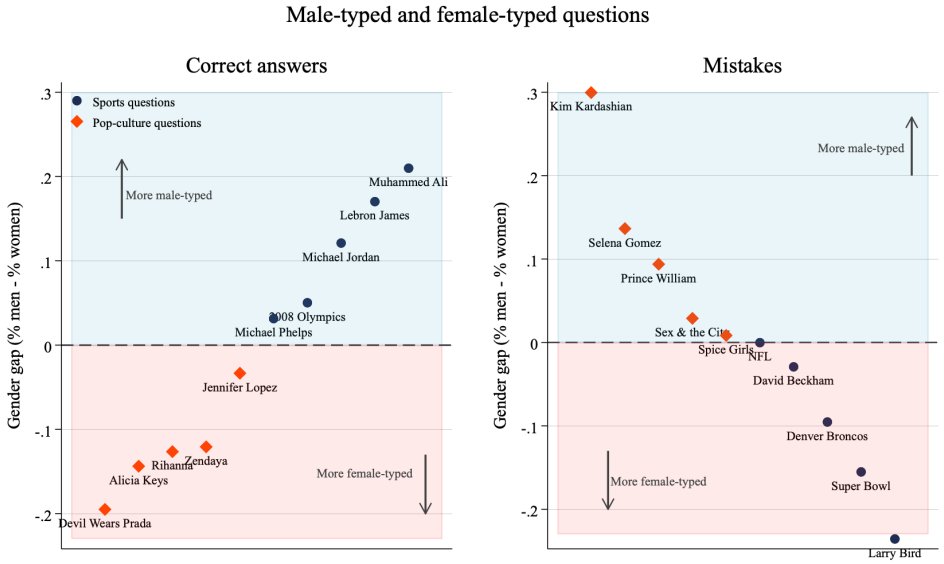
E.2 Additional Tables and Figures

Figure E4. Male-typed and female-typed signals

Panel (A): Share of questions answered correctly by domain. Female and Male typed domain.

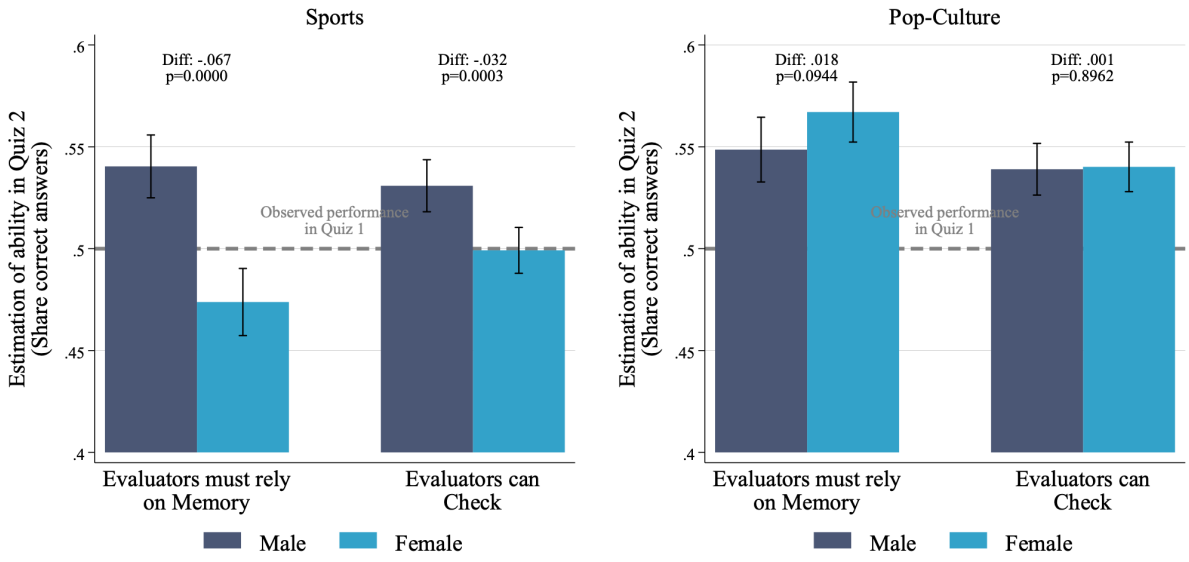


Panel (B): Female and male-typed signals (question-level)



Notes: This figure shows the questions answered correctly (left graph) and incorrectly (right graph) by the candidate ordered by the gender gap in correct answers (left graph) and incorrect answers (right graph) by a larger pool of 400 subjects who answered the questions prior to the main experiment. The left graph shows that the correct sports questions are male-typed, and the correct pop-culture questions are female-typed. The right graph shows that pop-culture mistakes are male-typed while sports mistakes are female-typed.

