

# Discrimination through Biased Memory\*

Francesca Miserocchi<sup>†</sup>

November 10, 2023

*[Click here for the most recent version](#)*

## Abstract

This paper shows that decision-makers make more stereotypical decisions when they struggle to recall individual-level information, penalizing women in male-dominated fields. Analyzing administrative data from Italian public schools, I find that when teachers need to assess a larger number of students, girls are less likely to be recommended for top-tier scientific high school tracks compared to boys with the same math standardized scores. Notably, this bias vanishes for teachers who report checking student data in class registers, relying less on memory alone. To directly assess the extent to which limitations and biases in recall generate more stereotypical decisions, I conducted two experiments. In the first, teachers provided track recommendations for a series of student profiles. When teachers cannot check individual data and must rely on memory, they recall a limited set of individual signals and disproportionately retrieve stereotype-consistent information. As a consequence, large gender-based disparities in track recommendations emerge, with girls 39% less likely to be recommended to STEM tracks than identical boys. Eliminating memory constraints by allowing teachers to check individual-level information reduces the gender gap by 80%, potentially mitigating the misallocation of talent. A second large-scale online experiment generalizes this mechanism. Taken together, the results highlight how memory limitations and biases amplify discriminatory behaviors and suggest that simple, cost-effective interventions facilitating access to individual-level information can mitigate such biases.

---

\*I thank the Unicredit Foundation and the Russell Sage Foundation for their funding of this project. I cannot thank Michela Carlana, Katie Coffmann, Ben Enke, Ed Glaeser, Larry Katz, and Amanda Pallais enough for their support and guidance in this project. I am also grateful to Spencer Kwon and Andrei Shleifer for providing essential feedback. This paper also benefited from comments from Chiara Aina, John Conlon, Antonio Coran, Ross Mattheis, Raphael Raux, Maya Roy, Awa Ambra Seck, Jesse Shapiro, Alice Wu, the participants in the labor workshop and seminar at Harvard, the SITE Experimental Conference, and the HEC Ph.D. student conference. AEA RCT numbers: AEARCTR-0008921 and AEARCTR-0010756. The project has obtained IRB approval from Harvard University.

<sup>†</sup>Harvard University. [fmiserocchi@g.harvard.edu](mailto:fmiserocchi@g.harvard.edu)

# 1 Introduction

Do decision-makers make more stereotypical decisions when they struggle to remember individual-level information about candidates, penalizing women in male-dominated fields? Memory limitations are widespread in high-stakes decision-making environments. Bail judges must make on-the-spot judgments to set bail conditions for multiple defendants (Arnold et al. (2018)). Employers must select candidates while juggling interviews and other tasks. Teachers often provide career advice to multiple students, drawing from accumulated information about their skills, attitudes, and interests. MBA instructors assess class participation based on semester-long observations. At the same time, a large body of research in psychology shows that people fail to recall information when they have many things on their mind (what is known as the interference theory of forgetting, Underwood (1957)).<sup>1</sup> While standard models of discrimination assume that decision-makers rely on observed information about candidates, they may recall a limited and biased set of information when they are cognitively overwhelmed.

To investigate whether biases in recall amplify biases in decisions, I focus on gender gaps in teachers' formal career advice to students about high school choice, in the context of Italian public middle schools. Despite the closing of the gender gap in overall university enrollment (Goldin et al. (2006)), women are still underrepresented in STEM fields of education in many countries (OECD (2022)).<sup>2</sup> I use administrative data from Italian public schools to test whether teachers make more stereotype-consistent decisions with respect to gender when experiencing higher mental burdens in a context with real and important consequences for students' careers. Then, I employ two experiments to provide direct evidence for a novel mechanism generating more stereotypical decisions. In the first experiment, middle school teachers provide track recommendations for a series of student profiles. When teachers cannot check student information and must rely on memory, they are 39% less likely to send girls to STEM tracks and recommend them to humanities, even though the students have identical capabilities and interests. By measuring the information recalled by teachers, I find that teachers recall a limited set of individual-level signals and disproportionately retrieve stereotype-consistent information (i.e. that the student is good at languages if she is a girl). Through a second experiment implemented in a large-scale online sample of US respondents,

---

<sup>1</sup>For a review of the interference literature, the reader is referred to Postman and Underwood (1973), Anderson and Neely (1996), Crowder (2014), Kahana (2012).

<sup>2</sup>While STEM is the predominant field of study for male graduates in 32 out of 37 OECD countries with data available, women are more likely to graduate from the fields of business, administration, education, and law (OECD (2022)). In the US, the male-to-female ratio among U.S. college majors in physics, engineering, and computer science is about 4-to-1 (Cimpian et al. (2020)).

I establish that this memory mechanism reflects a more general phenomenon occurring when decision-makers assess others' ability by retrieving information from memory. Together, the results indicate that memory constraints substantially amplify gender discrimination and that biases in recall drive biases in decisions.

I begin by developing a stylized model where a teacher assesses the probability that a student is good at math drawing from her memory database of experiences. Building on [Bordalo et al. \(2023\)](#), the belief formation process consists of two steps. Initially, the teacher retrieves both positive and negative math-related signals linked to the student. Then, she forms a belief about the student's ability by computing the ratio of recalled positive experiences to the total experiences retrieved. In recalling information, the teacher is susceptible to two biases. First, she tends to retrieve information aligning with gender stereotypes when she searches her memory. Second, thinking about a student triggers recall of information about same-gender students, which she uses as a proxy for the student's ability. The model delivers a set of predictions that I test in the data. Compared to a perfect memory benchmark, relying on memory leads to limited and biased recall of individual-level information, resulting in increased discrimination. This discrimination is driven by using memories of other same-gender students as proxies when struggling to recall individual information and by a biased recall of the individual student's signals.

I next turn to the empirical analysis and focus on high school track recommendations for 8th-grade students in Italy. Teachers face significant memory limitations at the time of the recommendation decision. They provide track recommendations for students in the same class in the same week and need to think about the skills and attitudes of many different students in a short time frame. Moreover, teachers may choose to check individual-level information about students in their notes and in the class register, or they may assess their students without reviewing information, relying only on what they recall at the time of the decision.

I report five key results. The first two results come from the analysis of teachers' actual track recommendations for their past students. For the first result, I use administrative data from Italian public schools on a sample of 8th-grade students from 2011 to 2015 matched with their math teachers. I compare girls and boys with similar objective math ability and assigned to the same math teacher in years in which she has different numbers of students to recommend, with controls for the total number of students assigned to the teacher to account for teachers' workload during the academic year. Most math teachers are assigned to one 8th-grade, one 7th-grade, and one 6th-grade class. Only 8th-grade students receive track recommendations, and track recommendations are usually assigned in December during a teaching meeting for students in the same class. Math teachers have a key role in assigning

track recommendations for the top-tier scientific high school tracks. Thus, the variation comes from within-teacher differences in 8th-grade class size and the number of 8th-grade classes across different years. The average number of 8th-grade students assigned to a math teacher is approximately 22, with the maximum being 48 (for teachers assigned to two 8th-grade classes).

Using this variation in the baseline specification with teacher and year fixed effects, the first result is that girls are less likely to be recommended for the top-tier scientific tracks than boys with the same standardized test scores in math and reading if their math teachers have a higher number of additional students to recommend. When teachers have a higher number of students, they are not generally more constrained, as they recommend a similar fraction of students to the scientific track (and there is no capacity limit on scientific track enrollment). However, boys are assigned to the math track instead of girls when teachers operate under a higher mental burden.

To ensure that this pattern is not driven by gender differences in student characteristics, I include a wide set of individual controls, controls for the quality of classmates, and their interaction by gender. Supply-side explanations such as girls participating less in larger classes are unlikely to explain the result as I obtain a qualitatively similar pattern when exploiting the variation in the number of students in other 8th-grade classes that the student's teacher has to recommend. Last, the pattern seems to be driven by a higher cognitive load specifically at the time of the recommendation decision rather than during the academic year (at the time of information acquisition). Only a higher number of 8th-grade students (who receive a recommendation) increases the gender gap in recommendations, not a greater number of students in 6th and 7th grade (who do not receive a recommendation).

Next, I implemented an original survey with approximately 600 middle school teachers and linked the survey teachers with administrative data on their past 8th-grade students in cohorts from 2016 to 2019. I compare students assigned to teachers who report that when assigning track recommendations they check their students' performance in the class register (relying less on memory) with those assigned to teachers who report that they rely on memory. The second result is that girls are less likely to be recommended to the scientific track than boys of similar math ability if their teachers report assigning recommendations by relying on their recollections of students' performance while there is no such gap for teachers who report checking students' performance in the register, relying less on memory alone.

In the second part of the paper, I implement two experiments to provide direct evidence that decision-makers make more stereotypical decisions when they struggle to recall individual-level information and to test for the existence and direction of biases in recall. In the first experiment, I ask middle school teachers to assign high school track recom-

recommendations to a series of hypothetical student profiles. The teachers observe four students' profiles describing the students' academic performance, interests, and other characteristics. After observing the profiles, they are asked to recall the performance and characteristics of each student and provide track recommendations. I randomize the gender of each profile (conveyed through the student's first name) and the teachers' memory capacity. Teachers in the *memory* condition cannot check the individual-level information when assigning recommendations and need to retrieve the characteristics of each student from memory. Since they observe the profiles of four students, signals about other students should compete for retrieval, imposing memory constraints. Teachers in the *baseline* condition are allowed to review the students' profiles before assigning recommendations, proxying a perfect memory benchmark.

I conducted the second experiment on a larger sample of US survey respondents recruited on Prolific. Participants are asked to evaluate the ability of a candidate in a male-typed domain (sports) and a female-typed domain (pop culture). Evaluators are asked to assess only one candidate, arguably minimizing the focus on gender. First, participants observe a candidate answering a set of trivia questions in sports and pop culture. The candidate correctly answers 50% of the questions in both domains. Then, they are asked to recall all the questions that their candidate answered correctly and incorrectly in the two domains and to estimate the candidate's ability in a new set of trivia questions. I randomize both candidate gender and decision-maker memory capacity similarly to in the first experiment. The controlled environment allows me to provide clear incentives for more accurate decisions.

The last three results are derived from the experiments. The first experimental result is that helping decision-makers recall all the available signals drastically reduces discrimination. When teachers cannot check the signals about the students but need to retrieve them from memory, female students are 39% less likely than identical male students to be recommended to scientific and technical tracks and are disproportionately recommended to humanities tracks. Allowing teachers to review each student's signals (proxying a perfect memory benchmark) reduces the gender gap by 80%. Likewise, in the Prolific experiment, the gender gap in each domain (favoring men in sports and women in pop culture) is 3 times larger in the memory condition than in the baseline condition where evaluators can review the information.

The second experimental result is on the direction of biases in recall. When teachers are prompted to think about a girl rather than an identical boy and do not check individual-level information, they selectively recall female-typed characteristics (e.g., "good at languages" if they observe a girl). Moreover, they recall a limited set of signals for each student. A similar result is confirmed in the second experiment. Conditional on observing identical performance

of a male or a female candidate, evaluators disproportionately recall stereotype-consistent signals: they recall a larger share of correct sports questions if their candidate is male and a larger share of correct pop culture questions if their candidate is female.

The third result from the experiments shows how biases in recall are associated with biased decisions. Conditional on recalling the same number of individual-level signals, the teachers who exhibit a higher gender bias in recall are those with a higher gender bias in students' track recommendations. In the second experiment, participants recalling a larger share of correct questions in the stereotype-consistent domain display more biased estimates of future performance.

The observational and experimental results indicate that decision-makers make more stereotypical decisions under greater mental burdens and that limitations and biases in recall act as a mechanism. Conditions imposing binding memory constraints, such as having more students on one's mind or not checking individual-level information, reduce recall and lead to a biased perception of signals, resulting in more stereotypical decisions. Bias can be reduced by helping decision-makers correctly recall individual-level information. In the observational data, gender bias in teachers' recommendations to the science track is not present in years when teachers have few students to recommend or if teachers consult the class register about students' performance. In the teachers' experiment, I similarly find that removing memory constraints eliminates approximately 80% of the gender gap in recommendations.

The paper is structured as follows. Section 2 outlines the contribution to the literature. Section 3 introduces a stylized model of discrimination based on recalled experiences. Section 4 describes the institutional setting. Sections 5 and 6 describe, respectively, the first and second results on gender gaps in track recommendations of teachers from the analysis of administrative data. In Section 7, I present the design and findings of the teacher experiment and, in Section 8, those of the Prolific experiment. Section 9 provides a discussion of the results and concludes.

## 2 Related Literature

This paper contributes to several lines of research. First, it relates to the literature on gender bias in subjective assessments and its consequences. [Benson et al. \(2021\)](#) find that subjective assessments of employees' potential contribute to gender gaps in promotions and pay. [Sarsons \(2017\)](#), [Sarsons et al. \(2021\)](#), and [Egan et al. \(2022\)](#) show that women receive worse evaluations for similar output and are penalized more for similar mistakes. Hiring methods involving higher degrees of employer discretion—such as interviews or oral tests—are associated with higher gender gaps ([Mocanu \(2022\)](#)). This paper builds upon this literature

highlighting how limited and biased recall of past information about candidates may act as a mechanism leading to gender bias in subjective assessments, especially if coupled with higher cognitive loads.

Second, this work speaks to the literature on discrimination and its sources.<sup>3</sup> The differential treatment of otherwise-identical candidates based on their group identity has been traditionally attributed to preferences (Becker (1957)) or beliefs (Phelps (1972), Arrow (1998), Bohren et al. (2019a), Bohren et al. (2019b)). Other works have explored nontraditional sources of discrimination, such as bias emerging from contrast effects (Kessler et al. (2022), Radbruch and Schiprowski (2021)) or from endogenous allocation of attention in selection decisions (Bartoš et al. (2016)). Tversky and Kahneman (1974) explain how higher mental burdens and quick decisions may lead decision-makers to rely on heuristics (i.e., simplified decision rules), and Bertrand et al. (2005) suggest that quick decisions may increase the role of implicit stereotypes, potentially leading to higher discrimination. Despite the wide descriptive evidence showing that fatigue and quick decisions are associated with more discriminatory behaviors, no prior studies in economics have provided experimental evidence that higher mental loads substantially exacerbate discriminatory behaviors in real-world decisions and connected it to information recall.<sup>4</sup> Regardless of whether mental loads are driven by making quick decisions, laboring under high workloads, or not checking information (and having many other things on one’s mind as a consequence), this paper suggests that limited and biased recall acts as a key mechanism.<sup>5</sup>

Third, this work relates to the literature on the consequences of teachers’ gender bias and on diversity in science (Carrell et al. (2010), Goldin (2014)). Teachers’ implicit bias affects students’ performance (Carlana (2019)), and boys choose more math and science-intensive secondary school tracks than girls of similar math ability (Buser et al. (2014)).<sup>6</sup>

---

<sup>3</sup>See Goldin and Rouse (2000), Bertrand and Mullainathan (2004), Aigner and Cain (1977), List (2004), Altonji and Pierret (2001), Fershtman and Gneezy (2001), Agan and Starr (2018), Doleac and Stein (2013), Kessler et al. (2019), and Kline et al. (2022). For a review of field studies on discrimination, see Bertrand and Duflo (2017).

<sup>4</sup>Research has shown that fatigue is associated with more discriminatory behaviors. For example, Kessler et al. (2019) find suggestive evidence of higher bias in employers who are more fatigued. Among employers evaluating CVs in these authors’ incentivized resume rating design, bias is higher for CVs evaluated at the end of each block, even though there are no significant effects for CVs evaluated in the second half of the study.

<sup>5</sup>Research in social psychology has studied recall of stereotypical and counterstereotypical information under different memory loads and was an inspiration for this research. Closely related to this work, Macrae et al. (1993) and Sherman and Frost (2000) show that subjects tend to retrieve stereotype-consistent information when they have higher cognitive loads. However, their studies employ smaller samples, are not in the context of gender, and do not investigate how selective recall leads to discrimination in subsequent real-world decisions. This paper extends this literature by providing large-scale laboratory and field evidence on how memory constraints exacerbate gender bias in real-world decisions.

<sup>6</sup>A large existing literature analyzes gender gaps in STEM fields – see Card and Payne (2021), Ceci et al.

Moss-Racusin et al. (2012) find that science faculty provide worse evaluations of female student profiles than of otherwise identical male profiles and that this could contribute to the gender disparity in science. This paper highlights how these disparities may emerge or increase because of evaluators’ limitations and biases in the recall of students’ signals.

I connect this broad discrimination literature with the growing literature in economics incorporating insights from psychology (Kahana (2012)) to study selective recall and memory-based belief formation. Mullainathan (2002), Bordalo et al. (2023), and Fudenberg et al. (2022) present models of how selective recall of past experiences shapes belief formation.<sup>7</sup> Esponda et al. (2023) implement a related experiment in an abstract setting and show that decision-makers distort their evaluation of new evidence in the direction of group stereotypes. I show that these memory biases affect real-world decisions.

### 3 Model

This section presents a stylized model of discrimination where decision-makers (DMs) draw on their memory database of experiences to form beliefs. We consider a teacher who needs to assess the probability that student  $i$  of gender  $g \in \{m, f\}$  is good at math. The belief formation process develops in two stages, similarly to in Bordalo et al. (2023) and Bordalo et al. (2022), but is adapted to describe discrimination. The decision-maker draws on her memory database looking for relevant experiences to estimate the probability that student  $i$  of gender  $g$  is good at math (which I denote as  $\hat{\pi}(i|g)$ ). Then, she uses the recalled experiences to form a belief.

This process can be affected by two biases. First, gender associations can direct the memory search, making it easier to retrieve stereotype-consistent information. Second, when counting positive and negative math experiences, the DM uses experiences with other, similar students to proxy the ability of student  $i$ . If the first bias is not in place, this framework delivers predictions similar to those derived from a belief-based discrimination framework where the DM uses information about other same-gender students to approximate individual ability and the DM’s perception of the individual signals is unbiased. If the second bias is not in place (the DM does not use other students as a proxy), discrimination arises only because the DM has a biased perception of the signals due to her biases in recall.

---

(2009), Hill et al. (2010), Ceci et al. (2014), Wiswall et al. (2014), and Kahn and Ginther (2017) for a review.

<sup>7</sup>These models have been validated with laboratory experiments on recall (Bordalo (2021a,b), Enke et al. (2020)). Moreover, recent works by Jiang et al. (2022), Charles (2022), and Andre et al. (2022) provide evidence on the role of memory in real-world decisions in the context of financial markets and for subjective assessment of the macroeconomy.



### 3.1 Setup

Table A1 summarizes the types of signals in the teacher’s memory database. There are  $N + 1$  students: student  $i$ ,  $N/2$  other girls, and  $N/2$  other boys. The teacher experiences  $K$  signals about each student in math. Thus, the memory database  $E$  is composed of  $K(N + 1)$  experiences. It is partitioned in the sets of positive and negative math signals about student  $i$  (respectively, the sets  $H_i$  and  $L_i$ ), other girls (the sets  $H_f$  and  $L_f$ ), and other boys (the sets  $H_m$  and  $L_m$ ). The set of total math signals about student  $i$  is  $I = H_i \cup L_i$ . Similarly,  $F = H_f \cup L_f$  is the set of math signals for other girls, and  $M = H_m \cup L_m$  is the set of signals for other boys.

Last, we define the true ratio of positive math signals for student  $i$ , other girls, and other boys as  $h_i \in [0, 1]$ ,  $h_f \in [0, 1]$ ,  $h_m \in [0, 1]$  respectively.  $h_i$  measures the observed ratio of positive math signals about the student  $i$  and is equal across gender ( $h_i = \frac{|H_i|}{|H_i| + |L_i|}$ ). We assume that  $h_m \geq h_f$ , i.e., that the teacher is exposed to signals (through her experiences, news, second-hand information, etc.) implying that boys are better in math than girls. This difference could be driven by past institutional discrimination (for instance, the teacher knows that most Nobel laureates in scientific subjects are male), implicit stereotypes affecting girls’ performance, or true differences in average ability.

We assume that  $h_i = 1/2$  (as in the experiment on Prolific, where the candidate answers 50% questions correctly) and that the ability of girls and boys is symmetric,  $h_m = (1 - h_f)$ .<sup>8</sup> These assumptions are useful to simplify the algebra and ensure that the total number of recalled signals about  $i$  does not depend on gender (only the type of signals recalled depends on gender), which is empirically observed in the experiments (Figure A31 and A43).<sup>9</sup>

### 3.2 Belief on Ability

The DM counts the fraction of positive math signals over the total math signals imputed to student  $i$ . We call  $R(H_i|f)$  the number of recalled positive math signals about  $i$  if student  $i$  is a girl and  $R(H_g|f)$  the number of recalled positive math signals about other girls (if  $H_g = H_f$ ) or other boys (if  $H_g = H_m$ ) if student  $i$  is a girl.  $R(I|f)$ ,  $R(F|f)$ ,  $R(M|f)$  are the total number of recalled signals (positive and negative) regarding student  $i$ , other girls, and other boys respectively. If recall was perfect, after observing signals in  $H_i$  (positive math signals of student  $i$ ), the number of positive signals recalled would be  $R(H_i|f) = |H_i|$ .

---

<sup>8</sup>This assumption means that the gender gap must be centered around 50%, for instance, if the gender gap is 40 percentage points we assume that girls answer 30% questions correctly and boys 70% (rather than 10-50).

<sup>9</sup>This ensures that the weight  $\theta(N)$  does not depend on the gender of student  $i$ , but only on the mental burden.

If student  $i$  is a girl, the believed probability that student  $i$  is good at math is:

$$\hat{\pi}(i|f) = \frac{\overbrace{R(H_i|f)}^{\text{Number of recalled positive signals of } i} + \sigma_1 \overbrace{R(H_f|f)}^{\text{Number of recalled positive signals about other girls}} + \sigma_2 \overbrace{R(H_m|f)}^{\text{Number of recalled positive signals about other boys}}}{R(I|f) + \sigma_1 R(F|f) + \sigma_2 R(M|f)} \quad (1)$$

Consider a female student named Susan. Assessment of the individual student is based on the number of recalled successes of Susan and students similar to Susan over the number of recalled successes and failures.  $\sigma_1 \in [0, 1]$  measures the fraction of signals of other girls imputed to Susan, while  $\sigma_2 \in [0, 1]$  measures the fraction of signals of boys imputed to Susan. We assume that  $\sigma_1 > \sigma_2$ , i.e. Susan is proxied with other girls more than with boys, as same-gender students are more similar to Susan.<sup>10</sup>

If recall of individual-level information is perfect (i.e.,  $R(H_i|f) = |H_i|$ ,  $R(L_i|f) = |L_i|$ ) and the DM does not proxy Susan with other students (i.e.,  $\sigma_1 = 0$  and  $\sigma_2 = 0$ ), then the DM assessment of Susan's ability equals the true observed ability of Susan ( $h_i$ ), and the assessment of the individual student would not depend on her gender (the assessment is  $h_i$  regardless of whether student  $i$  is Susan or John).

Biases arise for two reasons. First, a gender bias arises if  $\sigma_1 > 0$ , since the individual student is disproportionately proxied with other same-gender students and on average boys perform better at math than other girls. Bias is higher the higher is the difference between  $\sigma_1$  and  $\sigma_2$  (the more student  $i$  is proxied with same gender rather than opposite gender students). Second, biased assessments emerge if there are biases in which signals are recalled (e.g.,  $R(H_i|f) \neq |H_i|$ , etc.), which may affect both memories of student  $i$ 's signals and memories of other students' signals (which form the "prior").

### 3.3 Recalling Experiences

Recall of experiences is characterized by two properties. First, an experience  $e$  is more likely to be recalled if it is characteristics of the student's gender  $g$  (if it is more common among students of gender  $g$  than  $\bar{g}$ ). This notion is supported by research in psychology (Macrae

---

<sup>10</sup>One interesting implication of this could be that  $\sigma_1$  may depend on a girl's physical appearance being more or less similar to other girls'; it may be thus optimal for a girl to dress and behave more similarly to boys to be perceived as more competent in math. Moreover, Susan may be proxied with other same-gender students both rationally (as a form of statistical thinking, as in statistical discrimination models) or irrationally (by recalling signals of other girls and mistakenly attributing them to Susan). I do not aim to distinguish these two cases and posit only that signals of other students can be used to approximate Susan's ability, either rationally or irrationally.

et al. (1993), Sherman and Frost (2000)), finding higher recall for stereotype-consistent information under high memory loads. For instance, in one study, participants read information about a subject and are later asked to report what information they remember. When the memory load is higher, they disproportionately recall behaviors pretested to be "kind" if they were told that the subject is a priest rather than a skinhead.<sup>11</sup>

Second, an experience  $e$  is less likely to be recalled if many other experiences compete for recall, i.e., if the DM has many other things on her mind. This property reflects interference (Underwood (1957)), whereby the DM is less likely to recall an experience if she has many other, similar experiences in her mind.

We define a similarity function  $S(e|g)$  measuring whether experience  $e$  is consistent or inconsistent with the gender stereotype "*boys good at math, girls bad at math*".<sup>12</sup> The similarity of an experience  $e$  is equal to 1 if the experience is consistent with the stereotype. If student  $i$  is a girl,  $S(e|f) = 1$  if  $e \in \{L_i, H_m, L_f\}$  (negative math experiences about girls are consistent, positive are inconsistent), while  $S(e|f) = 1 - \Delta_s$  if  $e \in \{H_i, L_m, H_f\}$ . The probability that positive math experiences about Susan are recalled is:

$$r(H_i|f) = f(\underbrace{S(H_i|f)}_+, \underbrace{N}_-) \quad (2)$$

The probability of recalling positive math experiences is higher if they are consistent with the gender stereotype "boys good at math, girls bad at math" and is lower if the teacher has many other students in her mind (if  $N$  is higher). Additional details about the recall function can be found in Appendix A.1. In Appendix A.2, I consider a more general similarity function in which experiences have additional features.

**Assessment of ability.** The average assessment of ability for student  $i$  of gender  $g$  is:

$$\hat{\pi}(i|g) = \theta(N) \underbrace{\hat{p}(i|g)}_{\text{Perception of Signals}} + (1 - \theta(N)) \underbrace{\hat{p}(g)}_{\text{Belief on similar students (prior)}} \quad (3)$$

<sup>11</sup>In these studies, as a memory load subjects memorized 10-digit numbers before the recall task. Disproportionate recall of stereotype-consistent information can be attributed to the associative nature of memory and schema effects: the student's gender may guide the memory search toward experiences more characteristic of that gender (more common among students of gender  $g$  than  $\bar{g}$ ), which fulfill expectations.

<sup>12</sup>Here, I consider a simplified similarity function to isolate the mechanisms that I test in the empirical analysis. I consider a more standard similarity function in Section A.2 of the Appendix, in which I account for similarity with two other features: whether the experience is of student  $i$ , and whether the experience is of same-gender students.

where  $\hat{p}(i|g)$  is the fraction of recalled positive signals about student  $i$ ,  $\hat{p}(g)$  is the fraction of recalled signals of other similar students, and the weight  $\theta(N)$  depends on the number of other students. The derivation of equation 3 can be found in Appendix A.1. A higher number of other students inhibits the recall of signals about student  $i$  and leads the DM to rely more on other, similar students as a proxy. Biased assessments of ability arise both from biases in the perception of the observed information and from proxying student  $i$  with same-gender students.

**Discrimination.** Discrimination is defined as the difference in assessment for student  $i$  with ability  $h_i$  depending on her gender:

$$D(N) = \hat{\pi}(i|m) - \hat{\pi}(i|f) = \underbrace{\theta(N)(\hat{p}(i|m) - \hat{p}(i|f))}_{\text{Biased perception of signals}} + \underbrace{(1 - \theta(N))(\hat{p}(m) - \hat{p}(f))}_{\text{Gender gap in prior}} \quad (4)$$

If recall is not biased by stereotypes,  $\hat{p}(i|m) = \hat{p}(i|f)$ . In this case, we observe discrimination only if there are differences in the population, i.e.,  $h_m > h_f$ , and if the DM uses other students to approximate student  $i$ 's ability (the same prediction as that corresponding to statistical discrimination). If stereotypes affect what the DM recalls ( $\Delta_s \neq 0$ ), discrimination comes both from proxying the individual students with other students (biases in the prior) and from a biased perception of student  $i$ 's signals. With respect to existing models of belief-based discrimination (for example, [Bohren et al. \(2019b\)](#)), there are two differences. First, I hypothesize that beliefs are formed from previous experiences. Second, discrimination arises both from biases in the prior and from biases in the perception of individual-level signals.

**Perfect memory benchmark.** In the perfect memory benchmark, the DM is given a memory aid that helps her correctly recall the signals of student  $i$ . In this case,  $\bar{\theta}(N) > \theta(N)$  for every  $N$ ; i.e., the DM relies more on the signals since she recalls a higher number of signals. In this case, recall of signals is not biased, and discrimination depends only on pre-existing biases in prior beliefs:

$$D(N)^{PM} = (1 - \bar{\theta}(N))(\hat{p}(m) - \hat{p}(f)) \quad (5)$$

### 3.4 Predictions

This stylized model delivers three predictions that I test in the data:

1. *With respect to the perfect memory benchmark (memory aid to recall signals of student  $i$ ), discrimination is higher because the DM relies more on her prior and has a biased*

recall of signals:

$$D(N) - D(N)^{PM} = \underbrace{\theta(N)[\hat{p}(i|m) - \hat{p}(i|f)]}_{\text{Biased recall of signals}} + \underbrace{(\bar{\theta}(N) - \theta(N))[\hat{p}(m) - \hat{p}(f)]}_{\text{Relying more on prior}}$$

2. *Discrimination is higher when the teacher has many other students to recommend because she fails to recall individual-level signals and relies more on experiences with other students:  $\frac{\partial D(N)}{\partial N} > 0$ .*
3. *Conditional on observing the same information, the DM recalls a higher share of positive signals (higher recalled success ratio) if she observes a boy than if she observes a girl and assesses student  $i$  in a male-typed domain:  $\hat{p}(i|m) > \hat{p}(i|f)$  if  $\Delta_s > 0$ .*

## 4 Institutional Setting

In the Italian schooling system (similarly to many other systems characterized by early tracking), students choose their high school track at the end of grade 8. There are three main types of high school: academic, technical, and vocational. The top-tier academic tracks are composed of the scientific and classical tracks (I will refer to the first group as the "top-scientific" or simply "scientific" track, and the second group as the "top-classical" or "top-humanities" track throughout the analysis).<sup>13</sup> The top-scientific track's curriculum is predominantly focused on scientific subjects (math, physics, biology, chemistry), while the top-classical curriculum has fewer hours of instruction in scientific subjects and includes the study of Greek and Latin. Once students choose a track, they likely have few interactions with students attending other tracks, as classes in different tracks are usually located in different buildings and neighborhoods.

Italy is a country characterized by large and persistent gender gaps both in labor market outcomes and STEM education. In 2022, a lower percentage of women graduated in STEM fields (40.9% women compared to 59.1% men), despite there being no differences in graduation grades (AlmaLaurea (2023)). The high school track choice is the first important career decision in the Italian schooling system with long-term consequences for the subsequent field of study choice in university and for students' future earnings trajectories and gender gaps. While the majority of graduates in STEM fields attend top scientific high school tracks, those who attend the top-classical track in high school tend to graduate in

<sup>13</sup>Apart from the top-scientific and top-classical tracks, we group the other tracks as follows. There are other academically oriented but less demanding tracks that we group as medium-humanities (linguistic, humanities, artistic track); technical, technological, or economic track (technical); and vocational track.

humanities-related fields.<sup>14</sup>

Teachers provide formal career advice to 8th-grade students. In particular, they send a formal recommendation letter to the students' families with their suggested high school tracks. To assign track recommendations, teachers hold a class-specific teaching meeting in December, in which they assign track recommendations for each student in the class.<sup>15</sup> After sending the official track recommendations by mail, they usually meet the students' parents to communicate the reasons for their recommendations.<sup>16</sup>

Although students can choose against their teachers' track recommendations, track recommendations are important for students' choices. Approximately 80% of students recommended to enter the top-scientific track effectively choose that track, while only 20% of students recommended for the top-classical track end up choosing the scientific track.<sup>17</sup> If we control for students' ability in math and Italian and for their parents' education, being recommended to the scientific track increases the likelihood of choosing it by more than 50 percentage points (pp).<sup>18</sup>

Teachers have a considerable workload during the recommendation period. Class meetings to assign recommendations are usually held on the same week in each school. Each class is composed of 22 students on average, but class sizes vary from approximately 12 to 30 students. Moreover, math and literature teachers are usually assigned to one 8th-grade, one 7th-grade, and one 6th-grade class per year, but some of them may be assigned to two 8th-grade classes to accommodate their school's needs.

---

<sup>14</sup>For example, approximately 68% of graduates in industrial or IT engineering attended a top-scientific high school track, while only 12% of graduates in education-related fields and 14% of graduates in language-related fields attended a top-scientific high school track. Of graduates in humanities-related fields, 37% attended a top-classical track, while only 3% of graduates in IT engineering attended a top-classical high school track [AlmaLaurea \(2021\)](#).

<sup>15</sup>In principle, teachers can recommend multiple tracks to the same student, even though they are required to indicate either one specific track or "all tracks" as their first choice. In practice, teachers tend to recommend the school track that they consider the best fit for each student, and few students are recommended for more than one track. In the sample of teachers who completed the survey, approximately 6% of students received more than one recommendation, and 1.1% of students received "any choice" as a track recommendation.

<sup>16</sup>While all teachers can express their views during the teaching meeting, math, and Italian teachers are particularly involved in deciding track recommendations, as they spend more time with the students than teachers of other subjects (e.g., English or another language, gym, music, arts, technology). In particular, math teachers have a key role in assigning track recommendations for the scientific tracks.

<sup>17</sup>This statistic is computed based on the sample of survey teachers described in section 6, for which I observe recommendations and choices for all tracks.

<sup>18</sup>This computation is performed with data for teachers in the main sample, described in section 5. The regression results on the correlation between top-scientific track recommendation and choice, with controls for students' math ability and parental background, can be found in Table A4.

## 5 First Result on Track Recommendations for Past Students

In this section, I examine whether math teachers make more stereotypical recommendation decisions when they have more 8th-grade students to recommend. According to the stylized model in section 3, teachers recall fewer individual-level signals and disproportionately recall stereotype-consistent signals when they have many things on their minds, resulting in biased decisions.

### 5.1 Data and Summary Statistics

The main sample includes 8th-grade students in cohorts between academic years 2011-12 and 2015-16 in 92 Italian public middle schools assigned to their math teachers, as in [Carlana \(2019\)](#).<sup>19</sup> Crucially, I observe students' standardized test scores in math and Italian, which allows me to compare students with the same objective ability. Moreover, the sample includes several cohorts of students assigned to the same math teachers in different years, allowing me to compare similar students assigned to the same teachers in years in which they have many or fewer other students to recommend. Since I exploit within-teacher variation across years, I restrict the sample by excluding the math teachers that I observe for only one year. The final sample is composed of approximately 16,500 8th-grade students matched with their 316 math teachers.

Table A2 reports the summary statistics of students and their math teachers. Approximately 32% of students choose the top-tier scientific or classical tracks (27% choose the top-scientific and 5% the top-classical tracks). The top-performing students in math and Italian tend to choose the top-tier scientific or classical tracks (the average grade of students choosing these tracks is 8/10 in math and Italian).<sup>20</sup> When we focus on the students' math teachers, 79% are female, they are 47 years old on average, and 80% have a permanent contract.

---

<sup>19</sup>The initial sample also includes academic year 2016-17, but information on track recommendation is not available for that year, as explained in [Carlana \(2019\)](#).

<sup>20</sup>Students choosing the scientific tracks have an average math grade of 8/10 and an average literature grade of 7.8/10, and students choosing the top-classical track have an average math grade of 7.8/10 and an average literature grade of 8.3/10. Students choosing the medium humanities tracks have an average math grade of 7 and an average Italian grade of 7.4. Students choosing technical tracks have an average math grade of 6.8 and an Italian grade of 6.7. Students choosing the vocational tracks have an average math grade of 6.3 and an average Italian grade of 6.4.

**Gender gaps in recommendations to scientific track.** Girls are 3.3 pp (16%) less likely to be recommended to the top scientific high school track than boys. Figure A6 shows how the gender gap in scientific track recommendations varies with the inclusion of controls. Once I account for students’ standardized math and Italian test scores, students’ background, and differences between teachers and years by including teacher and year fixed effects, the gender gap goes down to 2.4 pp (12%). Even in the 21st century and in a high-stakes setting for students, young girls with equivalent test scores and backgrounds are 12% less likely than boys to be recommended for the scientific track. These gender gaps in track recommendations are likely to amplify gaps in students’ choices since girls are not less likely to follow the scientific track recommendations than boys, as shown in Table A4.

**Identifying variation.** In most cases, teachers are assigned to three classes per year (one 6th-, one 7th-, and one 8th-grade class). However, it is sometimes the case that teachers are assigned to two 8th-grade classes in one year to accommodate the school’s needs: in our sample, 12% of math teachers have more than one 8th-grade class for at least one year. Classes are composed of 22 students on average, but they can vary in size from 12 to 29 students. I exploit the within-teacher variation in the number of 8th-grade students to be recommended in different years, which comes both from variation in the teacher’s 8th-grade class sizes across years and from the number of students in other 8th-grade classes that the teacher is assigned to. On average, the within-teacher variation in the number of students to be recommended across years is approximately 6 additional students, with the maximum variation being an additional 31 students to recommend for math teachers with two 8th-grade classes (Figure A1 shows the minimum and maximum number of students to be recommended by math teachers across years).

## 5.2 Empirical Strategy

I compare the track recommendations for girls and boys of similar objective ability assigned to the same teachers in years in which the teacher has many vs. fewer other 8th-grade students to recommend. In particular, I estimate the following equation at the student level:

$$Y_{ijt} = \alpha_0 + \sum_{s=2}^4 \beta_s 1(\#Students \text{ Bin}_{jt} = s) + \sum_{s=2}^4 \gamma_s 1(\#Students \text{ Bin}_{jt} = s) \cdot Female_i + \quad (6)$$

$$+ \beta_1 Female_i + \mu_t + \nu_j + \delta_0 \mathbf{X}_{it} + \delta_1 \mathbf{Z}_{it} + \delta_2 \#Students \text{ Tot}_{.jt} + \varepsilon_{ijt}$$

where  $Y_{ijt}$  is a dummy equal to 1 if student  $i$  assigned to math teacher  $j$  in year  $t$  is



recommended to the scientific track,  $\text{Female}_i$  is a dummy equal to 1 if the student is female,  $\nu_j$  is a vector of teacher fixed effects to account for teacher-specific characteristics affecting recommendations that do not vary over time, and  $\mu_t$  is a vector of year fixed effects. I divide the total number of other 8th-grade students that the students' teachers have to recommend into 4 bins (fewer than 19, 20–24, 25–29, and more than 30 students).  $1(\#\text{Students Bin}_{jt} = s)$  are dummies equal to 1 if teacher  $j$ 's number of students to recommend in year  $t$  is in bin  $s$ . Standard errors are clustered at the teacher level.<sup>21</sup>

The baseline specification includes standardized test scores in math and Italian ( $\mathbf{Z}_{it}$ ) and a baseline set of student controls  $\mathbf{X}_{it}$ : mother's education level; whether the father works in a high-, medium-, or low-income occupation; whether the student is a first- or second-generation immigrant, and the number of years that the student spent with the math teacher (whether the student was assigned to the math teacher for 1, 2, or 3 years), and average classmates characteristics (share of girls, share of immigrants, share of classmates with highly educated mother, average classmates' math and reading scores).<sup>22</sup> Moreover,  $\#\text{Students Tot.}_{jt}$  controls for the total number of students assigned to teacher  $j$  in year  $t$  (including students in 6th and 7th grades).

I also estimate the following, more compact specification, with the stronger assumption that a higher number of students to be recommended impacts gender gaps in track recommendations in a linear way:

$$Y_{ijt} = \alpha_0 + \beta \#\text{Students}_{jt} + \gamma \#\text{Students}_{jt} \cdot \text{Female}_i + \alpha_2 \text{Female}_i + \mu_t + \nu_j + \delta_0 \mathbf{X}_{it} + \delta_1 \mathbf{Z}_{it} + \delta_2 \#\text{Students Tot.}_{jt} + \varepsilon_{ijt} \quad (7)$$

where  $\#\text{Students}_{jt}$  is the number of 8th-grade students that teacher  $j$  has to recommend in year  $t$  and the other variables are defined as before. Our coefficient of interest  $\gamma$  ( $\gamma_s$  in equation 6) identifies the increase in the gender gap in the probability of being recommended to the scientific track if the teacher has one additional student to recommend in a given year (one standard deviation is approximately 6 additional students). In order to interpret  $\gamma$  as the effect of an increase in teachers' cognitive load at the time of the recommendation decision, the following assumption needs to be satisfied:

**Assumption 1.** *Conditional on the total number of students assigned to the teacher, the full set of controls, and fixed effects, a higher number of 8th-grade students affects gender*

<sup>21</sup>Clustering the standard errors at the teacher-by-year level leads to similar results.

<sup>22</sup>Most math teachers teach the same cohort of students for 3 years in middle school. However, in some cases, students change teachers over the years.