Figure E5. Predicted Quiz 2 Performance in Sports and Pop Culture



Notes: The figure shows mean predicted Quiz 2 performance (share of questions answered correctly) by treatment conditions, separately for sports (left panel) and pop culture (right panel). Bars show group means; caps show 95 percent confidence intervals. The dashed line at 0.5 indicates observed Quiz 1 performance, which is identical across candidates and genders. The male candidate is stereotype-consistent in sports; the female candidate is stereotype-consistent in pop culture. Annotations report the mean gender difference (Female − Male) and its p-value within each condition, with standard errors clustered at the evaluator level.

Table E2. Balance table Prolific experiment

Variable	(1) Baseline	(2) SD	(3) Memory	(4) SD	(5) P-val.	(6) Std. Diff.
Baseline vs. Memory treatment						
Age	34.634	(12.159)	35.884	(13.089)	(0.116)	0.070
Gender (1=male)	0.509	(0.500)	0.491	(0.500)	(0.572)	-0.025
Republican	0.121	(0.327)	0.116	(0.321)	(0.810)	-0.011
More than high school	0.873	(0.334)	0.884	(0.321)	(0.598)	0.023
Number rejections on Prolific	2.515	(3.418)	2.639	(3.198)	(0.551)	0.027
Number of approvals on Prolific	813.590	(606.760)	816.834	(611.055)	(0.933)	0.004
Employed (full-time or part-time)	0.653	(0.477)	0.636	(0.482)	(0.672)	-0.025
Student	0.237	(0.426)	0.213	(0.410)	(0.493)	-0.040
Prior belief performance sports quiz 1	5.592	(1.960)	5.521	(1.945)	(0.560)	-0.026
Prior belief performance pop-culture quiz 1	5.795	(1.663)	5.941	(1.756)	(0.176)	0.060
Male vs. Female treatment						
Variable	(1) Male	(2) SD	(3) Female	(4) SD	(5) P-val.	(6) Std. Diff.
Age	35.311	(12.649)	35.210	(12.650)	(0.898)	-0.006
Gender (1=male)	0.527	(0.500)	0.472	(0.500)	(0.078)*	-0.078
Republican	0.138	(0.345)	0.099	(0.299)	(0.053)*	-0.086
More than high school	0.870	(0.337)	0.887	(0.317)	(0.397)	0.038
Number rejections on Prolific	2.679	(3.164)	2.472	(3.452)	(0.320)	-0.044
Number of approvals on Prolific	819.056	(600.057)	811.242	(617.949)	(0.838)	-0.009
Employed (full-time or part-time)	0.625	(0.485)	0.662	(0.474)	(0.344)	0.055
Student	0.214	(0.411)	0.235	(0.425)	(0.554)	0.035
Observations	514		496			

Notes: This table reports summary statistics and balance tests for the 1,010 evaluators recruited through Prolific. The upper panel compares evaluators in the Baseline and Memory conditions; the lower panel compares evaluators assigned to the male and female candidate. Columns 1 and 3 report group means; columns 2 and 4 report standard deviations; column 5 reports the p-value from a t-test of equality of means; column 6 reports the standardized difference).

Table E3. Prolific Experiment: Recalled Questions on Quiz 1

	Recalled N correct			Recalled N incorrect			Share recalled correct		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	All	Male	Female	All	Male	Female	All	Male	Female
Stereotype-Inconsistent × Memory	-0.0247	-0.0719	0.0170	0.172**	0.236**	0.132	-0.0533**	-0.0827**	-0.0380
	(0.0786)	(0.113)	(0.115)	(0.0753)	(0.105)	(0.110)	(0.0221)	(0.0354)	(0.0286)
Stereotype-Inconsistent	-0.0431	-0.0142	-0.0584	-0.0506	-0.0954	0.0185	0.00588	0.0230*	-0.0108
	(0.0454)	(0.0657)	(0.0673)	(0.0408)	(0.0597)	(0.0615)	(0.00932)	(0.0137)	(0.0134)
Memory	-2.299***	-2.339***	-2.230***	-2.194***	-2.250***	-2.092***	0.00468	0.0147	-0.00667
	(0.0991)	(0.150)	(0.137)	(0.102)	(0.155)	(0.138)	(0.0138)	(0.0222)	(0.0182)
Mean Control	4.159	3.953	4.345	4.163	3.953	4.353	0.496	0.491	0.501
Stereotype-Inconsistent × Memory as %	-0.594%	-1.818%	0.391%	4.120%	5.960%	3.021%	-10.731%	-16.821%	-7.587%
R-squared	0.427	0.421	0.442	0.382	0.360	0.414	0.0251	0.0443	0.0416
Obs.	2018	998	986	2018	998	986	1812	846	933
domain FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample	No pilot	No pilot	No pilot	No pilot	No pilot	No pilot	No pilot	No pilot	No pilot