*gaps in recommendations only by imposing a higher cognitive load on teachers.*

**Potential violations of the exclusion restriction.** A first potential violation would occur if variations in class size have a differential effect on class participation by gender. If girls participate less in larger classes, their teachers may have fewer signals about them. We address this concern by estimating equation 7 with controls for class size and class size interacted by gender (column 7 of Table A5) and exploiting only the variation coming from the number of students in other 8th-grade classes. Table A8 shows that having more students in "other" 8th-grade classes also increases the gender gap in scientific track recommendations, suggesting that the effect is not explained by supply side explanations such as girls' lower participation in larger classes.

Another concern is that teachers may recommend a fixed number of 8th-grade students to the scientific track that differs by gender, even though there is no formal capacity limit. In this case, the gender gap would increase with the number of students to be recommended because teachers are capacity constrained and not because of higher cognitive load when they assign recommendations. However, a student's probability of being recommended to the scientific track does not decrease when the teacher has more 8th-grade students, as seen in column 1 of Table 1 and in Table A9. When they have more students, teachers send a similar fraction of students to the scientific track but replace girls with boys.<sup>23</sup>

A third potential violation is that teachers' higher workload during the academic year—not at the time of providing the recommendations—may drive the results.<sup>24</sup> However, I find that what matters is 8th-grade students, not the number of students in lower grades (Table A10). Teachers decide on recommendations in early December before they do the semester grading for everyone. The fact that a greater number of students to be recommended rather than a greater number of students during the academic year drives the result suggests that differential information acquisition during the academic year is unlikely to explain the result.

**Balance on observables.** I proceed by testing whether students are systematically different when their math teachers have many other students to recommend, estimating model 6 with students' characteristics as dependent variables. Figure A2, Figure A3 and Panels (b) and (c) of Figure 1 show that the number of other 8th-grade students whom the math teacher

---

<sup>23</sup>I also test if having a higher standardized score in Italian decreases the likelihood of being recommended to the scientific track when the teacher has more students to evaluate, but I find no evidence of this either (Table A9).

<sup>24</sup>Teachers' higher workloads during the academic year may lead them to differentially focus their attention on students of one gender (similarly to the attention discrimination mechanism à la Bartoš et al. (2016)) or may lead them to acquire fewer signals about all students.

needs to recommend does not seem to predict many observable student characteristics.<sup>25</sup> I observe some imbalances in the last number of students bin – even though these differences are not statistically significant.<sup>26</sup> Teachers in years with higher and lower student evaluation loads have similar characteristics, as shown in Figure A4.

### 5.3 Results

Estimates of the coefficients  $\gamma_q$  from equation 6 represent the differential impact of the number of other students to be recommended by their math teacher on the probability that girls relative to similar boys are recommended for the scientific track, with respect to the gender gap in years of low evaluation load (in the first number of students bin). These estimates, shown in the right graph of Panel (a) in Figure 1, indicate that gender gaps in scientific track recommendations expand when math teachers have many other students to recommend. These gender gaps are not present when math teachers have few other students to recommend, as one can see from the left graph in Panel (a). Panels (b) and (c) show that when their math teachers have more students to recommend, the gender gaps in math and reading ability (the standardized test score is performed at the end of the academic year) do not significantly increase, suggesting that the expanding gender gap in scientific track recommendations is unlikely to be driven by supply-side factors differentially affecting students' ability by gender. If I focus on the very good students in math (in the 10th decile of math ability), I obtain a qualitatively similar pattern (Figure A7), suggesting that even very talented girls in math may be subject to this bias.

The results are summarized in Table 1, where I estimate equation 6. The initial raw gender gap is 16.3% of the boys' mean. Once I linearly control for standardized math and Italian test scores and I include teacher and year-fixed effects in column 1, the adjusted gender gap becomes on average -2.3 pp (10%). In column 3, I include the gender interaction with my proxy of teachers' workload: when math teachers have a higher number of students to recommend, the adjusted gender gap in track recommendations to the scientific track increases (Table A5 estimates the linear model). The rate at which the gender gap increases decreases with the number of other students: having 5 additional students matters more when

---

<sup>25</sup>Both female and male students evaluated in years with higher evaluation loads seem to have similar standardized test scores in math and reading, similar socioeconomic backgrounds, and classmates of similar average quality (average share of same gender classmates, share of immigrants, average math and Italian ability of classmates).

<sup>26</sup>Girls whose math teachers have more than 30 students to recommend (in the last number of students bin) have slightly lower math scores and are less likely to be immigrants compared to boys. I control for standardized scores in math and reading, squared standardized scores, for students' and students' classmates' characteristics in the analysis.

the teacher starts out with 10-19 students than with 25-29 or more students. In column 4, I include students' controls and controls for the quality of classmates, and in column 5, I control for squared standardized test scores. In column 6, I interact the student controls with student gender, accounting for the fact that student characteristics may be evaluated differently depending on student gender. Having 10 additional students to be recommended still increases the gender gap by around 15% of the boys' mean.

**Where are the "missing" science-track girls sent?** In the main observational dataset, I have only information on recommendations for scientific and vocational tracks; thus, I cannot test whether the missing science-track girls are sent to humanities tracks in this dataset. However, I can use the dataset of survey teachers matched with their past students to assess whether I find a symmetric effect for the top-classical track. Since this dataset has fewer teachers in core subjects, I use it to supplement this part of the analysis. Table A14 shows how the gender gap in scientific track and classical track recommendations evolves with the number of students to be recommended, with controls for standardized test scores and teacher and year fixed effects. When their teachers have more other students to recommend, girls are less likely to be sent to the top-scientific track and more likely to be sent to the top-classical track.

**Heterogeneity, sensitivity, and robustness.** Which types of students are discriminated against when their math teachers have a higher cognitive load? Figure A8 shows the gender gap in recommendations to the top-scientific track by evaluation load and students' math ability. Students from the 7th and 8th deciles of math ability are the most affected, but those in the top deciles are also affected. If I further split the students by their Italian ability (Figures A9), I find suggestive evidence that the most impacted students are those good in both math and Italian (students with "mixed" signals).

Last, I perform the following series of sensitivity exercises. I assess that the result is robust to dropping one school and one year at a time and restricting the range of 8th-grade students. Moreover, I check that other relevant counterfactual decision rules would not lead to the same pattern (Table A7).<sup>27</sup>

---

<sup>27</sup>For each decision rule, I assign the corresponding track recommendation to students and estimate equation 7. I analyze the following decision rules, for which I estimate small and insignificant effects: a student is recommended to the scientific track if (i) she is in deciles 9–10 of math ability, (ii) she is among the top 30% of students in the class, (iii) she is among the best X% of students according to her teacher-specific leniency (average fraction of students whom each teacher sends to the scientific track). The thresholds are chosen to be close to the average true fraction of scientific track recommendations, which is 20%.

## 6 Second Result on Track Recommendations for Past Students

In this section, I test another factor imposing memory limitations: teachers may discriminate more when they think about their students *off the cuff*, i.e., by retrieving individual-level signals from memory rather than checking their students' performance in the class register (relying less on memory). A similar mechanism may be at play: when they do not refer to class registers for students' information but retrieve individual-level signals from memory, teachers may recall a limited and biased set of signals and be more likely to discriminate against girls in math.<sup>28</sup>

### 6.1 Data and Summary Statistics

Between February and March 2023, I administered an original survey to approximately 600 Italian public middle school teachers to further investigate how the recommendation process works and the mechanisms driving gender gaps.<sup>29</sup> Teachers participating in the survey first took part in the experiment described in section 7. Then, they were asked to provide additional details on the recommendation process. In particular, they were asked to list which actions they take when deciding the track recommendations (choosing from a predetermined set of actions) and to provide more details on the recommendation process through an additional set of questions and open-ended questions.

The surveyed teachers are matched with administrative data on their past 8th-grade students who completed middle school between 2016–17 and 2019–20. Out of 609 survey teachers with an 8th-grade class, 473 had an 8th-grade class between 2016 to 2019, and approximately 240 of them teach core subjects (math and Italian) and are thus primarily involved in the track recommendation decision. The detailed administrative data on their past students contain information on track recommendations and choice, standardized test scores, teacher-assigned grades and GPA, and student demographic characteristics.

Table A12 and Figure A11 compare survey teachers with math teachers in the main administrative sample. With respect to math teachers in the main administrative sample, teachers participating in the survey are on average more likely to be female, to have a full-time contract, and to be slightly older. The surveyed teachers are slightly less likely to

---

<sup>28</sup>This is also consistent with interference theory (Underwood (1957)). When they retrieve signals from memory, they automatically have many other things on their mind, which inhibits recall. The comparison group in which they check information is an "extreme" scenario proxying a perfect memory benchmark.

<sup>29</sup>The surveyed teachers had been previously recruited for a project on socioeconomic status and track choice.

be born in Northern regions, are less lenient in their recommendations to the top-scientific track, and have larger gender gaps. While all schools in the main observational sample are in the North, 27% of the surveyed teachers' schools are in the Center or South (Figure A12 shows a map of the surveyed teachers' schools). On average, approximately 50% of the teachers participating in the survey teach a humanities subject, 30% teach a scientific subject, and 20% teach another subject, while the main observational sample only focuses on math teachers.

**Actions taken in assigning track recommendations.** In the teachers' survey, I directly elicit which actions they take in the process of assigning recommendations. Teachers can choose from a predetermined list. When choosing among the actions "I check my students' performance in the class register", "Another teacher checks students' performance", or "I remember my students' performance without checking in class register", approximately 60% of teachers of core subjects (math and Italian) report that they remember their students' academic performance, and 35% report that they check their students' performance in the class register (Figure 2). Moreover, almost 70% of teachers report that, during the teaching meeting in which they assign recommendations, teachers do not explicitly check the class register but talk more broadly about the students' attitudes and interests (Figure A13). Overall, the descriptive evidence on the recommendation process suggests that (i) teachers base their recommendations not only on students' academic performance but also on their perception of students' interests and attitudes and that (ii) what teachers recall about students' skills, attitudes and interests plays a crucial role.

## 6.2 Empirical Strategy

I estimate the following equation:

$$1(\text{Scientific})_{ijt} = \beta_0 + \beta_1 \text{Memory}_j + \beta_2 \text{Female}_{ijt} + \beta_3 \text{Female}_{ijt} \times \text{Memory}_j + \gamma_1 \mathbf{Z}_{ijt} + \gamma_2 \mathbf{X}_j + \gamma_3 (\text{Female}_{ijt} \times \mathbf{Z}_{ijt}) + \gamma_4 (\text{Female}_{ijt} \times \mathbf{X}_j) + \nu_{c(ijt)} + \mu_t + \varepsilon_{ijt} \quad (8)$$

where  $1(\text{Scientific})_{ijt}$  is a dummy equal to one if student  $i$  assigned to teacher  $j$  in year  $t$  is recommended to the scientific track and  $\text{Memory}_j$  is a dummy equal to 1 if teacher  $j$  reports that, in the process of assigning track recommendations, she *remembers her students' academic performance without checking it in the class register*. We consider only students assigned to teachers in core subjects (math and Italian) since they have a crucial role in the recommendation process. The sample includes teachers who reported that they check, that

another teacher checks, or that they rely on memory. A unit of observation is an 8th-grade student assigned to teacher  $j$ . Since we include both math and Italian teachers in the baseline specification, one student can appear twice if both her math and Italian teachers answered the survey.  $\mathbf{X}_j$  is a vector of teacher controls including subject studied (sciences/humanities), subject taught (sciences/humanities/other), teacher’s age, father’s education, gender, type of contract (fixed term/permanent), and indicators for schools in the North and teachers born in the North.  $\mu_t$  indicates year fixed effects and  $\nu_{c(ijt)}$  class fixed effects, to account for class-specific factors affecting all students within the class.  $\mathbf{Z}_{ijt}$  is a vector of student-level controls including 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. Our coefficient of interest  $\beta_3$  measures the increase in the gender gap for teachers who report that *do not check student performance but rely on their recollections*. In order to interpret  $\beta_3$  as the effect of relying more on memory rather than checking student information, the following assumption needs to be satisfied:

**Assumption 1.** *Conditional on the student level controls and class fixed effects, omitted student- or teacher-level variables affecting gender gaps in track recommendations should be orthogonal to the teacher’s reporting that she does not check but relies on memory.*

**Characteristics of teachers who check vs. rely on memory.** One key concern is that teachers who do not check and rely more on memory may have other characteristics driving gender gaps in their recommendations. As one may expect, teachers who do not check are more likely to be older and from schools in the South (Figure A14). To ensure that these differences do not drive the result, I control for a wide set of teacher characteristics, implicit stereotypes that the teachers hold as measured by an implicit association test (IAT), and their interaction with student gender.

### 6.3 Results

Table 2 displays the  $\beta_2$  coefficient from the estimation of model 8. With controls for students’ standardized test scores and year fixed effects (column 2), teachers who remember students’ performance are around 4.6 pp less likely to send girls to the scientific track than they are similar boys, while no gap is present for teachers who rely less on memory. When I control for student and teacher characteristics (including implicit stereotypes held by teachers) and account for class-specific factors with class fixed effects, the gap remains almost unchanged (columns 3–5). The results remain similar when I allow the teacher and student character-

istics to have a differential effect by gender (column 6). The effects are large: remembering students’ performance increases the gender gap by 22–24.5% of the boys’ mean, and checking students’ performance eliminates the gender gap. Figure 3 plots the fraction of students recommended to the scientific track by gender and bin of math ability separately for teachers who recall and those who check performance. For every decile of math ability above the 5th, girls are less likely to be recommended to the scientific track than boys with similar ability if their teachers recall their performance. In contrast, the gender gap in every ability decile is absent if teachers check the performance.

**Do teachers discriminate more when they have a higher number of students to recommend and they *do not check*?** Figure 3 shows the coefficients from estimating equation 8 separately for teachers who (i) remember their students’ performance and have many students to recommend (more than 23, the median), (ii) remember and have few students to recommend, (iii) check and have many or (iv) check and have few students to recommend. The fact that a higher number of students to be recommended has a larger effect on teachers who report relying more on memory supports the interpretation of the result in section 5 as the effect of a higher cognitive load. Gender discrimination increases when teachers have more limited memory capacity. In such cases, the missing science-track girls are partially allocated to the top-classical track.

## 7 Teachers’ Experiment

From the observational evidence, it is difficult to rule out all possible confounders; for example, while I present evidence that different class dynamics by gender, resource, or attention constraints during the academic year and teachers’ perceived recommendation thresholds are unlikely to explain the patterns, I cannot rule out the presence of other confounders that are unobservable to the econometrician. Moreover, I cannot directly test whether teachers recall a smaller and more biased set of characteristics about their students when they think about their students off the cuff (rather than checking their information) or when they have many other students on their minds.

I implement two experiments to show that decision-makers make more stereotypical decisions with binding memory constraints in a setting without confounders and to provide direct evidence for a memory mechanism. I conduct the first experiment with Italian public middle school teachers who participated in the survey; 448 teachers from 69 middle schools completed the experiment. In this section, I outline the experimental design, describe the empirical approach employed, and present the findings.



## 7.1 Design

Teachers observe objective information about a series of hypothetical student profiles and then provide track recommendations. The gender of each student and memory loads are cross-randomized across teachers. A summary of the experimental design is presented in Figure 4. The experiment proceeds as follows. First, teachers are asked to provide track recommendations for a series of hypothetical student profiles. Then, they observe 4 student profiles one by one. After reading the profiles, the teachers are asked to report what they remember about the academic performance and interests of each student and to provide a track recommendation. I elicit their recollections of student characteristics through open-ended questions to mimic a decision process in which teachers assess their students by retrieving experiences about them from memory without checking student-specific information.

**Choice of student profiles.** Each student profile contains information about (i) the student's academic performance in math and Italian, (ii) the student's interests, and (iii) some additional information on the student's background. I design these student profiles intentionally to include two profiles (Carlo/Carla and Marco/Anna) with "extreme" signals, where one excels in humanities but struggles in math and the other exhibits the opposite pattern. Additionally, I include two profiles (Roberto/Roberta and Francesco/Francesca) with "mixed" signals, where one excels in both mathematics and humanities and one has average grades in both subjects. Below are detailed descriptions of the four 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/Francesca is a good student but not excellent. He/She got a 7 in math*

and Italian, and his/her GPA is around 8. He/She is also very passionate about languages, and he/she got an 8 in English. He/She spent 3 weeks in Ireland in the summer where he/she substantially improved his/her ability to speak English, which is considerably above average. He/She cares a great deal about his/her group of friends, and both his/her parents are 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 he/she often forgets to do homework. He/She often disrupts lectures by chatting with his/her friends. He/She is very intuitive and talented in math, where he/she got an 8, while he/she got a 6 in Italian. He/She is passionate about fixing bikes with his/her older brother, and he/she loves playing video games.

**Randomizing memory constraints.** After observing the student profiles one by one, teachers in the *baseline* group make their recommendation decisions with the profiles in front of them (proxying a perfect memory benchmark), while teachers in the *memory* group cannot review the profiles and need to retrieve the students' characteristics from memory. Intuitively, teachers in the memory group likely struggle to recollect the student characteristics, as the task of processing four profiles simultaneously creates a cognitive load that hinders their ability to recall information. After observing the students' profiles, teachers in both the *baseline* and *memory* groups are asked to recall the characteristics of each profile and to provide a track recommendation for each student. The recall task is financially incentivized: teachers participate in a lottery for a bonus prize, and recalling more characteristics increases the likelihood of winning the prize. Figure A17 displays the recall task's experimental screen for the baseline and memory groups.

The randomization of memory constraints was designed with dual objectives. One aim was to replicate common real-life scenarios where individuals rely on recollection of information when making decisions. The other objective was to establish a benchmark in the form of the *perfect memory* condition. This benchmark serves as a reference point, enabling an assessment of the extent to which memory limitations can impact decision-making processes and amplify bias. Originally, I had planned and preregistered an additional treatment involving a second randomization of memory constraints by introducing an extra cognitive load. However, because of unexpected challenges such as a lower response rate and higher attrition, I implemented only the memory treatment to enhance statistical power.

**Technical challenges.** The baseline sample consists of all teachers who reached the experiment part of the survey (and were thus randomized into an experimental condition) and provided a recommendation for at least one student profile. Initially, I did not force teachers to complete all parts of the survey for ethical reasons, and this led to attrition and differential attrition of teachers in the memory treatment. As I acknowledged this problem, I introduced the "force response" option so that teachers had to provide recommendations for all students to proceed with the survey. This reduced the differential attrition and drastically reduced the number of teachers choosing to provide recommendations for only some students.<sup>30</sup> In the final sample, there is a differential attrition rate of 5% (23 teachers) across the memory and baseline groups, while no such differential attrition is observed in the gender randomization group.

This differential attrition could potentially introduce bias into the results in two opposing directions. On one hand, the missing teachers could be individuals who paid minimal attention to the student profiles and, if they had responded, might have provided even more stereotypical recommendations because of their limited recollection of students' signals. In this scenario, the results would be biased downwards, underestimating the true effect. The opposite, more concerning possibility is that the missing teachers are those who are particularly conscientious about providing appropriate recommendations and, had they participated, might have provided gender-neutral or counterstereotypical recommendations.

However, it appears unlikely that this second scenario could bias the results significantly. This is supported by the observation that the baseline sample is balanced across a wide range of teacher characteristics between the control and memory groups despite the differential attrition (Table 3). In particular, teachers in the two groups are similar in terms of education, socioeconomic background, age, and gender, and they teach similar subjects. They also have similar attitudes regarding the importance of different factors in deciding recommendations, and they have similar implicit association test (IAT) scores, indicating that the missing teachers are not differentially likely to be gender biased. To further ensure that attrition does not significantly bias the results, I take these additional steps: I replicate the analysis on the sample of all teachers while treating "no recommendation" as an additional outcome, and I include teachers' fixed effects in the main regressions to further ensure that the results are not driven by differences across teachers. The results remain very similar, making it unlikely that the differential attrition biases the results.

---

<sup>30</sup>In the final sample, 430 teachers provided recommendations to all students, 18 teachers provided recommendations for only some students, and 55 teachers viewed the profiles (and were thus randomized) but did not complete the survey (12% of attrition). Out of these 55 teachers, 16 teachers dropped out from the baseline group while 39 dropped out from the memory group (differential attrition of 23 teachers).

## 7.2 Methodology

The coefficient  $\beta_3$  in the regression below measures the increase in the gender gap with limited memory capacity:

$$Y_{ij} = \beta_0 + \beta_1 \text{Memory}_{ij} + \beta_2 \text{Female}_{ij} + \beta_3 \text{Memory}_{ij} \times \text{Female}_{ij} + \mu_i + \beta_4 \mathbf{X}_j + \varepsilon_{ij} \quad (9)$$

The observation is a student  $i$  evaluated by teacher  $j$ , and the outcomes  $Y_{ij}$  are dummies indicating the track recommendations received and whether the teacher recalled female- or male-typed characteristics. The vector  $\mathbf{X}_j$  contains teacher-level controls including teacher birth year, gender, subject taught (humanities, sciences, 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, and  $\mu_i$  is student fixed effects.

Since each teacher evaluates four student profiles and each profile’s gender is randomized across teachers, we can include teacher fixed effects  $\nu_j$  in our model, estimating the within-teacher increase in discrimination:

$$Y_{ij} = \beta_0 + \beta_1 \text{Memory}_{ij} + \beta_2 \text{Female}_{ij} + \beta_3 \text{Memory}_{ij} \times \text{Female}_{ij} + \mu_i + \nu_j + \beta_4 \mathbf{X}_j + \varepsilon_{ij} \quad (10)$$

In this case,  $\beta_3$  measures the within-teacher increase in the gender gap in recommendations for male vs. female students, accounting for student profile characteristics with student fixed effects. The main hypothesis is that  $\beta_3 < 0$  for recommendations to math-intensive tracks while  $\beta_3 > 0$  for humanities tracks.

**Classifying students’ attributes.** Figure A15 displays the characteristics of each student profile. I identify the keywords in teachers’ free recall text referring to each student’s characteristic described in the profiles and create dummies indicating whether each characteristic is mentioned by the teachers when they retrieve information about student  $i$ .

This procedure allows the measurement of two memory biases: selectively recalling certain student characteristics while neglecting others, and retrieving characteristics belonging to one student for another. Both biases are important to document since teachers may naively use all recalled characteristics to form their overall impressions of the students. Our main measure of recall includes all characteristics that the teacher reported when prompted to think about a student, regardless of whether they originally belonged to that student profile or to another hypothetical student.

### 7.3 Results

**Gender gaps in recommendations.** Figure 5 shows the fraction of students recommended to the scientific and humanities tracks by student gender separately for teachers making decisions based on their recollections or based on the actual information (Figure A18 shows all tracks separately, and Figure 7 shows results for two student profiles as an example). The difference in the gender gap for teachers who retrieve and those who review information is dramatic: the gender gap in recommendations to humanities tracks goes from 4.7 pp if teachers review the students’ signals (baseline group in the right graph in Figure 5) to 22.5 pp if teachers retrieve students’ signals from memory (memory group in the right graph in Figure 5), showing how relying on memory drastically increases gender discrimination. The left graph of Figure 5 shows a symmetric effect for recommendations to scientific and technical tracks. Having the possibility to review individual-level information before making the recommendation decision decreases the gender gap by approximately 80%, indicating that helping teachers correctly recall student information can substantially reduce inequitable decisions.

Table 4 reports coefficients  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  from the estimation of models 9 and 10. In the baseline condition, where teachers can check the information when making their decisions, girls are 4 pp (8.2% of the mean for boys) less likely to be recommended for the top-scientific or technical tracks, while they are 4.6 pp (9.7% of the mean for boys) more likely to be recommended to the top-classical and medium-humanities tracks than boys with same abilities and interests. When teachers cannot check student information and need to retrieve it from memory, these gender gaps are more than 5 times larger (there is a decrease in male-typed recommendations for girls of 37% of the mean for boys). Columns 2–3 and 6–7 progressively add controls and teacher fixed effects, showing that these gaps remain large. Columns 4 and 8 include teachers who did not provide a recommendation. The increase in gender gaps with binding memory constraints remains statistically significant, economically large, and similar across these specifications.

Figure A18 and Table A15 show how the gender gaps in recommendations for each track change with teachers’ memory constraints. The increase in gender gaps with teachers’ limited memory capacity is due to a reallocation of girls from the top-scientific and medium-technical tracks (columns 1 and 2) to the medium-humanities tracks (column 3). Interestingly, there is no gender gap in top-scientific track recommendations when teachers base their decision on the complete information set about students (column 1), a results that closely matches findings on past recommendations in the administrative data. When teachers can check the student information, the gap in top-scientific recommendations is close to zero, while it

becomes large and statistically significant (12 pp, a 52% decrease for girls with respect to the mean for boys) when teachers cannot check the student information and need to retrieve it from memory.

To what extent do teachers' choices for hypothetical students correlate with choices for past students? On average, teachers provide track recommendations that are consistent with the abilities and interests described in the profiles. Appendix D.4 and D.5 show gender gaps for past students with similar grades as each student profile in the administrative data, showing that we either observe the same modal track recommendation or that when the profiles' track recommendations differ, they do so in the direction of the additional student characteristics described in the experiment. Moreover, teachers in the *Memory* condition who display larger gaps in their recommendations to hypothetical students are the ones displaying larger gaps in their past actual track recommendations, as shown in Figure A19. Notably, teachers who did not experience gender gaps in track recommendations for their past students do not exhibit gender gaps even when placed in the *Memory* group, in line with the model's hypothesis that when teachers struggle to recall individual signals, they start using past experiences with other same-gender students as a proxy.

Finally, when analyzing results for each student profile, it emerges that teachers in the memory group still retain a perception of the student's characteristics, despite the emergence of gender gaps, as their most common track recommendation remains consistent with the baseline condition, as illustrated in Figure 7. For instance, when teachers facing binding memory constraints encounter an outstanding student who has participated in a math competition, they predominantly place them in the scientific track, even though gender gaps emerge.<sup>31</sup>

**Limited and biased recall of student characteristics.** As explained in the simple model in section 3, both limited and biased recall can increase discrimination. While recalling a limited (and unbiased) set of signals always increases discrimination, the direction of biases in recall is ex ante ambiguous. Research in psychology suggests that, under high memory loads, recall of stereotypical information prevails while, under low memory loads, recall of counterstereotypical information prevails (Macrae et al. (1993), Sherman and Frost (2000)).<sup>32</sup>

Moreover, across different contexts, women are often depicted differently than men. Fe-

---

<sup>31</sup>In other words, teachers with memory constraints disproportionately place Roberto/Roberta on the scientific track and Marco/Anna on the technical track, indicating that they still have a perception of the different profiles' characteristics.

<sup>32</sup>On the one hand, surprise effects may induce people to retrieve more stereotype-inconsistent characteristics (e.g., teachers may be more likely to recall an exceptional math performance for a girl than a boy, as it contrasts with the norm). On the other hand, it may be easier to retrieve information that aligns with the stereotype (e.g., recall a negative math performance for a girl).

male economists are disproportionately described with attributes related to their physical appearance while male economists with attributes related to their competence (Wu (2018)). In educational contexts, gender bias in student evaluations of teaching is often present (Boring (2017), MacNell et al. (2015)), and female and male professors tend to be described with different attributes in the online forum *Rate my Professor* (for instance, words such as *organized*, *emotional* and *kids* are more common for women while *brilliant*, *smart* are more common for men).<sup>33</sup> To my knowledge, this is the first paper documenting how gender biases emerge in recall holding fixed the underlying characteristics of subjects, and their relationship with biases in decisions in an applied setting.

Figure 6 displays the recall dynamics for each student characteristic in the memory group. On average we observe disproportionate recall of characteristics related to humanities and emotions for girls, and tech, science, and outdoor for boys, even though this average effect masks more nuanced recall patterns, and in the case of the extremely male-typed student profile, there is some suggestive evidence that surprise effects may be at play as well.

I proceed by classifying as female-typed those characteristics related to emotions and humanities and as male-typed those characteristics related to technology, science, the outdoors, and not being diligent. Female-typed characteristics are displayed in pink and male-typed characteristics in blue in Figure A15. Panel (b) in Figure 5 displays teachers' probability of recalling at least one female-typed or male-typed characteristic by student gender and teacher memory capacity. In the baseline group, female-typed characteristics have a 57% and male-typed characteristics a 59% probability of being recalled when prompted to think about a student. When teachers cannot check student information, both female- and male-typed characteristics are approximately 20 pp less likely to be recalled. Moreover, the student's gender affects what characteristics are retrieved. With binding memory constraints, teachers are 8.8 pp more likely to recall female-typed characteristics if they observe a girl (17% of the mean for boys in the baseline group) and 4.5 pp less likely to retrieve male-typed characteristics. Table 5 reports the coefficients when I progressively add controls and teacher fixed effects (columns 2–3 and 6–7) and when I include only teachers who recalled at least one student characteristic (columns 4 and 8). Figure A39 and Tables A16 and A17 of the Appendix report a similar analysis for teachers' recall of students' grades. Grades are recalled less frequently in the memory condition, while I do not find significant gender biases in the recalled grades.<sup>34</sup>

---

<sup>33</sup>Ben Shmidt created an online interactive tool that can be used to visualize differences in words used to describe male and female teachers in about 14 million reviews from RateMyProfessor.com (the interactive tool can be found at this [link](#)).

<sup>34</sup>Teachers in the baseline condition report students' Italian and math grades 95% of times. The fact that a minority of teachers in the baseline group still do not report students' grades suggests that not all

Do these biases in recall reflect false memories or selective memories of characteristics truly belonging to the students? When thinking about Roberta rather than Roberto, teachers may disproportionately recall that the student is good in humanities, which was a true characteristic of Roberta. They may also mistakenly recall that Roberta loves reading, which was a characteristic belonging to Carla. The baseline measures of recall report all characteristics recalled by the teacher, regardless of whether they are selective memories (case 1) or false memories (case 2). Figure A36 displays the recall dynamics separately for selective and false memories. Reassuringly, most memories reflect selective memories, indicating that teachers are not mainly guessing, even though some false memories are present as well.

**Biases in decisions and biases in recall.** The foundational assumption of the simple model in section 3 is that decision-makers form beliefs based on their recollections of information and make decisions based on those beliefs. However, if decision-makers are aware of their biases in recall, they may not rely on their memories. I assess whether decision-makers' recollections predict their decisions. Figure 8 shows how gender gaps in female-typed recommendations (in orange) and male-typed recommendations (in blue) evolve according to the number of signals recalled, with teacher fixed effects and controls for teachers' bias in recall. When I control for the biases in the set of recalled signals, teachers who recall fewer signals fall back on the stereotypical decision, suggesting that teachers' pre-existing gender categories affect their choices when they struggle to recall information.

Moreover, teachers heavily rely on the recalled signals. The binned scatterplots in Figure 8 describe how humanities track recommendations and scientific track recommendations are influenced by teachers' biases in recall (measured as the difference between female-typed and male-typed signals recalled), with controls for student and teacher fixed effects and for the total number of recalled signals. Recalling one additional female-typed rather than male-typed characteristic decreases the probability of making a scientific track recommendation by approximately 20%, indicating how small biases in recall may translate into large biases in decisions when decision-makers are on the margin.

---

teachers in the baseline group are fully attentive. The likelihood of recalling any Italian or math grade is respectively 17.9 pp (18.8%) and 18.6 pp (19.5%) lower for teachers in the memory group. Table A17 shows that among those who recall the students' grades in Italian, teachers report higher recalled Italian grades if they are assigned to female students, but the difference is not statistically significant. I speculate that one possible reason for the absence of a significant gender bias in the recall of grades can be due to the different way in which grades were elicited. Grades were elicited by choosing from a dropdown menu, while characteristics were elicited through a "free recall" task (by asking teachers to recall without any specific prompts). Sherman and Frost (2000) suggest that the recall advantage of stereotypical information seems to be due to retrieval advantages as it is found in "free recall tasks", while it is not found in "recognition tasks" where subjects are presented with different options and do not need to freely recall.



## 8 Prolific Experiment

Does the memory mechanism reflect a more general phenomenon occurring when decision-makers assess others’ ability by retrieving information from memory? In this section, I report results from a large-scale experiment in which decision-makers assess the ability of a candidate in a male-typed and a female-typed domain. With respect to the teacher experiment, in this setting, the decision-makers (i) assess only one candidate, which arguably reduces the focus on gender, and (ii) are financially incentivized to provide the most accurate evaluation possible for their candidate (they are incentivized to estimate their candidate’s true ability).

### 8.1 Design

In the experiment, 1239 participants, henceforth referred to as evaluators, were recruited through the online platform Prolific and completed the experiment.<sup>35</sup> The experiment uses a 2x2 across-subject design similar to the experimental design in the teacher experiment. I vary the candidate’s gender and evaluators’ memory capacity.<sup>36</sup> Figure 9 shows a graphic representation of the general setup of the experiment. The flowchart shows what participants do in each part and when the treatment interventions take place. I describe the experimental design and the two interventions in detail below.

*Part A: Prior on ability of the assigned candidate.* In Part A, evaluators are told that a large number of candidates have been asked to answer 80 trivia questions in two different domains (sports and pop culture), and the evaluators are asked to evaluate the ability of one of these candidates, whose name is either John or Susan (candidate gender is conveyed by the name and is randomized across evaluators). Then, they are asked to estimate how many of 10 randomly selected questions their candidate answered correctly in each domain. Since evaluators do not know anything about their candidate apart from her gender, this is a way to elicit evaluators’ prior on their candidate’s ability.

---

<sup>35</sup>Prolific was designed by social scientists to attain more representative samples online; Prolific samples have been shown to perform well relative to other subject pools (Gupta, Rigotti, and Wilson 2021). 234 participants were recruited in the pilot study, and 1005 participants were part of the main study. The pilot study had an identical design, but there were some minor differences, described in Appendix G. I present the pooled results and the results excluding the pilot for completeness.

<sup>36</sup>The experiment consists of three incentivized parts and an initial part that collects demographic information. All participants receive general instructions informing them that one part of the experiment is preselected for a bonus payment. They receive a 3\$ participation fee, and they earn points in each part. Their likelihood of receiving the bonus payment increases with the points that they earn in the preselected part (which is unknown to them).

*Part B: Evaluators observe the candidate answering a set of questions.* Next, evaluators observe their candidate's answers to 10 new questions in each domain. Evaluators know that they do not have to answer the questions themselves and that they have to pay attention to the questions and to their candidate's answers. They are shown one question at a time, along with their candidate's answer and the correct answer (Figure A41 shows an example of a question). The question order is randomized. Figure A40 shows the questions ordered according to the gender gap in answers by a set of subjects recruited prior to the main experiment. As expected, successes are more frequent among men and mistakes more frequent among women on average in sports, while the opposite is true for pop culture.<sup>37</sup>

Evaluators observe signals from real candidates who answered a larger set of questions (80 questions) in the two domains prior to the main experiment. I preselected two candidates of different genders who performed equally in the smaller set of 20 questions and would thus appear identical to evaluators. In this way, I can measure the candidate's actual performance in the broader set of questions and incentivize evaluators to provide their best estimate of their candidate's ability without answering strategically.

*Part C: Evaluators recall signals and evaluate candidates in a new task.* After observing the signals, participants are asked to recall their candidate's correct and incorrect answers and to evaluate his or her ability in a new, similar task. In particular, evaluators have to write down all the questions that they remember their candidate answered correctly and incorrectly in the "correct answers" and "incorrect answers" boxes (decision screens are displayed in Figure A42).<sup>38</sup> They can report either the question title, the question answer, or a uniquely identifiable question topic. The recall task is explicitly designed as an open-ended question to mimic real-life situations in which decision-makers think about a candidate off the cuff without any specific prompt. Then, they are asked to estimate their candidate's performance in a new randomly drawn set of 10 questions per domain.

---

<sup>37</sup>I selected the domains based on prior research by [Bordalo et al. \(2019\)](#), which shows that females tend to outperform boys in questions related to domains like Kardashian and Disney, while males perform better in domains such as sports, video games, cars, and math.

<sup>38</sup>The main difference between the two groups is the link provided to evaluators in the baseline condition to review the signals before recalling them. There is another minor difference between the two groups regarding financial incentives, which might attenuate the results. In particular, evaluators in the baseline conditions do not have financial incentives in the free recall task since the task is trivial for them, as they have the signals in front of them. This choice is motivated by the fact that retrieving signals from memory is mentally more costly for evaluators in the limited-memory condition and one may be worried that they might not make an effort to recall the questions, while it is a trivial task for evaluators in the baseline group. Removing financial incentives for the recall task in the no-memory-constraints condition is a conservative choice that might attenuate the effects of the limited memory treatment.

*Treatment interventions.* I use a  $2 \times 2$  across-subject design. First, I vary the candidate’s gender by varying the first name (either John or Susan). Candidates appear identical to the evaluators apart from gender, they answer half of the questions correctly by domain, and they answer the same questions correctly and incorrectly. Second, we vary evaluators’ memory capacity similarly to in the teachers’ experiment. Evaluators in the baseline group are provided with a link to review their candidate’s answers before the recall task (proxying a perfect memory benchmark), while evaluators in the memory group cannot review the signals and need to retrieve them from memory.

## 8.2 Methodology

I estimate the following equation:

$$Y_{idj} = \beta_0 + \beta_1 \text{Memory}_j + \beta_2 \text{Consistent}_{id} + \beta_3 \text{Memory}_j \cdot \text{Consistent}_{id} + \gamma X_{idj} + \mu_d + \mu_j + \varepsilon_{idj} \quad (11)$$

The observation is a candidate  $i$  evaluated by evaluator  $j$  in domain  $d$  (sports or pop culture). There are two observations for each candidate, as candidates are evaluated in both domains.  $\text{Memory}_j$  is a dummy equal to 1 if evaluator  $j$  is assigned to the memory treatment, and  $\text{Consistent}_{id}$  is a dummy equal to 1 if the candidate is female and the domain is pop culture or if the candidate is male and the domain is sports.  $\mu_d$  and  $\mu_j$  are, respectively, domain and evaluator fixed effects. The coefficient of interest  $\beta_3$  measures the increase in the gender gap in the outcome if the evaluator observed the stereotype-consistent rather than the inconsistent candidate and she has more limited memory capacity.  $\beta_2$  is the gender gap in the baseline condition without memory constraints.

The outcomes are  $Y_{idj} = \{A_{idj}, \text{Recall Share}_{idj}\}$ .  $A_{idj}$  measures evaluators’ assessment of their candidate’s ability in the new set of questions in domain  $d$  (i.e., the estimated share of correct answers). The variable  $\text{Recall Share}_{idj}$  measures the share of correct questions that the evaluator retrieves in domain  $d$  (the recalled answers in the observed set of questions).

Starting from the open-ended questions in which participants reported which questions they remember the candidate answering correctly and incorrectly, a research assistant cleaned the open-ended answers as follows. As written in the text of the experiment, participants could refer to a question by a uniquely identifiable question title, question topic, or question answer. For instance, for the pop culture question asking “*What are the names of Kim Kardashian’s children?*”, participants could refer to the question by reporting “*question on Kim Kardashian*”, “*Kim’s children*”, or “*Saint, North, Psalm [etc.]*”. As long as the participant

clearly referred to the question on Kim Kardashian, the research assistant cleaned the answer to indicate the reference to that question.

The questions were chosen to be fairly easy to remember, even though the exposure to many signals made it difficult for participants to recall all questions. The recalled share of correct answers in domain  $d$  is then computed as the ratio of correct answers retrieved in domain  $d$  over the total answers retrieved:

$$\text{Recall Share}_{idj} = \frac{R(\text{correct})_{idj}}{R(\text{correct})_{idj} + R(\text{incorrect})_{idj}}$$

By directly eliciting recall, I can examine what comes to mind (what evaluators retrieve) if their candidate has a different gender, and I can directly assess whether gender influences what information is retrieved about candidates.

### 8.3 Results

**Assessment of ability in the new task.** How is the assessment of the candidate’s competence affected by decision-makers’ memory constraints? Figure 10 shows the average assessment of the stereotype-consistent vs. stereotype-inconsistent candidate for evaluators in the memory and baseline conditions. When evaluators need to recall the information about the candidate and cannot review it, they evaluate the stereotype-consistent candidate to be approximately 10% more competent than the stereotype-inconsistent candidate. In contrast, when evaluators have full access to the candidate’s information, they evaluate the stereotype-consistent candidate to be approximately 3.3% more competent. Columns 1 to 3 of Table 6 show that the coefficients remain stable once we add controls (column 2) and if we include evaluator fixed effects (column 3).

**Limited and biased recall.** Limitations and biases in information recall act as a mechanism for explaining the increase in gender gaps. When evaluators need to retrieve information from memory and cannot check it, they recall approximately half of the individual-level signals and disproportionately recall stereotype-consistent signals. Evaluators in the baseline treatment recall on average 8 out of 10 signals for each domain while evaluators in the memory treatment recall on average 4 signals by domain (Figure A43). Moreover, the candidate’s gender influences which signals are recalled. The right graph in Figure 10 (a) shows that biases start from recall: when we look at which types of signals are recalled by evaluators, the recalled share of positive signals is higher if they observe a stereotype-consistent candidate. They recall more correct sports questions than mistakes if they observe a man instead of a

woman, while the opposite holds for pop culture (Figures A45 and A47 report the results separately by domain). Table 7 shows that the coefficient on biased recall remains similar once we add controls (column 2) and if we include evaluator fixed effects (column 3).