Notes: This table reports estimates of equation 5, using counts of recalled questions as dependent variables. Columns 1 to 3 report the number of questions recalled as correct, columns 4 to 6 report the number of questions recalled as incorrect and columns 7 to 9 report the share of all recalled questions placed in the correct box. One observation is an evaluator–candidate–domain pair. Stereotype-Inconsistent equals one when the candidate's gender is inconsistent with the domain (female in sports or male in pop culture). Memory equals one for evaluators in the Memory condition. The key coefficient of interest is Stereotype-Inconsistent × Memory: a positive coefficient in columns 4 to 6 indicates that evaluators under memory constraints place more questions in the incorrect box for stereotype-inconsistent candidates (e.g. for female in sports). All columns include domain fixed effects and demographic controls. Standard errors are clustered at the evaluator level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table E4. Prolific Experiment: Recalled Questions on Quiz 1, alternative measure of recall

	Recalled N correct			Recalled N incorrect			Share recalled correct		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	All	Male	Female	All	Male	Female	All	Male	Female
Stereotype-Inconsistent × Memory	0.00841	-0.00109	0.0106	0.171**	0.255**	0.114	-0.0489**	-0.0645*	-0.0454
	(0.0786)	(0.111)	(0.116)	(0.0752)	(0.106)	(0.109)	(0.0217)	(0.0343)	(0.0286)
Stereotype-Inconsistent	-0.0406	-0.0434	-0.0208	-0.0798**	-0.131**	-0.00830	0.0105	0.0235*	-0.00177
	(0.0452)	(0.0651)	(0.0673)	(0.0403)	(0.0578)	(0.0616)	(0.00906)	(0.0135)	(0.0129)
Memory	-2.280***	-2.315***	-2.205***	-2.125***	-2.188***	-2.021***	0.00151	0.00644	-0.00430
	(0.0982)	(0.152)	(0.133)	(0.103)	(0.160)	(0.137)	(0.0136)	(0.0217)	(0.0179)
Mean Control	4.229	4.024	4.416	4.243	4.020	4.450	0.495	0.492	0.497
Stereotype-Inconsistent × Memory as %	0.199%	-0.027%	0.240%	4.036%	6.336%	2.568%	-9.874%	-13.112%	-9.126%
R-squared	0.418	0.399	0.452	0.365	0.333	0.413	0.0224	0.0416	0.0383
Obs.	2018	998	986	2018	998	986	1828	853	942
domain FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample	No pilot	No pilot	No pilot	No pilot	No pilot	No pilot	No pilot	No pilot	No pilot

Notes: This table replicates Table D2 using an alternative coding of recall in which a question is counted as recalled if at least one of the two independent coders classified it, rather than requiring agreement from both. All columns include domain fixed effects and demographic controls. Pilot participants are excluded. Standard errors are clustered at the evaluator level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table E5. False memories, evaluator by domain analysis

	False: N. truly incorrect recalled as correct			False: N. truly correct recalled as incorrect		
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Male	Female	All	Male	Female
Stereotype-Inconsistent × Memory	-0.0163	0.00287	-0.0350	0.0436**	0.0603*	0.0264
	(0.0180)	(0.0271)	(0.0252)	(0.0212)	(0.0322)	(0.0295)
Stereotype-Inconsistent	0.00462	0.0102	-0.000626	-0.00617	-0.0235*	0.0131
	(0.00752)	(0.0111)	(0.0107)	(0.00922)	(0.0131)	(0.0143)
Memory	0.0755***	0.0770**	0.0824***	0.0399**	0.0667**	0.0259
	(0.0210)	(0.0374)	(0.0253)	(0.0189)	(0.0300)	(0.0239)
Mean Control	0.0159	0.0118	0.0210	0.0319	0.0276	0.0378
R-squared	0.0539	0.108	0.0728	0.0473	0.0819	0.0762
N. Obs	2018	998	986	2018	998	986
Domain FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes

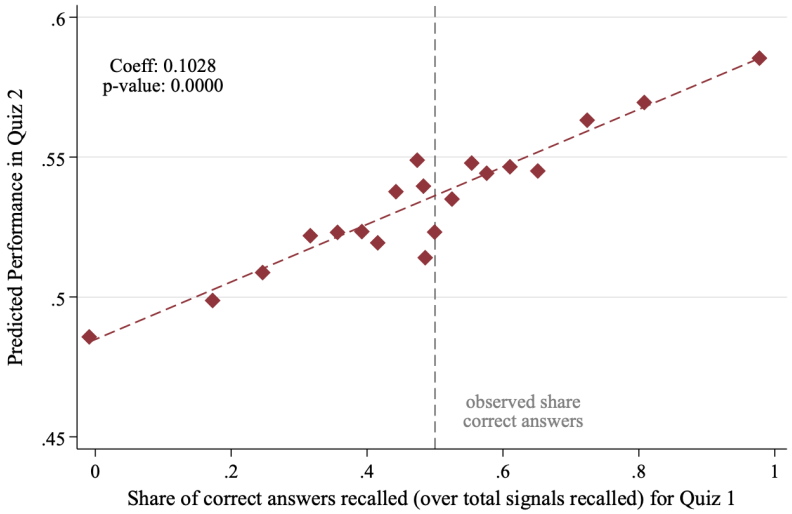
Notes: One observation is an evaluator–candidate–domain pair. The table reports estimates for false memories (misattribution): in columns 1 to 3, the outcome is the number of truly incorrect questions placed in the correct box (thus remembered as correct) and in columns 4 to 6, it is the number of truly correct questions placed in the incorrect box (remembered as incorrect). Stereotype-Inconsistent is 1 if the candidate is female in sports or male in pop culture. All columns include domain fixed effects and demographic controls. Standard errors are clustered at the evaluator level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table E6. Number of recalled questions and predicted performance in Quiz 2