**Biases in recall and biases in decisions.** Similarly to in the teachers' experiment, I test whether, conditional on evaluators recalling the same number of signals, recalling a higher share of positive signals predicts decision-makers' assessment of ability. The right graph in Figure 11 shows that a one-standard-deviation increase in the share of positive signals recalled leads to a 0.20-standard-deviation increase in the assessment of ability.<sup>39</sup>

## 9 Conclusion

In this paper, I have developed and empirically tested a framework that explains why discrimination against women in male-dominated fields can persist even when decision-makers have access to ample candidate information. Decision-makers have limited memory capacity: they recall a limited set of individual-level information and disproportionately recall stereotype-consistent information, leading to more biased decisions. Limitations and biases in recall are driven by various factors prevalent in our daily lives. These factors include neglecting to check individual-level information or experiencing a heavy mental burden because of excessive workload or multitasking and can even occur in the absence of explicit time constraints, as in cases when decision-makers opt not to verify individual-level information because they believe that they remember it.

I focus on the formal career advice that Italian public school teachers provide their students at the end of middle school. When I compare students with similar backgrounds and abilities in math and Italian, girls are less likely to be recommended for the top-tier scientific high school track than boys. Such disparities are larger when teachers have many other students to recommend and are not present for teachers who report checking students' information in class registers when assigning track recommendations. I then directly assess whether limitations and biases in the recall of individual-level information can act as a mechanism. When they have binding memory constraints, teachers recall a limited number of student-specific signals, disproportionately recall stereotype-consistent signals, and send students onto stereotype-consistent tracks. Facilitating access to individual-level information at the time of the decision reduces these gender gaps by approximately 80%. Last, I

---

<sup>39</sup>Figure A44 in the appendix reports the results from the second check, showing that a lower number of recalled signals (with controls for biases in the signals recalled) is associated with larger gender gaps.

implement a large-scale online experiment to assess the extent to which memory limitations and biases amplify gender gaps in decision-makers' assessment of the ability of candidates.

In the setting studied, bias is not driven by a fraction of decision-makers always making unequal decisions but rather by the same decision-makers making more or less biased decisions depending on the environment in which they operate. One implication of the findings is that factors imposing memory constraints on evaluators should also be considered in assessing what policy should be implemented to reduce bias. This is particularly relevant for situations in which a great deal of information is present but information may be difficult to process—for instance, in cases of promotions, nominations, recommendations, and hiring processes with subjective assessments. One relevant example of an intervention reducing evaluators' memory constraints to improve equity is the introduction of scribes into MBA courses (course assistants who take notes on students' participation). Class participation is an important component of final grades in MBA courses, but instructors may have difficulties remembering students' actual class participation without the support of such scribes, instead falling back on stereotypes.

This paper opens several avenues for future research. I show how the same objective information is perceived differently by evaluators when they are under greater mental burdens, leading to stereotypical decisions that penalize girls in science. One unanswered question relates to how variations in past experiences affect decision-makers' current treatment of minorities. Moreover, biases in recall may differ by candidate's ability. While the disproportionate recall of stereotypical information could prevail when decision-makers observe mixed signals, surprise effects may prevail when they observe extreme signals—such as a girl who is extremely good in math—potentially leading to discrimination reversals. Overall, taking into account factors affecting the recall of individual-level information in theoretical and empirical analyses of discrimination could provide a deeper and more realistic understanding of its dynamics.

## References

- Agan, Amanda and Sonja Starr. Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, 133(1):191–235, 2018.
- Aigner, Dennis J and Glen G Cain. Statistical theories of discrimination in labor markets. *Ilr Review*, 30(2):175–187, 1977.
- AlmaLaurea, . Xxiii indagine profilo dei laureati 2021. *Rapporto 2022*, 2021.

- AlmaLaurea, . Focus gender gap 2023. *Rapporto 2023*, [Link to "Focus Gender Gap 2023"](#), 2023.
- Altonji, Joseph G and Charles R Pierret. Employer learning and statistical discrimination. *The quarterly journal of economics*, 116(1):313–350, 2001.
- Anderson, Michael C and James H Neely. Interference and inhibition in memory retrieval. In *Memory*, pages 237–313. Elsevier, 1996.
- Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart. Subjective models of the macroeconomy: Evidence from experts and representative samples. *The Review of Economic Studies*, 89(6):2958–2991, 2022.
- Arnold, David, Will Dobbie, and Crystal S Yang. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932, 2018.
- Arrow, Kenneth J. What has economics to say about racial discrimination? *Journal of economic perspectives*, 12(2):91–100, 1998.
- Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, 106(6):1437–75, 2016.
- Becker, Gary S. The economics of discrimination. 1957.
- Benson, Alan, Danielle Li, and Kelly Shue. Potential” and the gender promotion gap. Technical report, Working paper, 2021.
- Bertrand, Marianne and Esther Duflo. Field experiments on discrimination. *Handbook of economic field experiments*, 1:309–393, 2017.
- Bertrand, Marianne and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.
- Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. Implicit discrimination. *American Economic Review*, 95(2):94–98, 2005.
- Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope. Inaccurate statistical discrimination: An identification problem. Technical report, National Bureau of Economic Research, 2019a.

- Bohren, J Aislinn, Alex Imas, and Michael Rosenberg. The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436, 2019b.
- Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Beliefs about gender. *American Economic Review*, 109(3):739–73, 2019.
- Bordalo, Pedro, Giovanni Burro, Katherine B Coffman, Nicola Gennaioli, and Andrei Shleifer. Imagining the future: memory, simulation and beliefs about covid. Technical report, National Bureau of Economic Research, 2022.
- Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer. Memory and probability. *The Quarterly Journal of Economics*, 138(1):265–311, 2023.
- Boring, Anne. Gender biases in student evaluations of teaching. *Journal of public economics*, 145:27–41, 2017.
- Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. Gender, competitiveness, and career choices. *The quarterly journal of economics*, 129(3):1409–1447, 2014.
- Card, David and A Abigail Payne. High school choices and the gender gap in stem. *Economic Inquiry*, 59(1):9–28, 2021.
- Carlana, Michela. Implicit stereotypes: Evidence from teachers’ gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224, 2019.
- Carrell, Scott E, Marianne E Page, and James E West. Sex and science: How professor gender perpetuates the gender gap. *The Quarterly journal of economics*, 125(3):1101–1144, 2010.
- Ceci, Stephen J, Wendy M Williams, and Susan M Barnett. Women’s underrepresentation in science: sociocultural and biological considerations. *Psychological bulletin*, 135(2):218, 2009.
- Ceci, Stephen J, Donna K Ginther, Shulamit Kahn, and Wendy M Williams. Women in academic science: A changing landscape. *Psychological science in the public interest*, 15(3):75–141, 2014.
- Charles, Constantin. Memory moves markets. *Available at SSRN 4019728*, 2022.
- Cimpian, Joseph R, Taek H Kim, and Zachary T McDermott. Understanding persistent gender gaps in stem. *Science*, 368(6497):1317–1319, 2020.

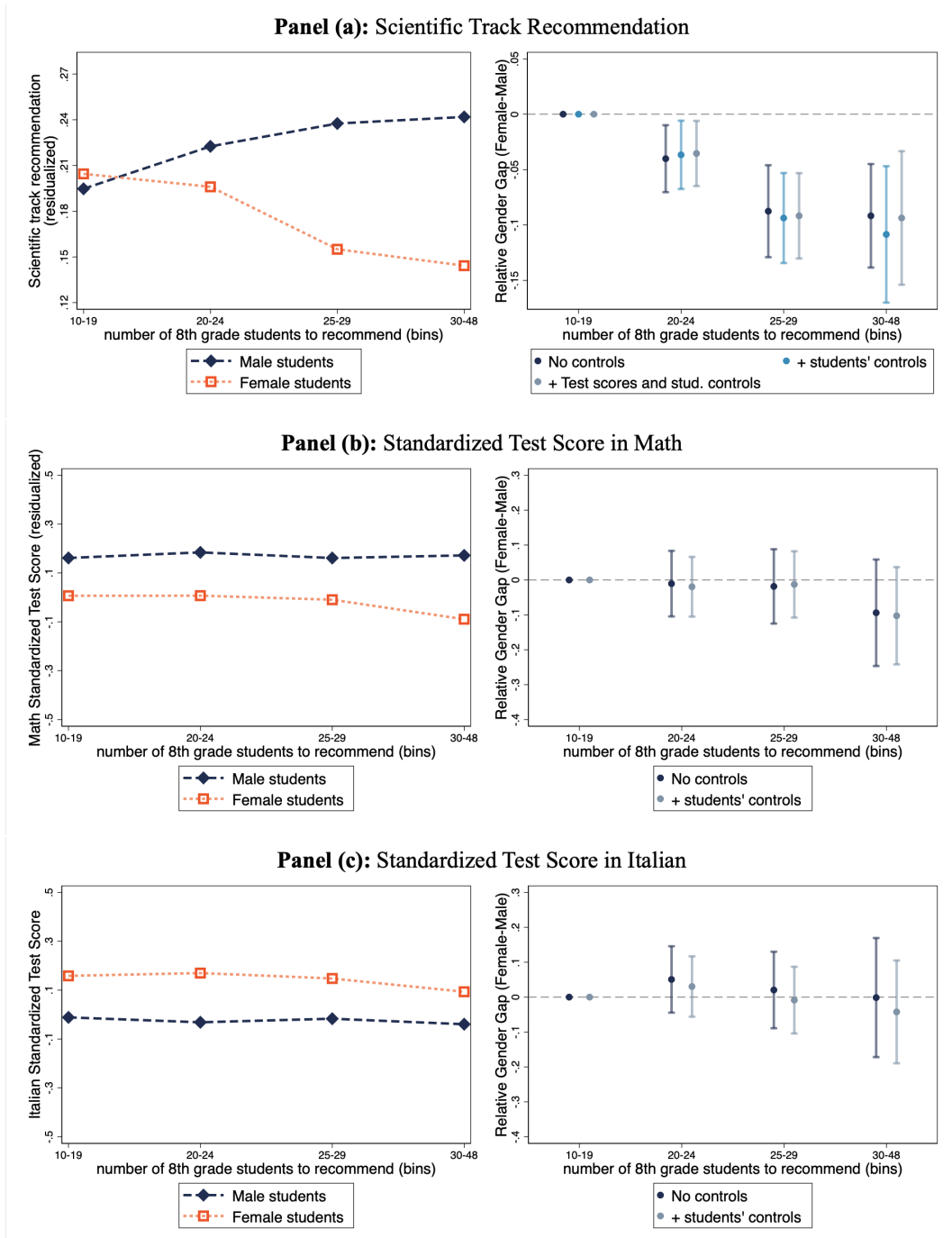


- Crowder, Robert G. *Principles of learning and memory: Classic edition*. Psychology Press, 2014.
- Doleac, Jennifer L and Luke CD Stein. The visible hand: Race and online market outcomes. *The Economic Journal*, 123(572):F469–F492, 2013.
- Egan, Mark, Gregor Matvos, and Amit Seru. When harry fired sally: The double standard in punishing misconduct. *Journal of Political Economy*, 130(5):1184–1248, 2022.
- Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann. Associative memory and belief formation. Technical report, National Bureau of Economic Research, 2020.
- Esponda, Ignacio, Ryan Oprea, and Sevgi Yuksel. Seeing what is representative. *The Quarterly Journal of Economics*, page qjad020, 2023.
- Fershtman, Chaim and Uri Gneezy. Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics*, 116(1):351–377, 2001.
- Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack. Selective memory equilibrium. *Available at SSRN 4015313*, 2022.
- Glover, Dylan, Amanda Pallais, and William Pariente. Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *The Quarterly Journal of Economics*, 132(3):1219–1260, 2017.
- Goldin, Claudia. A grand gender convergence: Its last chapter. *American economic review*, 104(4):1091–1119, 2014.
- Goldin, Claudia and Cecilia Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American economic review*, 90(4):715–741, 2000.
- Goldin, Claudia, Lawrence F Katz, and Ilyana Kuziemko. The homecoming of american college women: The reversal of the college gender gap. *Journal of Economic perspectives*, 20(4):133–156, 2006.
- Hill, Catherine, Christianne Corbett, and Andresse St Rose. *Why so few? Women in science, technology, engineering, and mathematics*. ERIC, 2010.
- Jenkins, John G and Karl M Dallenbach. Obliviscence during sleep and waking. *The American Journal of Psychology*, 35(4):605–612, 1924.

- Jiang, Zhengyang, Hongqi Liu, Cameron Peng, and Hongjun Yan. Investor memory and biased beliefs: Evidence from the field. *Available at SSRN*, 2022.
- Kahana, Michael Jacob. *Foundations of human memory*. OUP USA, 2012.
- Kahn, Shulamit and Donna Ginther. Women and stem. Technical report, National Bureau of Economic Research, 2017.
- Kessler, Judd B, Corinne Low, and Colin D Sullivan. Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review*, 109(11):3713–3744, 2019.
- Kessler, Judd B, Corinne Low, and Xiaoyue Shan. Lowering the playing field: Discrimination through contrast effects. 2022.
- Kline, Patrick, Evan K Rose, and Christopher R Walters. Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, 137(4):1963–2036, 2022.
- List, John A. The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics*, 119(1):49–89, 2004.
- MacNell, Lillian, Adam Driscoll, and Andrea N Hunt. What’s in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40:291–303, 2015.
- Macrae, C Neil, Miles Hewstone, and Riana J Griffiths. Processing load and memory for stereotype-based information. *European Journal of Social Psychology*, 23(1):77–87, 1993.
- McGeoch, John A. Forgetting and the law of disuse. *Psychological review*, 39(4):352, 1932.
- Mocanu, Tatiana. Designing gender equity: Evidence from hiring practices and committees. Technical report, Working paper, 2022.
- Moss-Racusin, Corinne A, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. Science faculty’s subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479, 2012.
- Mullainathan, Sendhil. A memory-based model of bounded rationality. *The Quarterly Journal of Economics*, 117(3):735–774, 2002.
- OECD, . Education at a glance 2022: Oecd indicators. 2022.
- Phelps, Edmund S. The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661, 1972.

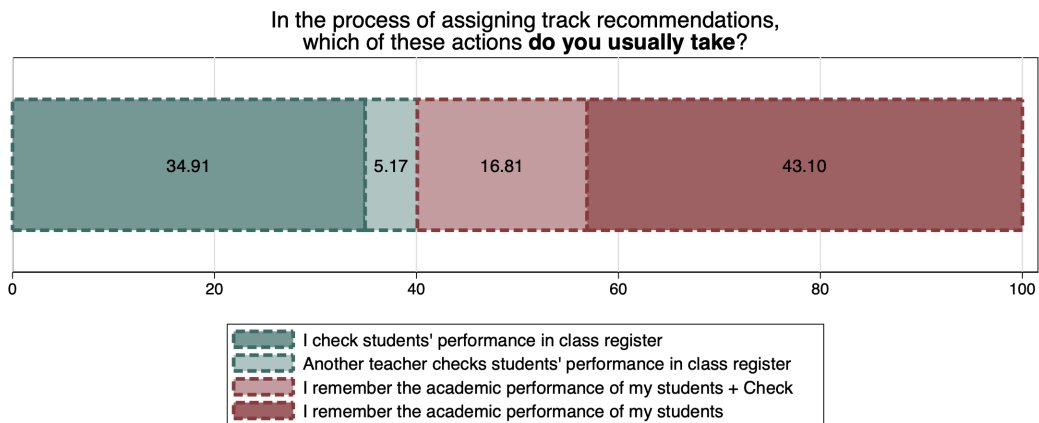
- Postman, Leo and Benton J Underwood. Critical issues in interference theory. *Memory & Cognition*, 1(1):19–40, 1973.
- Radbruch, Jonas and Amelie Schiprowski. Interview sequences and the formation of subjective assessments. 2021.
- Sarsons, Heather. Interpreting signals in the labor market: evidence from medical referrals. *Job Market Paper*, 2017.
- Sarsons, Heather, Klarita Gërkhani, Ernesto Reuben, and Arthur Schram. Gender differences in recognition for group work. *Journal of Political Economy*, 129(1):101–147, 2021.
- Sherman, Jeffrey W and Leigh A Frost. On the encoding of stereotype-relevant information under cognitive load. *Personality and Social Psychology Bulletin*, 26(1):26–34, 2000.
- Tversky, Amos and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.
- Underwood, Benton J. Interference and forgetting. *Psychological review*, 64(1):49, 1957.
- Wiswall, Matthew, Leanna Stiefel, Amy Ellen Schwartz, and Jessica Boccardo. Does attending a stem high school improve student performance? evidence from new york city. *Economics of Education Review*, 40:93–105, 2014.
- Wu, Alice. Gendered language on the economics job market rumors forum. *AEA Papers and Proceedings*, 108:175–179, 2018.

**Figure 1:** Gender gaps and number of 8th-grade students to recommend (*Past students*)



*Notes:* The left graphs show the residualized trend in the probability of being assigned to the scientific track (Panel a), math (Panel b) and reading (Panel c) standardized test scores 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 variable on the students’ controls, teacher and year-fixed effects, and plotting the residuals by gender after adding back the mean of the dependent variable. The right graphs show the  $\gamma_q$  coefficients from specification 6, without controls (dark blue estimates), and including controls for math and year fixed effects, test scores (in Panel a), and the full set of students’ level controls described in equation 6. The caps show 95% confidence intervals with clustering by teacher.

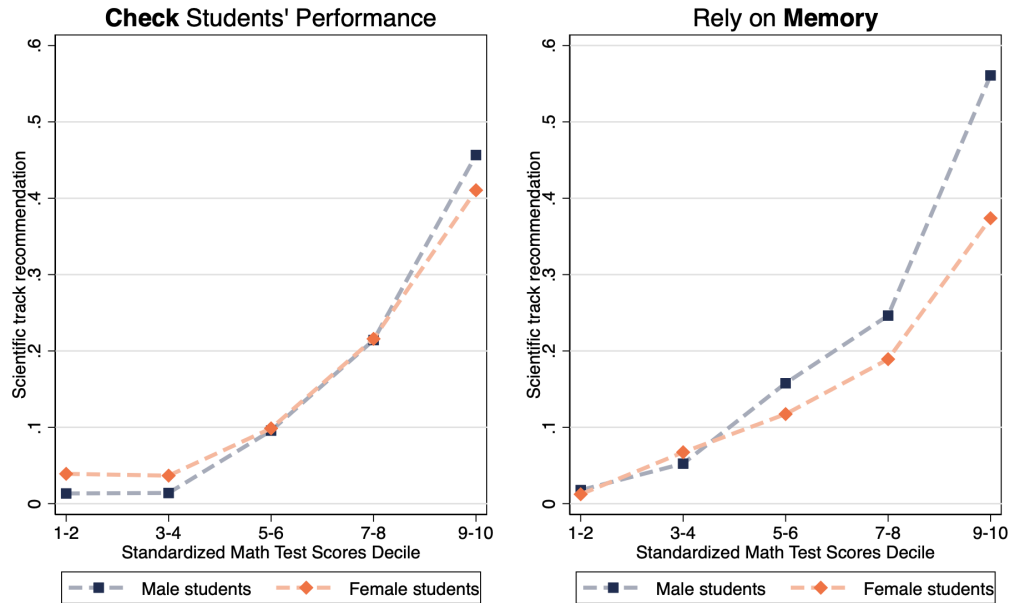
**Figure 2:** Actions taken when assigning track recommendations



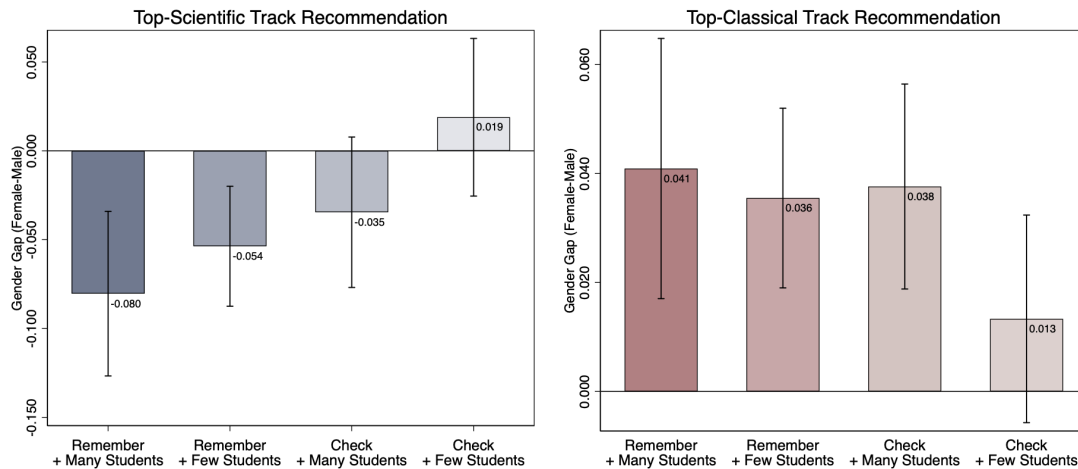
*Notes:* The graph shows teachers' answers to the following survey question: "In the process of assigning track recommendations, which of these actions do you usually take? (A) I check my students' performance in the class register, (B) Another teacher checks students' performance in the class register, (C) I remember my students' performance without checking it in the class register". The sample includes the 240 survey teachers in core subjects (math and Italian).

**Figure 3:** Gender gaps and binding memory constraints (*Past Students*)

**Panel (a):** Gender gaps for teachers who check student performance vs. rely on memory

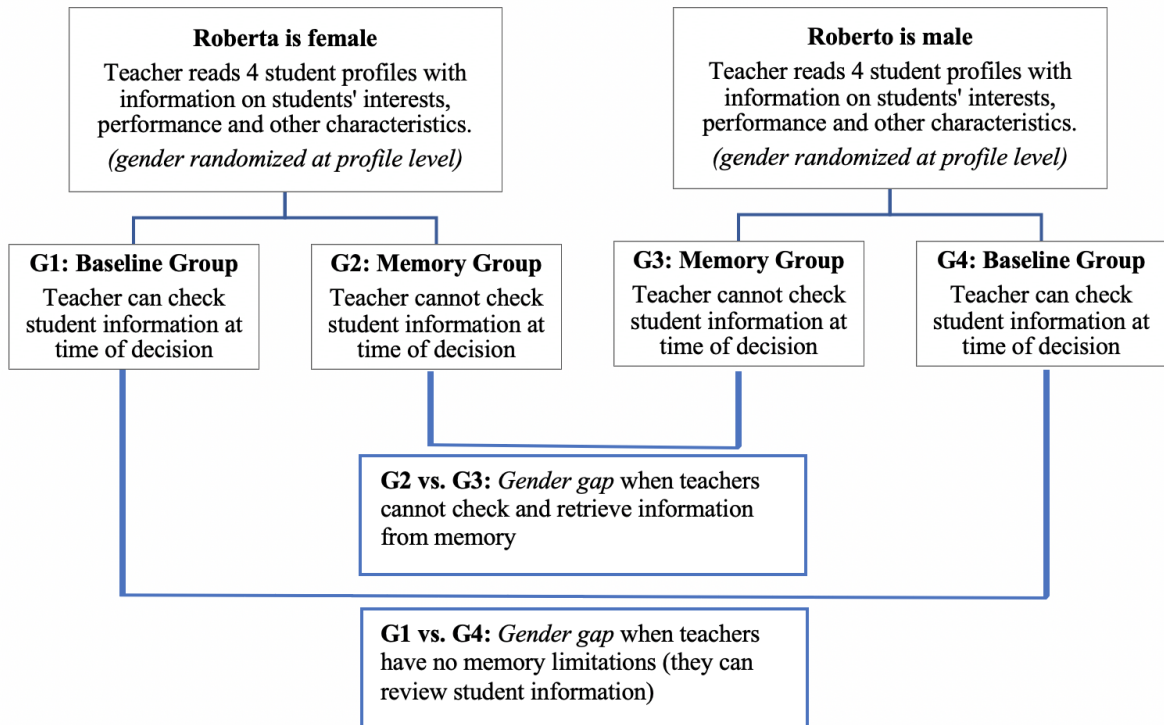


**Panel (b):** Gender Gap and Binding Memory Constraints



*Notes:* Panel (a) displays the fraction of male and female students recommended to the scientific track by their standardized math test scores separately for teachers reporting that they check their students' academic performance in class register (left figure) and do not check but rely on memory (right figure). Figures in Panel (b) display conditional gender gaps (controlling for standardized test scores and the same students and teacher controls as in model 8) in recommendations to the scientific track (left) and classical track (right) for students assigned to teachers who have many vs. few students to recommend (above vs. below the average) and do not check student performance relying more on memory, have many vs. few students to recommend and report that they check student performance. The caps in Panel (b) show 95% confidence intervals with clustering by teacher. The sample includes survey teachers in core subjects matched with their past 8th-grade students in 2016-19.

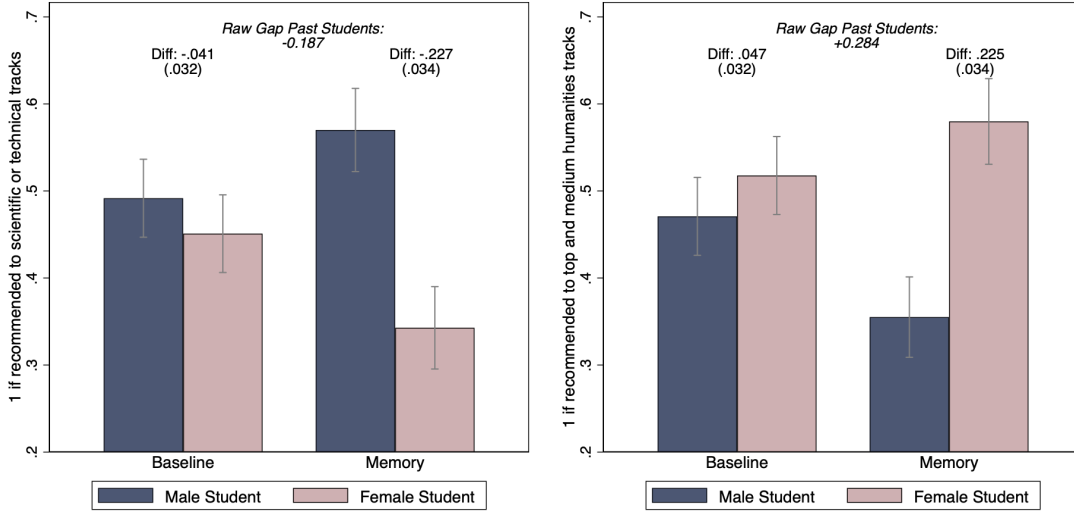
**Figure 4:** Teacher Experiment Flow



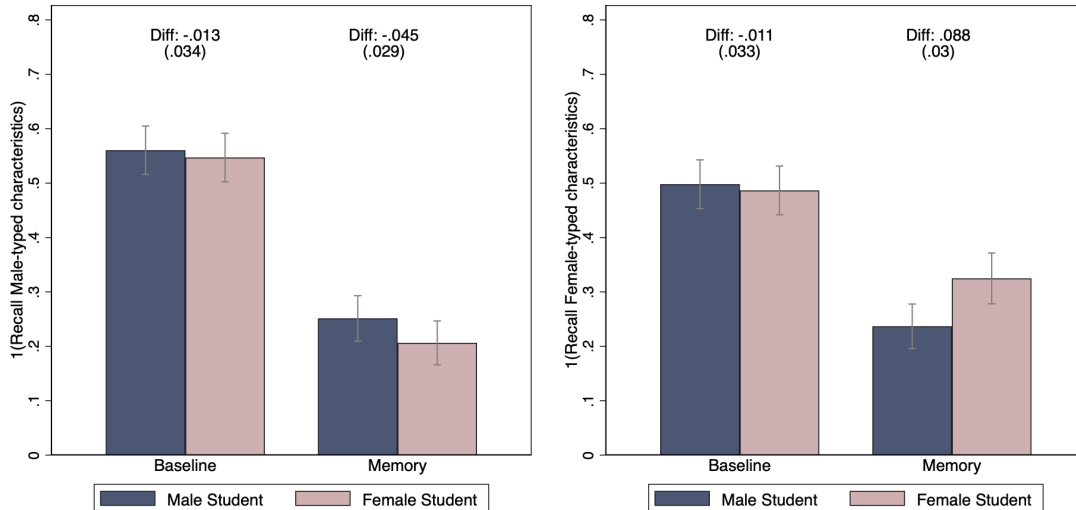
*Notes:* This figure shows the teacher experiment design. The groups G1, G2, G3, G4 denote the experimental groups. I randomized whether teachers can review student information or need to retrieve it from memory (*Baseline* vs. *Memory* conditions) across teachers, while student’s gender is randomized at the student profile level. Each teacher observes 4 student profiles (student profiles are shown in section 7). After observing the student profiles one by one, teachers are required to recall the characteristics of each student and provide a high school track recommendation (decision screens shown in Figure A17).

**Figure 5:** Experiment with Teachers: Gender gaps in decisions and recall of individual information and binding memory constraints (*hypothetical students*)

**Panel (a):** High school track recommendations



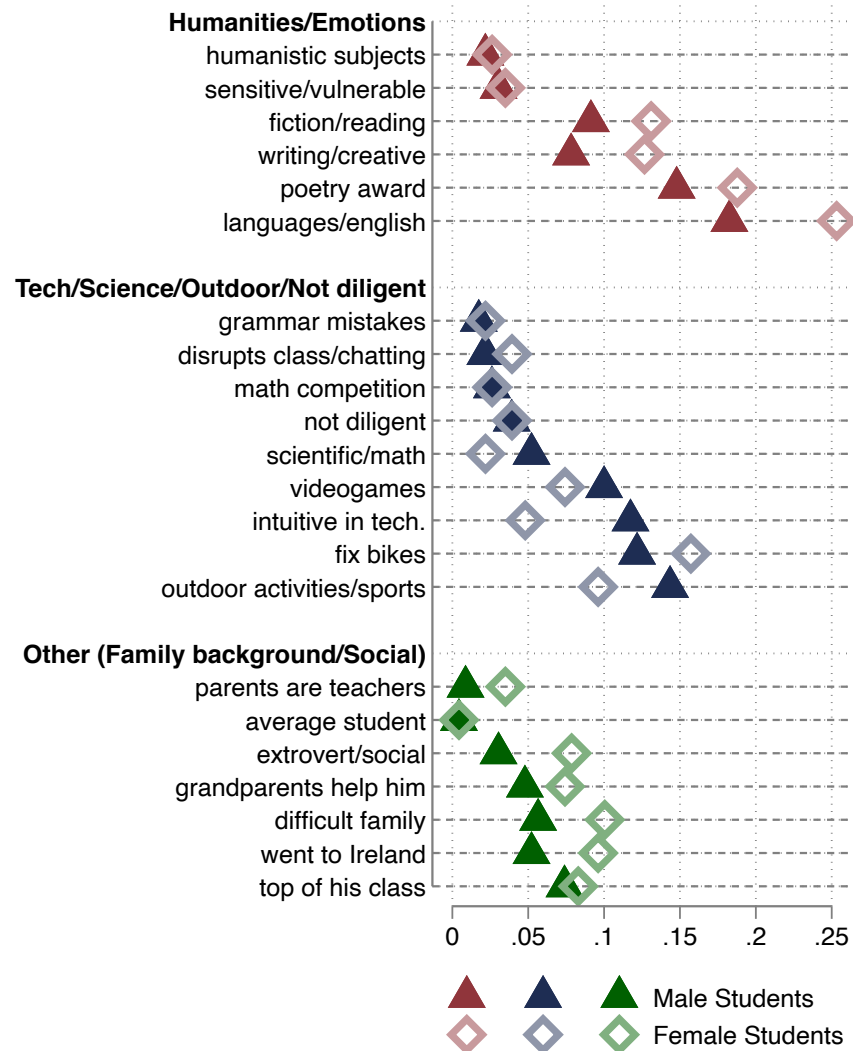
**Panel (b):** Recalled students' characteristics



*Notes:* The figures in Panel (a) show the probability that a student is assigned to scientific or technical high school tracks (left) and humanities tracks (right), for female and male students and for teachers in the memory and in the baseline condition. The figures in Panel (b) show the probability that teachers recall female-typed characteristics (left graph) and male-typed characteristics (right graph), for teachers in the memory and baseline conditions. Characteristics are classified as female or male typed as described in section 7. Both teachers in the baseline and memory conditions observed the same students' profiles. Teachers in the memory condition need to retrieve students' characteristics from their memory (they do not have the profiles in front of them when they make decisions), while teachers in the baseline condition can review students' characteristics before providing recommendations. The baseline sample included 448 teachers from 68 middle schools who completed the teachers' experiment.



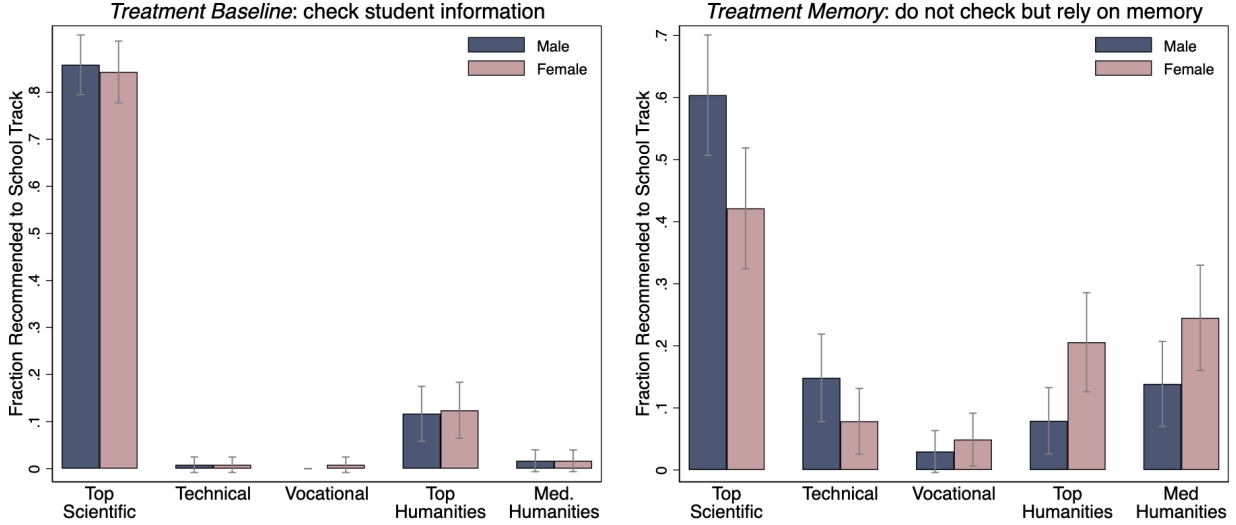
**Figure 6:** Experiment with Teachers: Recalled Characteristics, by Student's Gender (all characteristics belonging to students' profiles)



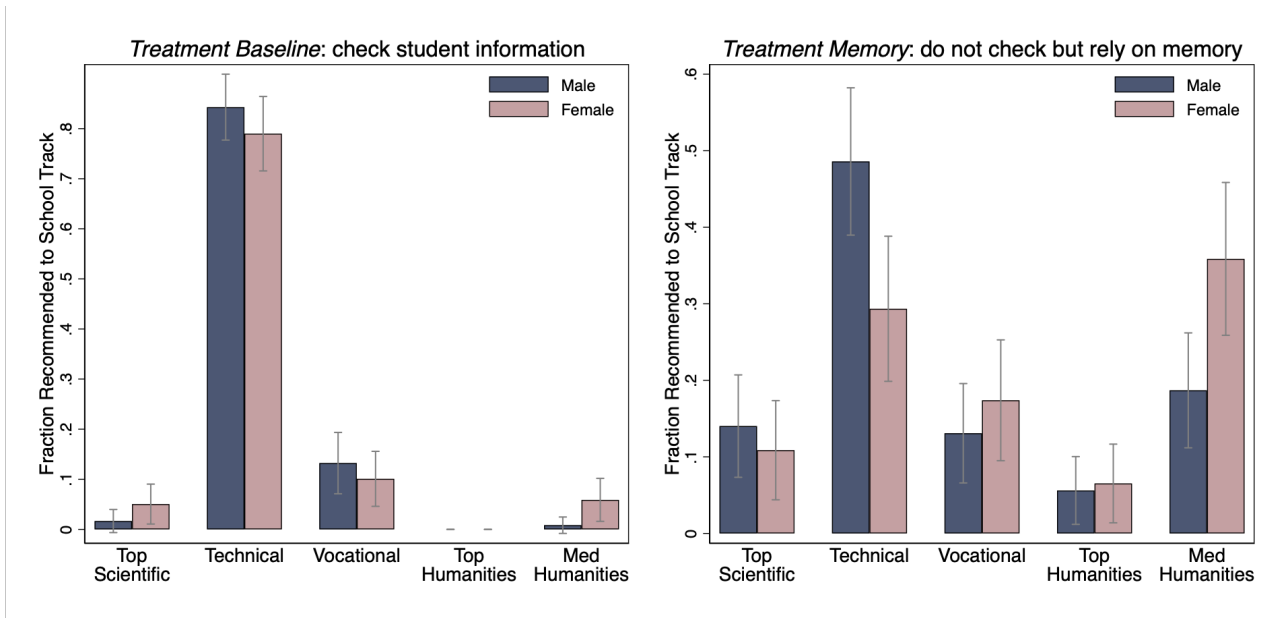
*Notes:* This figure shows the probability that each characteristic is recalled by teachers when they are prompted to think about boys or identical girls. The left graph includes characteristics related to math, sports, and mechanics. The graph in the center includes characteristics related to emotions, reading, and humanities. The graph on the right includes other characteristics, mostly related to family background. A characteristic is coded as recalled if the teacher reports the characteristics as describing a student, regardless of whether it is a correct memory (it truly belongs to the student) or a false memory (it belongs to another student). Figure A36 displays separately correct memories (recall of characteristics truly belonging to the student) and false memories (recall of characteristics that the teacher mistakenly attributes to the student).

**Figure 7:** Teacher Experiment: Examples from two student profiles

**Panel (a):** hypothetical student *Roberto/a*: excellent student both in math and humanities, participated in math competition



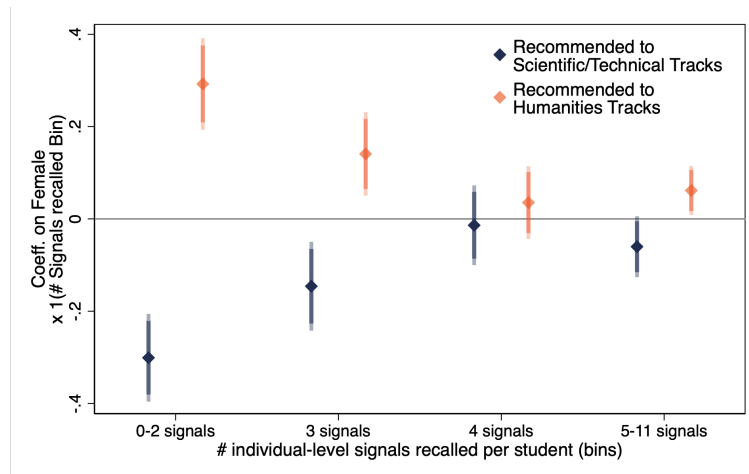
**Panel (b):** hypothetical student *Marco/Anna*: grade 6/10 in Italian and 8/10 in math and technology, passionate about videogames



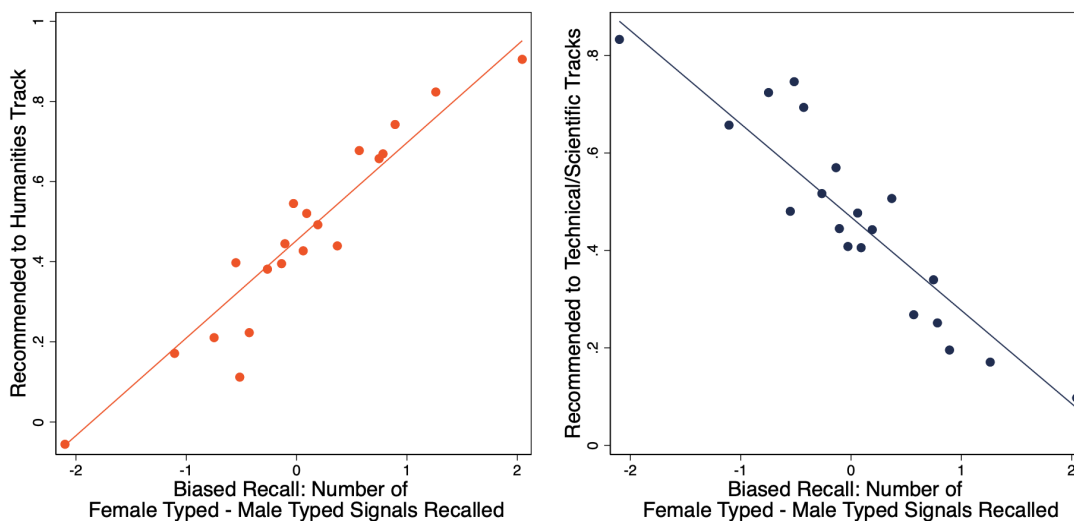
*Notes:* This figure shows the probability that the hypothetical student profiles named Roberto/Roberta (in Panel (a)) and Marco/Anna (in Panel (b)) are recommended to each high school track by teachers assigned to the baseline and memory treatments. Teachers in the memory condition cannot check student information at the time of the decision and need to retrieve it from memory, while teachers in the baseline condition can review students' characteristics when providing recommendations. The baseline sample included 448 teachers from 68 middle schools who completed the teacher experiment.