	Predicted Performance in Quiz 2	
	(1)	(2)
	Post Score	Post Score
Selective: truly correct, recalled as correct (number)	0.0154***	0.0161***
	(0.00356)	(0.00537)
Selective: truly incorrect, recalled as incorrect (number)	-0.0132***	-0.0180***
	(0.00353)	(0.00538)
False: truly incorrect, recalled as correct (number)	0.0186	0.0623***
	(0.0125)	(0.0193)
False: truly correct, recalled as incorrect (number)	-0.0503***	-0.0337**
	(0.0117)	(0.0151)
R-squared	0.104	0.622
N. Obs	1014	1014
Domain FE	Yes	Yes
Evaluator FE	No	Yes
Demographic controls	Yes	Yes
Mean DV	0.533	0.533

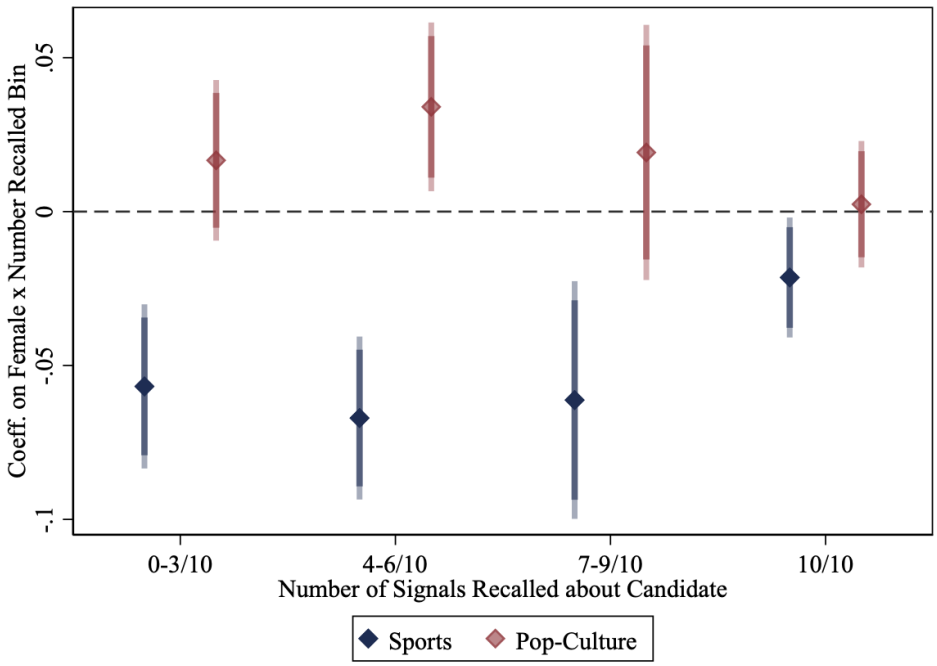
Notes: This table estimates the relationship between the content of recalled Quiz 1 performance and predicted Quiz 2 performance. One observation is an evaluator–domain pair. The dependent variable is the evaluator's predicted share of correct answers in Quiz 2 for the assigned domain. The four independent variables decompose recall into its components: truly correct questions recalled as correct (selective recall of successes), truly incorrect questions recalled as incorrect (selective recall of failures), truly incorrect questions recalled as correct (false memories of success), and truly correct questions recalled as incorrect (false memories of failure). Column 1 includes domain fixed effects and demographic controls without evaluator fixed effects, and column 2 additionally includes evaluator fixed effects. Standard errors are clustered at the evaluator level. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Figure E6. Prolific Experiment: Predicted Performance in Quiz 2 and biased recall of Quiz 1 Performance



Notes: The figure shows the binned scatterplot of the relationship between the estimated ability in the new task and the recalled share of correct questions in the old task (controlling for the number of signals recalled, evaluator and domain fixed effects). The sample includes evaluators in the Memory treatment of the experiment with the US survey sample recruited through Prolific. The graph shows the β_1 coefficient from estimating the following equation, where i is a candidate evaluated by j in domain d: $\text{Y}_{ijd} = \beta_0 + \beta_1 \text{Recalled Share Correct}_{ijd} + \beta_2 \text{Tot. Recall}_{ijd} + \mu_d + \nu_j + \varepsilon_{ij}$.