**Figure 8:** Experiment with Teachers: Gender Gap in Track Recommendations, Number and Types of Signals Recalled

**Panel (a):** Gender gap in Track Recommendations and number of individual signals recalled per student



**Panel (b):** Gender Gap in Track Recommendations and Biases in Recall



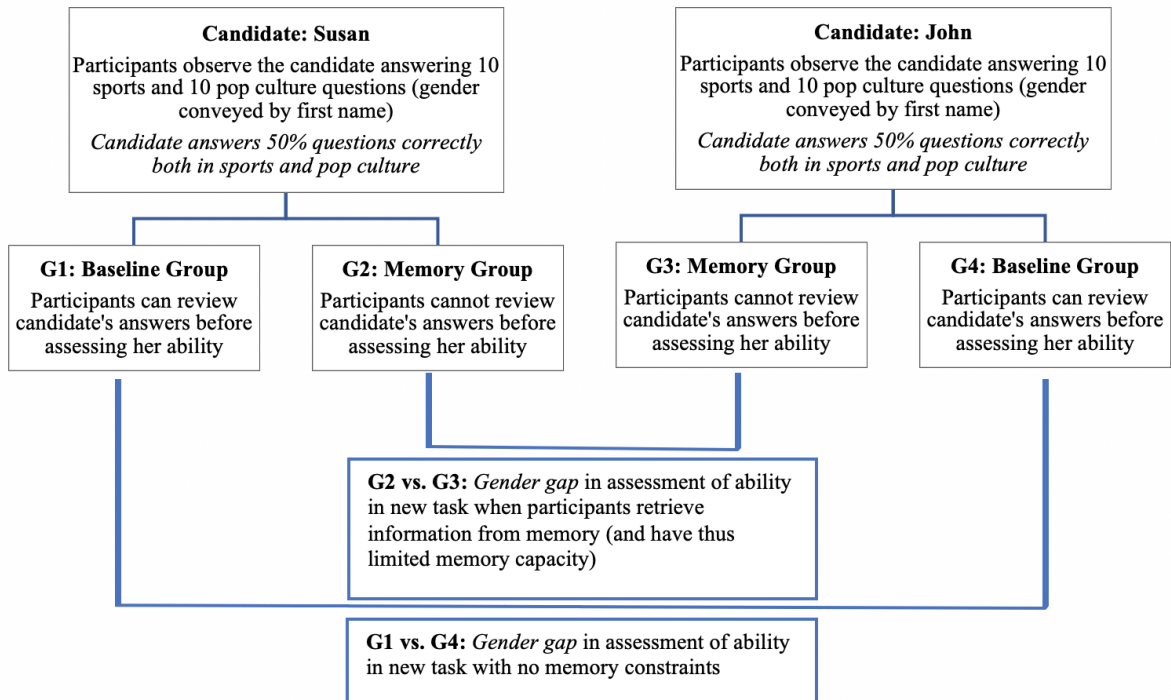
*Notes:* Panel (a) shows the gender gap in recommendations to humanities or scientific/technical tracks by the number of individual-level signals recalled by the teachers, controlling for the gender gaps in the recalled signals (number of female typed minus number of male typed signals recalled), and profile and teacher fixed effects. It shows coefficients  $\gamma_s$  from the following equation:

$$Y_{ij} = \beta_0 + \sum_{s=1}^4 \beta_s 1(\text{Recall}=s) + \sum_{s=1}^4 \gamma_s 1(\text{Recall}=s) \cdot F_{ij} + \delta_0 \text{Gap Recall}_{ij} + \mu_i + \nu_j + \varepsilon_{ij}.$$

Panel (b) shows binned scatterplots of the relationship between humanity track recommendations (left) and scientific track recommendations (right) and the gender gap in recall per student, measured as the number of female-typed minus male-typed signals recalled. I control for the total number of signals recalled and include teacher and profile fixed effects. It shows the  $\beta_1$  coefficient from estimating the following equation:

$$Y_{ij} = \beta_0 + \beta_1 \text{Gap Recall}_{ij} + \beta_2 \text{Tot Recall}_{ij} + \mu_i + \nu_j + \varepsilon_{ij}.$$

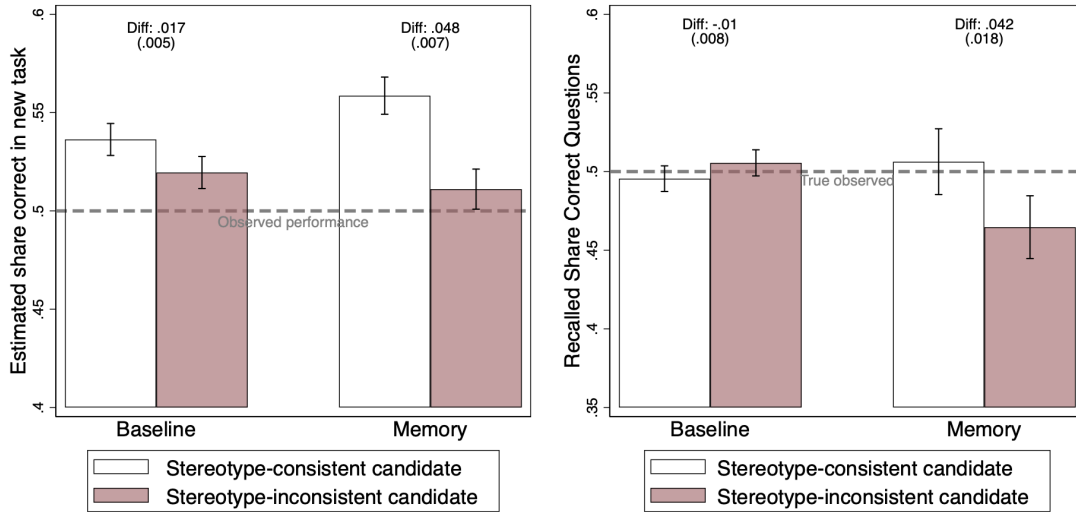
**Figure 9:** Experiment with US survey sample flow



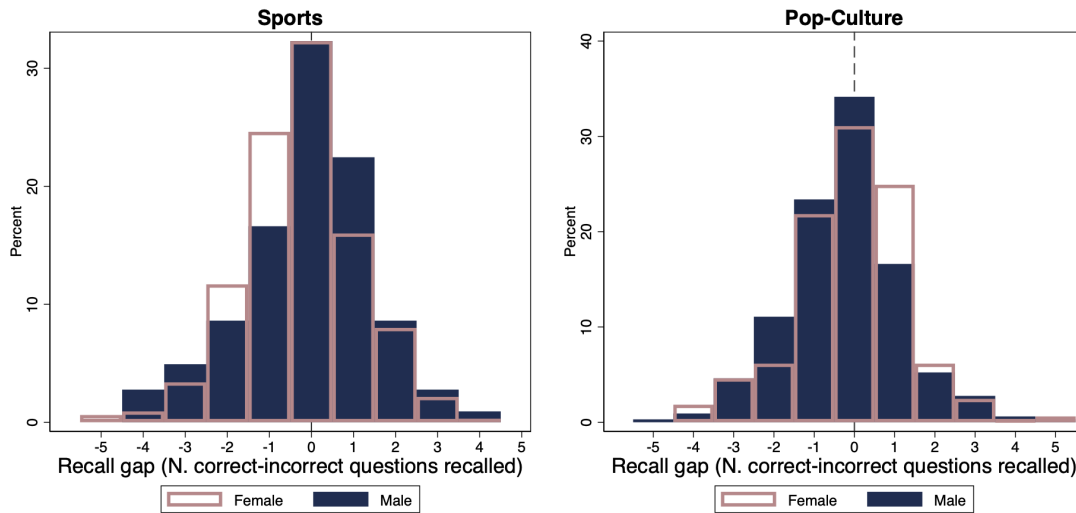
*Notes:* This figure shows the Prolific experiment design. The groups G1, G2, G3, G4 denote the experimental groups. I randomized the candidate's gender (conveyed through her first name) and whether participants can review their candidate answers before assessing their ability in the new task or whether 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 A40 and A41. The decision screen for the recall task is shown in Figure A42.

**Figure 10:** Prolific Experiment: Assessment of Ability in New Task and Recall of Performance in Observed Task

**Panel (a):** Estimated share correct questions in the new task (left) and share of correct questions in the observed task among the questions recalled (right)

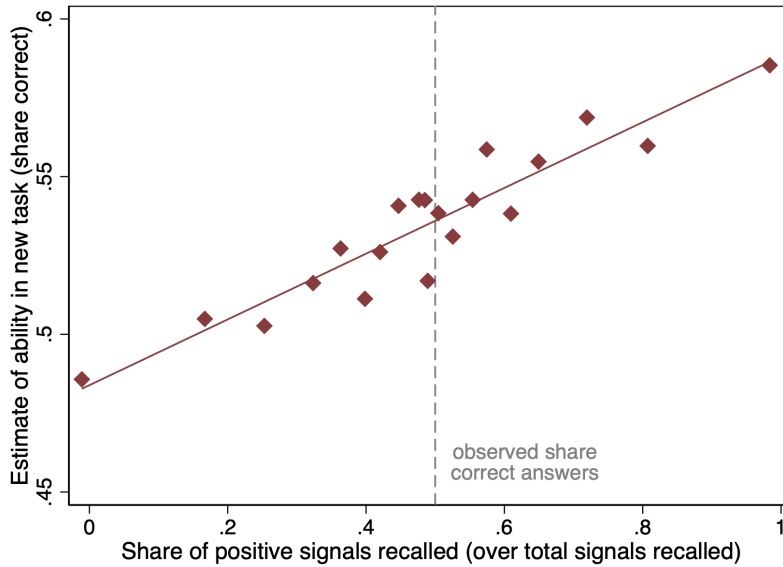


**Panel (b):** Number of correct minus incorrect questions recalled by candidate's gender and domain (*Memory* treatment)



*Notes:* The figures in Panel (a) show the gender gap in the assessment of ability (left) and in the recalled share of correct answers in the observed task (right) for participants observing a stereotype-consistent vs. inconsistent candidate in the memory and baseline conditions. I stack the two domains so that one observation is the assessment of a candidate in a domain (there are two observations per candidate, one per domain). The male candidate ("John") is the stereotype-consistent candidate in sports, while the female candidate ("Susan") is the stereotype-consistent candidate in pop culture. Panel (b) shows the number of correct minus incorrect questions recalled by candidate's gender and domain. The sample includes the 1239 US participants recruited through Prolific.

**Figure 11:** Prolific Experiment: Estimation of ability in new task and biases in signals recalled



*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  $\beta_1$  coefficient from estimating the following equation, where  $i$  is a candidate evaluated by  $j$  in domain  $d$ :  $Y_{ijd} = \beta_0 + \beta_1 \text{Recalled Share Correct}_{ijd} + \beta_2 \text{Tot. Recall}_{ijd} + \mu_d + \nu_j + \varepsilon_{ij}$ .

## 10 Tables

**Table 1:** Number of Students to Recommend and Top-Scientific Track Recommendation

|   | DV: Scientific Track Recommendation |                         |                        |                        |                        |                        |
|---|-------------------------------------|-------------------------|------------------------|------------------------|------------------------|------------------------|
|   | (1)                                 | (2)                     | (3)                    | (4)                    | (5)                    | (6)                    |
| Female                                    | -0.0334***<br>(0.00656)             | -0.0229***<br>(0.00629) | 0.0240*<br>(0.0127)    | 0.0228*<br>(0.0127)    | 0.0158<br>(0.0124)     | 0.0834*<br>(0.0445)    |
| 1(20-24 Stud. to Recommend)               |                                     | 0.0124<br>(0.0131)      | 0.0303**<br>(0.0140)   | 0.0268**<br>(0.0133)   | 0.0218<br>(0.0134)     | 0.0215<br>(0.0134)     |
| 1(25-29 Stud. to Recommend)               |                                     | -0.00480<br>(0.0171)    | 0.0407**<br>(0.0196)   | 0.0344*<br>(0.0195)    | 0.0329*<br>(0.0191)    | 0.0221<br>(0.0195)     |
| 1(30-48 Stud. to Recommend)               |                                     | 0.00816<br>(0.0252)     | 0.0570*<br>(0.0294)    | 0.0390<br>(0.0287)     | 0.0386<br>(0.0269)     | 0.0359<br>(0.0281)     |
| 1(20-24 Stud. to Recommend) × Female      |                                     |                         | -0.0362**<br>(0.0151)  | -0.0354**<br>(0.0149)  | -0.0281*<br>(0.0146)   | -0.0235<br>(0.0152)    |
| 1(25-29 Stud. to Recommend) × Female      |                                     |                         | -0.0927***<br>(0.0198) | -0.0917***<br>(0.0196) | -0.0873***<br>(0.0193) | -0.0656***<br>(0.0203) |
| 1(30-48 Stud. to Recommend) × Female      |                                     |                         | -0.0984***<br>(0.0309) | -0.0937***<br>(0.0307) | -0.0853***<br>(0.0301) | -0.0876***<br>(0.0303) |
| Mean DV Males                             | 0.205                               | 0.224                   | 0.224                  | 0.224                  | 0.224                  | 0.224                  |
| Female as %                               | -16.271%                            | -10.218%                |                        |                        |                        |                        |
| # Stud. to Recommend × Female as %        |                                     |                         | -2.073%                | -1.875%                | -1.875%                | -1.513%                |
| 1(20-24 Stud. to Recommend) × Female as % |                                     |                         | -16.142%               | -15.777%               | -12.504%               | -10.455%               |
| 1(25-29 Stud. to Recommend) × Female as % |                                     |                         | -41.274%               | -40.855%               | -38.892%               | -29.232%               |
| 1(30-48 Stud. to Recommend) × Female as % |                                     |                         | -43.822%               | -41.719%               | -37.976%               | -38.999%               |
| R-squared                                 | 0.00181                             | 0.322                   | 0.323                  | 0.331                  | 0.354                  | 0.335                  |
| # Students                                | 18123                               | 16486                   | 16486                  | 16486                  | 16486                  | 16486                  |
| # Teachers                                | 316                                 | 316                     | 316                    | 316                    | 316                    | 316                    |
| Math Teacher FE                           |                                     | ✓                       | ✓                      | ✓                      | ✓                      | ✓                      |
| Year FE                                   |                                     | ✓                       | ✓                      | ✓                      | ✓                      | ✓                      |
| Std.test scores                           |                                     | ✓                       | ✓                      | ✓                      | ✓                      | ✓                      |
| Stud. Controls                            |                                     |                         |                        | ✓                      | ✓                      | ✓                      |
| Quality Classmates                        |                                     |                         |                        | ✓                      | ✓                      | ✓                      |
| Squared std. test scores                  |                                     |                         |                        |                        | ✓                      | ✓                      |
| All controls × Female                     |                                     |                         |                        |                        |                        | ✓                      |

*Notes:* This table shows coefficients  $\gamma_s$  and  $\beta_s$  from estimation of model 6. One observation is a student assigned to a math teacher in a given year. The dependent variable is a dummy equal to 1 if the student is recommended for the scientific track. The indicator variables measure bins of the total number of students that the math teacher has to recommend in a given year. The sample includes the students matched with their math teachers from the main observational sample (same sample as [Carlana \(2019\)](#)). Students' controls include students' standardized math and Italian test scores, students' mother education dummies, students' father occupation dummies, immigrant status, class size, 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 2:** Gender gap in scientific track recommendations for teachers who report checking students' performance vs. relying on memory

|                               | DV: Scientific Track Recommendation |                       |                       |                       |                       |                       |
|-------------------------------|-------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|                               | (1)                                 | (2)                   | (3)                   | (4)                   | (5)                   | (6)                   |
| Female                        | -0.0499***<br>(0.0108)              | -0.0214<br>(0.0147)   | -0.0201<br>(0.0147)   | -0.0202<br>(0.0147)   | -0.0167<br>(0.0154)   | -0.00848<br>(0.0568)  |
| Memory                        |                                     | 0.0373**<br>(0.0168)  | 0.0325**<br>(0.0160)  | 0.0327**<br>(0.0161)  | 0.0253**<br>(0.0109)  | 0.0243**<br>(0.0103)  |
| Memory $\times$ Female        |                                     | -0.0457**<br>(0.0202) | -0.0471**<br>(0.0202) | -0.0469**<br>(0.0201) | -0.0495**<br>(0.0212) | -0.0488**<br>(0.0203) |
| Mean control                  | 0.202                               | 0.202                 | 0.202                 | 0.202                 | 0.202                 | 0.202                 |
| Female as a %                 | -24.721%                            |                       |                       |                       |                       |                       |
| Memory $\times$ Female as a % |                                     | -22.638%              | -23.316%              | -23.235%              | -24.505%              | -24.200%              |
| Observations                  | 8584                                | 8584                  | 8584                  | 8584                  | 8584                  | 8584                  |
| teachers                      | 170                                 | 170                   | 170                   | 170                   | 170                   | 170                   |
| R <sup>2</sup>                | 0.188                               | 0.189                 | 0.197                 | 0.197                 | 0.258                 | 0.277                 |
| Year FE                       | ✓                                   | ✓                     | ✓                     | ✓                     | ✓                     | ✓                     |
| std. Test Scores              | ✓                                   | ✓                     | ✓                     | ✓                     | ✓                     | ✓                     |
| Stud. Controls                | ✓                                   | ✓                     | ✓                     | ✓                     | ✓                     | ✓                     |
| Teacher Controls              |                                     |                       | ✓                     | ✓                     | ✓                     | ✓                     |
| IAT                           |                                     |                       |                       | ✓                     | ✓                     | ✓                     |
| Class FE                      |                                     |                       |                       |                       | ✓                     | ✓                     |
| Squared Std. Test Scores      |                                     |                       |                       |                       |                       | ✓                     |
| All Controls $\times$ Female  |                                     |                       |                       |                       |                       | ✓                     |

*Notes:* This figure reports coefficients  $\beta_1, \beta_2, \beta_3$  from the estimation of model 8. 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.



**Table 3:** Teachers' Experiment: Balance

| Variable  | (1)<br>Control | (2)<br>SD | (3)<br>Treated | (4)<br>SD | (5)<br>Diff. | (6)<br>P-val. |
|---|----------------|-----------|----------------|-----------|--------------|---------------|
| <b>Teachers background</b>                        |                |           |                |           |              |               |
| Female  | 0.838          | (0.369)   | 0.816          | (0.388)   | -0.022       | (0.544)       |
| School in Northern Italy                          | 0.730          | (0.445)   | 0.715          | (0.453)   | -0.015       | (0.719)       |
| Age   | 56.324         | (127.571) | 57.879         | (137.605) | 1.556        | (0.901)       |
| Highest level of education: PhD                   | 0.110          | (0.313)   | 0.100          | (0.301)   | -0.010       | (0.753)       |
| Highest level of education: Master                | 0.840          | (0.367)   | 0.858          | (0.350)   | 0.018        | (0.619)       |
| Highest level of education: BA or less            | 0.050          | (0.219)   | 0.042          | (0.201)   | -0.008       | (0.698)       |
| Married/cohabitant                                | 0.717          | (0.452)   | 0.802          | (0.399)   | 0.085        | (0.045)**     |
| Low education mother                              | 0.243          | (0.430)   | 0.204          | (0.404)   | -0.039       | (0.348)       |
| Years in teaching                                 | 16.758         | (10.337)  | 17.020         | (10.697)  | 0.261        | (0.794)       |
| Permanent contract                                | 0.808          | (0.395)   | 0.811          | (0.392)   | 0.003        | (0.930)       |
| Humanistic subject                                | 0.477          | (0.501)   | 0.498          | (0.501)   | 0.020        | (0.667)       |
| Scientific subject                                | 0.290          | (0.455)   | 0.300          | (0.459)   | 0.009        | (0.834)       |
| Other subject                                     | 0.228          | (0.421)   | 0.193          | (0.396)   | -0.035       | (0.368)       |
| Viceprincipal as role in the school               | 0.033          | (0.180)   | 0.029          | (0.168)   | -0.004       | (0.799)       |
| Responsible for orientation as role in the school | 0.066          | (0.249)   | 0.077          | (0.268)   | 0.011        | (0.656)       |
| <b>Attitudes</b>                                  |                |           |                |           |              |               |
| Ability to motivate difficult students (0-5)      | 3.702          | (0.839)   | 3.653          | (0.818)   | -0.049       | (0.547)       |
| Importance of Grades                              | 46.217         | (16.502)  | 45.680         | (17.728)  | -0.537       | (0.750)       |
| Importance of Attitudes                           | 34.167         | (12.864)  | 35.634         | (14.759)  | 1.467        | (0.280)       |
| Importance of Parents                             | 19.615         | (19.723)  | 18.686         | (18.977)  | -0.930       | (0.626)       |
| IAT   | 0.391          | (0.766)   | 0.375          | (0.708)   | -0.015       | (0.846)       |
| Remember students' performance                    | 0.412          | (0.493)   | 0.430          | (0.496)   | 0.018        | (0.708)       |
| Check grades during the teaching meeting          | 0.344          | (0.476)   | 0.295          | (0.457)   | -0.049       | (0.292)       |
| Do not check grades but think more holistically   | 0.656          | (0.476)   | 0.632          | (0.483)   | -0.024       | (0.612)       |
| Observations                                      | 241            |           | 207            |           | 448          |               |

*Notes:* This table shows a balance of observable characteristics for teachers who completed the experiment.