Figure E7. Predicted Quiz 2 Performance and Number of Signals Recalled



Notes: The figure shows the relationship between the total number of Quiz 1 signals recalled and the gender gap in predicted Quiz 2 performance, conditional on recall bias—measured as the difference between the number of questions recalled as correct and incorrect in the domain—and including evaluator and domain fixed effects. Recalling fewer signals is associated with a higher predicted ability gap in the stereotype-consistent direction: evaluators who recall fewer signals predict higher Quiz 2 performance for male candidates in sports and for female candidates in pop-culture. Caps show 95 percent confidence intervals clustered at the evaluator level.

F Conceptual Framework

A teacher evaluates student i belonging to group $g \in \{M, F\}$. The teacher recommends the scientific track if and only if assessed math ability exceeds a threshold $\bar{\theta}$.

True ability. All students have identical ability distributions regardless of gender:

$$\theta_i \sim \mathcal{N}(\mu, \sigma_\theta^2), \quad \text{for all } g \tag{6}$$

Teacher's prior. The teacher holds a biased prior about group ability:

$$\mu_M = \mu + \delta, \quad \mu_F = \mu - \delta, \quad \delta \geq 0 \tag{7}$$

with prior precision $\tau_a > 0$. Since true ability is identical across groups, $\delta > 0$ reflects a biased belief rather than a true ability difference.

Signals. Over the course of the term, the teacher observes T signals about student i:

$$s_{it} = \theta_i + \varepsilon_{it}, \quad \varepsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_\varepsilon^{-1}), \quad t = 1, \ldots, T \tag{8}$$

Each signal s_{it} is accompanied by a context variable x_{it}—reflecting the circumstances of the interaction, the subject matter, and the student's gender—that affects recall but carries no information about θ_i.

Cognitive load. The teacher evaluates S students simultaneously. Higher S increases interference among competing memories, reducing recall volume (Kahana, 2012). Memory constraints also activate stereotype-consistent retrieval: when unable to consult records, the teacher uses a gender cue to retrieve signals from memory, retrieving with higher probability those signals that are more consistent with the gender stereotype.

Memory Technology

At evaluation time, the teacher retrieves signals about student i from memory. A signal s_{it} is *stereotype-consistent* for group g if, according to the teacher's beliefs, it is disproportionately likely to come from a student of group g relative to group g'—the representativeness principle of Bordalo et al. (2023). Under the teacher's believed signal distributions, with $\mu_M > \mu_F$, the likelihood ratio satisfies:

$$\frac{f(s_{it} \mid g = M)}{f(s_{it} \mid g = F)} = \exp\left(\frac{2\delta}{\sigma_s^2} \cdot (s_{it} - \mu)\right) \tag{9}$$

which exceeds one if and only if $s_{it} > \mu$. Stereotype-consistency therefore reduces to:

$$\text{SC}(s_{it}, g) = \begin{cases} 1 & \text{if } s_{it} \geq \mu \text{ and } g = M \\ 1 & \text{if } s_{it} \leq \mu \text{ and } g = F \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Above-average signals are stereotype-consistent for male students; below-average signals for female students. Under the symmetric signal distribution, $\Pr(\text{SC} = 1) = \frac{1}{2}$ holds exactly for both groups. In the teacher experiment, the empirical counterpart is characteristics disproportionately associated with one gender in the administrative data. In the Prolific experiment, it is the questions answered correctly (incorrectly) at disproportionately higher rates by one gender in the relevant domain. The probability of recalling signal s_{it} for a student of group g is:

$$r(s_{it} \mid g) = \underbrace{\rho(S)}_{\text{interference}} \cdot \underbrace{\frac{\exp\big(\beta \cdot \text{SC}(s_{it}, g)\big)}{\exp\big(\beta \cdot \text{SC}(s_{it}, g)\big) + \exp\big(\beta \cdot (1 - \text{SC}(s_{it}, g))\big)}}_{\text{stereotype-consistent retrieval}} \tag{11}$$

The two components enter multiplicatively and are independently parametrised.

Component 1: Interference. $\rho(S) = \exp(-\lambda S)$ captures the reduction in recall volume as cognitive load

increases: $\rho(0) = 1$, $\rho'(S) < 0$, and $\lim_{S\to\infty} \rho(S) = 0$. The parameter $\lambda > 0$ governs the rate of interference.[31]

Component 2: Stereotype-consistent retrieval. The logistic term captures the similarity-based retrieval principle of Bordalo et al. (2025): signals more consistent with the active gender cue are retrieved with higher probability. The parameter $\beta \geq 0$ governs the strength of this bias: when $\beta = 0$ retrieval is unbiased conditional on volume; as $\beta \to \infty$ only stereotype-consistent signals are recalled. The odds ratio of recalling a stereotype-consistent versus stereotype-inconsistent signal is:

$$\frac{r\big(\mathrm{SC}(s_{it}, g) = 1 \mid g\big)}{r\big(\mathrm{SC}(s_{it}, g) = 0 \mid g\big)} = \exp(\beta) \tag{12}$$

so β is identified empirically by the log odds ratio of recalling stereotype-consistent versus stereotype-inconsistent signals in the free recall data.

Recalled signals. Let K_i denote the total number of signals recalled about student i. Since $\Pr(\mathrm{SC} = 1) = \frac{1}{2}$ holds exactly, the average logistic weight in equation (11) equals $\frac{1}{2}$ regardless of β, and:

$$\mathbb{E}[K_i] = \tfrac{1}{2} \cdot T \cdot \rho(S) \tag{13}$$

strictly decreasing in S and independent of β. The two components thus operate on distinct margins: $\rho(S)$ governs recall volume and β governs recall composition, with no cross-effects in expectation. Denote the set of recalled signals $R_i \subseteq \{s_{i1}, \ldots, s_{iT}\}$ with $|R_i| = K_i$.

Belief Updating

Naive Bayesian updating. The teacher is unaware of her recall bias. She treats the recalled sample R_i as if it were a random draw from the full signal history and applies Bayes' rule:

$$\hat{\theta}_i = (1 - \omega_{K_i})\, \mu_g + \omega_{K_i}\, \bar{s}_i^R \tag{13}$$

where the precision weight on recalled signals is:

$$\omega_{K_i} = \frac{K_i \tau_\varepsilon}{\tau_a + K_i \tau_\varepsilon} \in [0, 1) \tag{14}$$

increasing in K_i, and the average recalled signal is:

$$\bar{s}_i^R = \frac{1}{K_i} \sum_{s_{it} \in R_i} s_{it} \tag{15}$$

[31] The exponential form captures the well-documented exponential decline in recall probability as the number of competing memory traces increases (Kahana, 2012), applied here to the number of students evaluated simultaneously. It delivers the same qualitative mechanism as the ratio-based recall function of Bordalo et al. (2023) —recall of a target strictly decreasing in the number of competing experiences. The key advantage over the ratio formula is that it separates the volume channel from the composition channel, a separation the ratio formula does not allow, since both channels are driven by the same similarity function $S(\cdot)$. This separation enables the two-channel identification in Proposition 1.

Three features of equation 13 are worth noting. First, when $K_i = 0$ the teacher has no recalled evidence and falls back entirely on the group prior: $\hat{\theta}_i = \mu_g$. This is the *fall-back-on-prior* channel—active whenever recall volume is low, because a lower weight ω_{K_i} means the teacher must rely more heavily on the prior μ_g. Second, as $K_i \to \infty$ the weight $\omega_{K_i} \to 1$ and individual signals dominate the prior. Third, the teacher weights recalled signals as if they were unbiased draws, so the bias in $\bar{s}_i^R$ translates directly into a biased assessment—this is the *biased signal recall* channel.

Lemma 1 (Recalled signal mean)**.** *Under the recall technology in equation* (11)*, the expected average recalled signal for student i of group g satisfies:*

$$\mathbb{E}\left[\bar{s}_i^R \,\middle|\, g\right] = \mu + \kappa_g \cdot B(\beta, \sigma_\theta, \tau_\varepsilon) \tag{16}$$

where $\kappa_M = +1$*,* $\kappa_F = -1$*, and:*

$$B(\beta, \sigma_\theta, \tau_\varepsilon) \equiv \frac{e^\beta - 1}{e^\beta + 1} \cdot 2\sigma_s \varphi(0) \geq 0, \qquad \sigma_s^2 \equiv \sigma_\theta^2 + \tau_\varepsilon^{-1} \tag{17}$$

with $\varphi(0) = 1/\sqrt{2\pi}$ *the standard normal density at zero. B is strictly increasing in* β*: when* $\beta = 0$ *retrieval is unbiased and* $B = 0$*; as* $\beta \to \infty$ *only stereotype-consistent signals are recalled and* $B \to 2\sigma_s\varphi(0)$*. B is also increasing in* σ_s*: greater signal dispersion implies a larger gap between the mean of the stereotype-consistent and inconsistent parts of the signal distribution, amplifying the distortion in the average recalled signal.*

Proof. Since $s_{it} = \theta_i + \varepsilon_{it}$ with $\theta_i \sim \mathcal{N}(\mu, \sigma_\theta^2)$ and $\varepsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_\varepsilon^{-1})$, marginalising over θ_i gives $s_{it} \sim \mathcal{N}(\mu, \sigma_s^2)$. Under the SC definition with threshold μ, $\Pr(\text{SC} = 1) = \Pr(s_{it} \geq \mu) = \frac{1}{2}$ holds exactly by symmetry of the normal distribution.

By Bayes' rule:

$$\Pr(\text{SC} = 1 \mid \text{recalled}, g = M) = \frac{\Pr(\text{recalled} \mid \text{SC} = 1) \cdot \Pr(\text{SC} = 1)}{\Pr(\text{recalled})} \tag{18}$$

The recall probabilities from equation (11) are $\Pr(\text{recalled} \mid \text{SC} = 1) = \rho(S) \cdot \frac{e^\beta}{e^\beta + 1}$ and $\Pr(\text{recalled} \mid \text{SC} = 0) = \rho(S) \cdot \frac{1}{e^\beta + 1}$. Since $\Pr(\text{SC} = 1) = \Pr(\text{SC} = 0) = \frac{1}{2}$:

$$\Pr(\text{recalled}) = \frac{e^\beta}{e^\beta + 1} \cdot \tfrac{1}{2} + \frac{1}{e^\beta + 1} \cdot \tfrac{1}{2} = \tfrac{1}{2} \tag{19}$$

Substituting into (18):

$$\Pr(\text{SC} = 1 \mid \text{recalled}, g = M) = \frac{e^\beta}{e^\beta + 1}, \qquad \Pr(\text{SC} = 0 \mid \text{recalled}, g = M) = \frac{1}{e^\beta + 1} \tag{20}$$

By the law of iterated expectations:

$$\begin{aligned}\mathbb{E}[s_{it} \mid \text{recalled}, g = M] &= \mathbb{E}[s_{it} \mid \text{SC} = 1, g = M] \cdot \Pr(\text{SC} = 1 \mid \text{recalled}, g = M) \\ &\quad + \mathbb{E}[s_{it} \mid \text{SC} = 0, g = M] \cdot \Pr(\text{SC} = 0 \mid \text{recalled}, g = M) \\ &= \mathbb{E}[s_{it} \mid s_{it} \geq \mu] \cdot \frac{e^{\beta}}{e^{\beta}+1} + \mathbb{E}[s_{it} \mid s_{it} < \mu] \cdot \frac{1}{e^{\beta}+1} \end{aligned} \tag{21}$$

where the last equality uses that, conditional on SC, the recall weight is constant and carries no additional information about the value of s_{it}.

For $s_{it} \sim \mathcal{N}(\mu, \sigma_s^2)$, standard truncated normal formulas give:

$$\mathbb{E}[s_{it} \mid s_{it} \geq \mu] = \mu + \sigma_s \cdot \frac{\varphi(0)}{1-\Phi(0)} = \mu + 2\sigma_s\varphi(0) \tag{22}$$

$$\mathbb{E}[s_{it} \mid s_{it} < \mu] = \mu - \sigma_s \cdot \frac{\varphi(0)}{\Phi(0)} = \mu - 2\sigma_s\varphi(0) \tag{23}$$

where $\Phi(0) = \frac{1}{2}$ and $1 - \Phi(0) = \frac{1}{2}$ by symmetry.

Substituting into (21):

$$\begin{aligned}\mathbb{E}[s_{it} \mid \text{recalled}, g = M] &= \frac{e^{\beta}(\mu + 2\sigma_s\varphi(0)) + (\mu - 2\sigma_s\varphi(0))}{e^{\beta}+1} \\ &= \frac{\mu(e^{\beta}+1) + 2\sigma_s\varphi(0)(e^{\beta}-1)}{e^{\beta}+1} \\ &= \mu + \frac{e^{\beta}-1}{e^{\beta}+1} \cdot 2\sigma_s\varphi(0) \end{aligned} \tag{24}$$

For female students SC=1 corresponds to $s_{it} \leq \mu$, so the weight e^{β} applies to the lower tail. By the same argument:

$$\mathbb{E}[s_{it} \mid \text{recalled}, g = F] = \mu - \frac{e^{\beta}-1}{e^{\beta}+1} \cdot 2\sigma_s\varphi(0) \tag{25}$$

Since recalled signals are exchangeable—each s_{it} is drawn iid and recalled according to the same rule—the average recalled signal inherits the same expectation as any individual recalled signal:

$$\mathbb{E}\left[\bar{s}_i^R \,\middle|\, g\right] = \mu + \kappa_g \cdot \frac{e^{\beta}-1}{e^{\beta}+1} \cdot 2\sigma_s\varphi(0) \tag{26}$$

Setting $B \equiv \frac{e^{\beta}-1}{e^{\beta}+1} \cdot 2\sigma_s\varphi(0)$ completes the proof.

□

Discrimination and Memory Constraints

Definition 1 (Discrimination)**.** Discrimination is the expected assessment gap between a male and an otherwise identical female student:

$$\Delta \equiv \mathbb{E}[\hat{\theta}_i \mid g = M] - \mathbb{E}[\hat{\theta}_i \mid g = F] \tag{23}$$

Proposition 1 (Two-channel decomposition)**.** *Discrimination decomposes as:*

$$\Delta = \underbrace{(1 - \mathbb{E}[\omega_{K_i}]) \cdot 2\delta}_{C_1\text{: fall-back-on-prior}} + \underbrace{\mathbb{E}[\omega_{K_i}] \cdot 2B(\beta, \sigma_\theta, \tau_\varepsilon)}_{C_2\text{: biased signal recall}} \tag{24}$$

Both channels are non-negative for $\delta \geq 0$ and $\beta \geq 0$, and are separately identified by distinct empirical objects: C_1 by total recall volume (through ω_{K_i}) and C_2 by recall content (through B).

Proof. From the updating rule (13):

$$\mathbb{E}[\hat{\theta}_i \mid g] = (1 - \mathbb{E}[\omega_{K_i}])\, \mu_g + \mathbb{E}[\omega_{K_i}]\, \mathbb{E}\left[\bar{s}_i^R \,\middle|\, g\right] \tag{25}$$

Taking the difference across groups and applying Lemma 1:

$$\Delta = (1 - \mathbb{E}[\omega_{K_i}])(\mu_M - \mu_F) + \mathbb{E}[\omega_{K_i}]\left(\mathbb{E}\left[\bar{s}_i^R \,\middle|\, M\right] - \mathbb{E}\left[\bar{s}_i^R \,\middle|\, F\right]\right) = (1 - \mathbb{E}[\omega_{K_i}]) \cdot 2\delta + \mathbb{E}[\omega_{K_i}] \cdot 2B \tag{26}$$

□

Channel 1 (fall-back-on-prior). The weight $(1 - \omega_{K_i})$ on the group prior is large when the teacher recalls few signals. Since $\mu_M - \mu_F = 2\delta > 0$, defaulting to the prior generates discrimination proportional to stereotype strength. This channel is active even when $\beta = 0$: it requires only a biased prior and low recall volume. Empirically, C_1 is identified by the relationship between total recall volume and the recommendation gap (Figure **6**, panel a): teachers who recall fewer signals exhibit larger gender gaps, as assessed ability defaults more toward the biased group prior.

Channel 2 (biased signal recall). Even when the teacher recalls many signals (high ω_{K_i}), those signals are distorted in the stereotype-consistent direction when $\beta > 0$. This channel is active even when $\delta = 0$: it requires only biased retrieval, not a biased prior. Because the teacher applies Bayes' rule naively, the distorted average $\bar{s}_i^R$ enters the posterior directly and shifts assessments in a stereotype-consistent direction. Empirically, C_2 is identified by the gender gap in recall content (Figure **6**, panel b): teachers who evaluate female students disproportionately recall female-typed characteristics, and conditional on total recall volume, this compositional shift predicts lower science-track recommendations.

Proposition 2 (Memory constraints amplify discrimination)**.** *Under perfect memory ($\rho = 1$, $\beta = 0$):*

$$\Delta^{PM} = (1 - \omega^*) \cdot 2\delta, \quad \omega^* \equiv \frac{T\tau_\varepsilon}{\tau_a + T\tau_\varepsilon} \tag{27}$$

Under memory constraints ($\rho < 1$, $\beta > 0$):

$$\Delta > \Delta^{PM} \quad \textit{whenever } \delta > 0 \textit{ or } \beta > 0 \tag{28}$$

Proof. Under perfect memory, $K_i = T$ for all i, so $\mathbb{E}[\omega_{K_i}] = \omega^*$ and $B = 0$, giving $\Delta^{PM} = (1-\omega^*)\cdot 2\delta$. Under memory constraints, $\rho < 1$ implies $\mathbb{E}[K_i] = T\rho(S) < T$, so $\mathbb{E}[\omega_{K_i}] < \omega^*$. Computing the gap:

$$\Delta - \Delta^{PM} = \underbrace{(\omega^* - \mathbb{E}[\omega_{K_i}])\cdot 2\delta}_{\geq 0} + \underbrace{\mathbb{E}[\omega_{K_i}]\cdot 2B(\beta, \sigma_\theta, \tau_\varepsilon)}_{\geq 0} \tag{29}$$

The first term is strictly positive when $\delta > 0$, since $\omega^* > \mathbb{E}[\omega_{K_i}]$. The second term is strictly positive when $\beta > 0$, since $B > 0$ and $\mathbb{E}[\omega_{K_i}] > 0$. □

Proposition 2 maps directly onto the Memory versus Baseline experimental comparison. The Baseline condition approximates perfect memory ($\rho \approx 1$, $\beta \approx 0$), while the Memory condition imposes $\rho < 1$ and activates $\beta > 0$. The Memory × Female coefficient in Table **4** identifies $\Delta^{Mem} - \Delta^{Base}$.