**Table 4:** Teacher experiment: gender gaps in high school track recommendations and memory constraints

|                              | DV: Scientific or Technical |                       |                       |                       | DV: Top Classical or Medium Humanities |                       |                       |                      |
|------------------------------|-----------------------------|-----------------------|-----------------------|-----------------------|--|-----------------------|-----------------------|----------------------|
|                              | (1)                         | (2)                   | (3)                   | (4)                   | (5)                                    | (6)                   | (7)                   | (8)                  |
| Memory $\times$ Female       | -0.186***<br>(0.0395)       | -0.185***<br>(0.0394) | -0.220***<br>(0.0454) | -0.189***<br>(0.0408) | 0.175***<br>(0.0397)                   | 0.169***<br>(0.0396)  | 0.194***<br>(0.0459)  | 0.163***<br>(0.0415) |
| Female                       | -0.0405**<br>(0.0200)       | -0.0418**<br>(0.0204) | -0.0512**<br>(0.0238) | -0.0494**<br>(0.0240) | 0.0463**<br>(0.0186)                   | 0.0496***<br>(0.0186) | 0.0665***<br>(0.0220) | 0.0593**<br>(0.0231) |
| Memory                       | 0.0780***<br>(0.0275)       | 0.0787***<br>(0.0275) |                       |                       | -0.115***<br>(0.0260)                  | -0.111***<br>(0.0259) |                       |                      |
| Mean control (Baseline-Male) | 0.492                       | 0.492                 | 0.492                 | 0.460                 | 0.471                                  | 0.471                 | 0.471                 | 0.441                |
| Memory $\times$ Female as %  | -37.733%                    | -37.649%              | -44.753%              | -41.048%              | 37.111%                                | 35.960%               | 41.237%               | 37.082%              |
| R <sup>2</sup>               | 0.303                       | 0.308                 | 0.457                 | 0.471                 | 0.353                                  | 0.359                 | 0.481                 | 0.486                |
| Observations                 | 1761                        | 1761                  | 1757                  | 2012                  | 1761                                   | 1761                  | 1757                  | 2012                 |
| N. teachers                  | 448                         | 448                   | 444                   | 503                   | 448                                    | 448                   | 444                   | 503                  |
| Student FE                   | Yes                         | Yes                   | Yes                   | Yes                   | Yes                                    | Yes                   | Yes                   | Yes                  |
| Controls                     | No                          | Yes                   | Yes                   | Yes                   | No                                     | Yes                   | Yes                   | Yes                  |
| Teacher FE                   | No                          | No                    | Yes                   | Yes                   | No                                     | No                    | Yes                   | Yes                  |

*Notes:* This table shows coefficients  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  from estimation of equations 9 and 10. One observation is a student profile assigned to a teacher (each teacher observes 4 student profiles). Columns 1 to 3, and 5 to 7 include the baseline sample of teachers. Columns 4 and 7 also include the teachers who did not provide any recommendations. 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. Columns 3-4 and 7-8 include teacher fixed effects. Standard errors are clustered at the teacher level. \* $p < 0.1$ ; \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

**Table 5:** Teacher Experiment: Biased recall of students' characteristics

|                             | 1(Recall a female typed characteristics) |                       |                      |                     | 1(Recall a male typed characteristics) |                       |                     |                      |
|-----------------------------|--|-----------------------|----------------------|---------------------|--|-----------------------|---------------------|----------------------|
|                             | (1)                                      | (2)                   | (3)                  | (4)                 | (5)                                    | (6)                   | (7)                 | (8)                  |
| Memory $\times$ Female      | 0.0953**<br>(0.0371)                     | 0.0936**<br>(0.0366)  | 0.0809**<br>(0.0395) | 0.106**<br>(0.0460) | -0.0219<br>(0.0388)                    | -0.0251<br>(0.0382)   | -0.0558<br>(0.0401) | -0.122**<br>(0.0506) |
| Female                      | -0.0124<br>(0.0241)                      | -0.0159<br>(0.0236)   | -0.00222<br>(0.0244) | -0.0108<br>(0.0226) | -0.0130<br>(0.0290)                    | -0.0174<br>(0.0283)   | 0.00114<br>(0.0278) | 0.0117<br>(0.0278)   |
| Memory                      | -0.260***<br>(0.0271)                    | -0.259***<br>(0.0265) |                      |                     | -0.313***<br>(0.0302)                  | -0.312***<br>(0.0295) |                     |                      |
| Mean control                | 0.504                                    | 0.504                 | 0.504                | 0.544               | 0.590                                  | 0.590                 | 0.591               | 0.633                |
| Memory $\times$ Female as % | 18.922%                                  | 18.578%               | 16.054%              | 19.524%             | -3.715%                                | -4.245%               | -9.436%             | -19.224%             |
| R <sup>2</sup>              | 0.351                                    | 0.355                 | 0.560                | 0.603               | 0.277                                  | 0.289                 | 0.546               | 0.537                |
| Observations                | 1761                                     | 1761                  | 1757                 | 1404                | 1761                                   | 1761                  | 1757                | 1404                 |
| N. teachers                 | 448                                      | 448                   | 444                  | 382                 | 448                                    | 448                   | 444                 | 382                  |
| Student FE                  | Yes                                      | Yes                   | Yes                  | Yes                 | Yes                                    | Yes                   | Yes                 | Yes                  |
| Teacher Controls            | No                                       | Yes                   | Yes                  | Yes                 | No                                     | Yes                   | Yes                 | Yes                  |
| Teacher FE                  | No                                       | No                    | Yes                  | Yes                 | No                                     | No                    | Yes                 | Yes                  |

*Notes:* This table shows coefficients  $\beta_2$  and  $\beta_3$  from the estimation of equations 9 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). Columns 3 and 6 include teacher fixed effects. Teachers in the baseline sample are included. 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. \* $p < 0.1$ ; \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

**Table 6:** Prolific Experiment: Assessment of ability in new task

|  | DV: Estimated share correct new task |                        |                        |                        |                        |                        |
|--|--------------------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
|  | (1)                                  | (2)                    | (3)                    | (4)                    | (5)                    | (6)                    |
| Stereotype-Consistent $\times$ Memory      | 0.0307***<br>(0.00826)               | 0.0300***<br>(0.00806) | 0.0300***<br>(0.00800) | 0.0268***<br>(0.00928) | 0.0258***<br>(0.00905) | 0.0258***<br>(0.00897) |
| Stereotype-Consistent                      | 0.0168***<br>(0.00498)               | 0.0175***<br>(0.00496) | 0.0175***<br>(0.00493) | 0.0156***<br>(0.00545) | 0.0167***<br>(0.00542) | 0.0167***<br>(0.00538) |
| Memory                                     | -0.00845<br>(0.00665)                | -0.00752<br>(0.00665)  |                        | -0.00791<br>(0.00751)  | -0.00554<br>(0.00742)  |                        |
| Control Mean                               | 0.520                                | 0.520                  | 0.520                  | 0.519                  | 0.519                  | 0.519                  |
| Stereotype-Consistent $\times$ Memory as % | 5.915%                               | 5.781%                 | 5.781%                 | 5.155%                 | 4.965%                 | 4.965%                 |
| R-squared                                  | 0.0250                               | 0.0790                 | 0.630                  | 0.0193                 | 0.0736                 | 0.629                  |
| Obs.                                       | 2478                                 | 2478                   | 2478                   | 2010                   | 2010                   | 2010                   |
| # Evaluators                               | 1239                                 | 1239                   | 1239                   | 1005                   | 1005                   | 1005                   |
| domain FE                                  | No                                   | Yes                    | Yes                    | No                     | Yes                    | Yes                    |
| controls                                   | No                                   | Yes                    | Yes                    | No                     | Yes                    | Yes                    |
| evaluator FE                               | No                                   | No                     | Yes                    | No                     | No                     | Yes                    |
| Sample                                     | All                                  | All                    | All                    | No pilot               | No pilot               | No pilot               |

*Notes:* This table shows coefficients  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  from the estimation of equation 11, where the dependent variable is evaluators' assessment of candidate ability in a domain. One observation is a candidate evaluated in a domain (either sports or pop culture) by an evaluator. Each participant in the experiment (participants are "evaluators") evaluates one candidate in sports and pop culture. Memory is a dummy equal to 1 for evaluators in the memory treatment, while stereotype-consistent is a dummy equal to 1 if the candidate is a female and the domain is pop culture, or if the candidate is male and the domain is sports. The first three columns present results including all participants (1005 from the main experiment and 234 from the pilot), while columns 3 to 6 present results excluding the pilot. Controls include evaluators' gender, age, education, employment status, political affiliation, number of rejections and approvals on Prolific, and time spent on the survey. Columns 2 and 5 include domain fixed effects and controls, while columns 3 and 6 add evaluator fixed effects. The sample includes 1239 US survey participants recruited through Prolific. Standard errors are clustered at the evaluator level. \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**Table 7:** Prolific Experiment: Types of Recalled Questions on Observed Task

|  | (1)                    | (2)                    | (3)                   | (4)                    | (5)                    | (6)                  |
|--|------------------------|------------------------|-----------------------|------------------------|------------------------|----------------------|
| <b>Panel A: Recalled share of correct questions</b>    |                        |                        |                       |                        |                        |                      |
| Stereotype-Consistent × Memory                         | 0.0517***<br>(0.0194)  | 0.0520***<br>(0.0196)  | 0.0525***<br>(0.0200) | 0.0488**<br>(0.0215)   | 0.0487**<br>(0.0217)   | 0.0495**<br>(0.0221) |
| Stereotype-Consistent                                  | -0.0101<br>(0.00759)   | -0.0101<br>(0.00766)   | -0.00913<br>(0.00769) | -0.0113<br>(0.00885)   | -0.0113<br>(0.00895)   | -0.0100<br>(0.00894) |
| Memory   | -0.0409***<br>(0.0116) | -0.0406***<br>(0.0117) |                       | -0.0458***<br>(0.0130) | -0.0449***<br>(0.0131) |                      |
| Mean Control   | 0.506                  | 0.506                  | 0.506                 | 0.506                  | 0.506                  | 0.506                |
| Stereotype-Consistent × Memory as %                    | 10.224%                | 10.286%                | 10.376%               | 9.650%                 | 9.625%                 | 9.768%               |
| R-squared  | 0.00703                | 0.0198                 | 0.331                 | 0.00757                | 0.0215                 | 0.352                |
| N. Obs   | 2242                   | 2242                   | 2178                  | 1811                   | 1811                   | 1760                 |
| <b>Panel B: Recalled number of correct questions</b>   |                        |                        |                       |                        |                        |                      |
| Stereotype-Consistent × Memory                         | 0.0355<br>(0.0685)     | 0.0385<br>(0.0684)     | 0.0385<br>(0.0679)    | -0.0122<br>(0.0776)    | -0.00666<br>(0.0770)   | -0.00666<br>(0.0764) |
| Stereotype-Consistent                                  | 0.0153<br>(0.0402)     | 0.0123<br>(0.0400)     | 0.0123<br>(0.0397)    | 0.0360<br>(0.0442)     | 0.0300<br>(0.0438)     | 0.0300<br>(0.0434)   |
| Memory   | -2.393***<br>(0.0892)  | -2.394***<br>(0.0884)  |                       | -2.334***<br>(0.100)   | -2.334***<br>(0.0994)  |                      |
| Mean Control   | 4.205                  | 4.205                  | 4.205                 | 4.146                  | 4.146                  | 4.146                |
| Stereotype-Consistent × Memory as %                    | 0.844%                 | 0.916%                 | 0.916%                | -0.295%                | -0.161%                | -0.161%              |
| R-squared  | 0.367                  | 0.420                  | 0.904                 | 0.353                  | 0.409                  | 0.905                |
| N. Obs   | 2478                   | 2478                   | 2478                  | 2010                   | 2010                   | 2010                 |
| <b>Panel C: Recalled number of incorrect questions</b> |                        |                        |                       |                        |                        |                      |
| Stereotype-Consistent × Memory                         | -0.172***<br>(0.0662)  | -0.169**<br>(0.0664)   | -0.169**<br>(0.0659)  | -0.189**<br>(0.0742)   | -0.186**<br>(0.0746)   | -0.186**<br>(0.0739) |
| Stereotype-Consistent                                  | 0.0424<br>(0.0365)     | 0.0402<br>(0.0368)     | 0.0402<br>(0.0365)    | 0.0720*<br>(0.0397)    | 0.0689*<br>(0.0399)    | 0.0689*<br>(0.0396)  |
| Memory   | -2.119***<br>(0.0926)  | -2.118***<br>(0.0906)  |                       | -2.025***<br>(0.104)   | -2.029***<br>(0.102)   |                      |
| Control Mean   | 4.183                  | 4.183                  | 4.183                 | 4.124                  | 4.124                  | 4.124                |
| Stereotype-Consistent × Memory as %                    | -4.104%                | -4.051%                | -4.051%               | -4.579%                | -4.508%                | -4.508%              |
| R-squared  | 0.319                  | 0.373                  | 0.908                 | 0.295                  | 0.354                  | 0.910                |
| Obs.   | 2478                   | 2478                   | 2478                  | 2010                   | 2010                   | 2010                 |
| domain FE  | Yes                    | Yes                    | Yes                   | Yes                    | Yes                    | Yes                  |
| controls   | No                     | Yes                    | Yes                   | No                     | Yes                    | Yes                  |
| evaluator FE   | No                     | No                     | Yes                   | No                     | No                     | Yes                  |
| Sample   | All                    | All                    | All                   | No pilot               | No pilot               | No pilot             |

*Notes:* This table shows coefficients  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  from the estimation of equation 11, where the dependent variables are the recalled share of perceived correct questions (Panel (a)), the number of perceived correct questions (Panel (b)), and the number of perceived incorrect questions (Panel (c)). One observation is a candidate evaluated in a domain (either sports or pop culture) by an evaluator. Each participant in the experiment (participants are "evaluators") evaluates one candidate in sports and pop culture. Memory is a dummy equal to 1 for evaluators in the memory treatment, while stereotype-consistent is a dummy equal to 1 if the candidate is a female and the domain is pop culture, or if the candidate is male and the domain is sports. The first three columns present results including all participants (1005 from the main experiment and 234 from the pilot), while columns 3 to 6 present results excluding the pilot. Controls include evaluators' gender, age, education, employment status, political affiliation, number of rejections and approvals on Prolific, and time spent on the survey. Columns 2 and 5 include domain fixed effects and controls, while columns 3 and 6 add evaluator fixed effects. The sample includes 1239 US survey participants recruited through Prolific. Standard errors are clustered at the evaluator level. \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

# Appendix

## Table of Contents

---

|          |  |            |
|----------|--|------------|
| <b>A</b> | <b>Model Appendix</b>  | <b>63</b>  |
| A.1      | Derivations for Baseline Model . . . . .   | 63         |
| A.2      | More general similarity function (adding more features) . . . . .                    | 65         |
| <b>B</b> | <b>Additional Figures and Tables for First Result on Past Track Recommendations</b>  | <b>68</b>  |
| B.1      | Summary Statistics . . . . .   | 68         |
| B.2      | Main Result with Linear Relationship . . . . .                                       | 75         |
| B.3      | Raw Trends . . . . .   | 76         |
| B.4      | Gender gap in scientific track recommendations . . . . .                             | 77         |
| B.5      | Heterogeneity . . . . .  | 78         |
| B.6      | Sensitivity and Robustness . . . . .   | 82         |
| B.7      | Alternative Explanations . . . . .   | 83         |
| <b>C</b> | <b>Additional Figures and Tables for Second Result on Past Track Recommendations</b> | <b>88</b>  |
| C.1      | Summary Statistics . . . . .   | 88         |
| C.2      | Recommendation Process . . . . .   | 90         |
| C.3      | Characteristics of Teachers who Check vs. Rely on Memory . . . . .                   | 91         |
| C.4      | Alternative Comparison Groups . . . . .  | 92         |
| C.5      | Sensitivity and Robustness . . . . .   | 93         |
| C.6      | Where are the missing "scientific" girls sent? . . . . .                             | 93         |
| <b>D</b> | <b>Additional Figures and Tables for Teachers Experiment</b>                         | <b>94</b>  |
| D.1      | Experiment Design . . . . .  | 94         |
| D.2      | Recommendations to all tracks . . . . .  | 96         |
| D.3      | Heterogeneity . . . . .  | 98         |
| D.4      | Recommendations for past students with same grades as student profiles .             | 101        |
| D.5      | Results by Student Profile . . . . .   | 103        |
| D.6      | Biases in Recall . . . . .   | 112        |
| <b>E</b> | <b>Additional Figures and Tables for Prolific Experiment</b>                         | <b>118</b> |

|          |  |            |
|----------|--|------------|
| E.1      | Design . . . . .   | 118        |
| E.2      | Balance . . . . .  | 121        |
| E.3      | Sensitivity . . . . .                                    | 122        |
| E.4      | Biases in Recall . . . . .                               | 123        |
| E.5      | Bias in Assessment of Ability by Domain . . . . .        | 126        |
| <b>F</b> | <b>Prior Beliefs in Teacher and Prolific Experiments</b> | <b>127</b> |
| F.1      | Prolific Experiment . . . . .                            | 127        |
| F.2      | Teachers Experiment . . . . .                            | 128        |
| <b>G</b> | <b>Data Appendix</b>                                     | <b>129</b> |
| G.1      | Teacher Survey and Experiment . . . . .                  | 129        |

---

## A Model Appendix

### A.1 Derivations for Baseline Model

**Sample space and features.** This table presents the types of signals in the teacher’s memory database:

**Table A1:** Types of signals in the teacher’s memory database

| Subject                   | Type of Signals | Set of Signals<br>in the memory database | Fraction of Signals<br>given Subject | Number                        |
|---------------------------|-----------------|--|--------------------------------------|-------------------------------|
| Student $i$ of gender $g$ | Good at math    | $H_i$                                    | $h_i$                                | $h_i \cdot K$                 |
| Student $i$ of gender $g$ | Bad at math     | $L_i$                                    | $1 - h_i$                            | $(1 - h_i) \cdot K$           |
| Other girls               | Good at math    | $H_f$                                    | $h_f$                                | $h_f \cdot K \cdot S/2$       |
| Other girls               | Bad at math     | $L_f$                                    | $1 - h_f$                            | $(1 - h_f) \cdot K \cdot S/2$ |
| Other boys                | Good at math    | $H_m$                                    | $h_m$                                | $h_m \cdot K \cdot S/2$       |
| Other boys                | Bad at math     | $L_m$                                    | $1 - h_m$                            | $(1 - h_m) \cdot K \cdot S/2$ |

**Recall function.** The probability that experiences in the set  $H_i$  (positive math experiences) are recalled given that Susan is observed is:

$$r(H_i|f) = \frac{\sum_{e \in H_i} S(e|f)}{\sum_{e \in E} S(e|f)} = \frac{\overbrace{h_i - \Delta_s h_i}^{\text{high math Susan}}}{\underbrace{1 - \Delta_s h_i}_{\text{high+low math Susan}} + \underbrace{\frac{N}{2}(1 - \Delta_s h_f)}_{\text{high+low girls}} + \underbrace{\frac{N}{2}(1 - \Delta_s(1 - h_m))}_{\text{high+low boys}}} \quad (12)$$

The probability of recalling positive experiences about Susan is increasing in the share of true positive experiences ( $h_i$ ), and decreasing in the number of other students S.

On the other hand, the probability that experiences in the set  $H_i$  (positive math experiences) are recalled given that John is observed is:

$$r(H_i|m) = \frac{\sum_{e \in H_i} S(e|m)}{\sum_{e \in E} S(e|m)} = \frac{\overbrace{h_i}^{\text{high math John}}}{\underbrace{1 - \Delta_s(1 - h_i)}_{\text{high+low math John}} + \underbrace{\frac{N}{2}(1 - \Delta_s h_f)}_{\text{high+low girls}} + \underbrace{\frac{N}{2}(1 - \Delta_s(1 - h_m))}_{\text{high+low boys}}} \quad (13)$$

**Assessment of ability.** In this stylized setting, the average number of positive math signals recalled if a girl is observed is  $R(H_i|f) = \sum_{e \in E} 1(e \in H_i) \cdot r(e|f) = r(H_i|f)$ . The average assessment of individual ability if the student is a girl is:

$$\begin{aligned} \hat{\pi}(i|f) &= \frac{(h_i - \Delta_s h_i) + \sigma_1 \frac{N}{2}(h_f - \Delta_s h_f) + \sigma_2 \frac{N}{2}(h_m)}{(1 - \Delta_s h_i) + \sigma_1 \frac{N}{2}(1 - \Delta_s h_f) + \sigma_2 \frac{N}{2}(1 - \Delta_s(1 - h_m))} \\ &= \frac{(1 - \Delta_s h_i)}{(1 - \Delta_s h_i) + \sigma_1 \frac{N}{2}(1 - \Delta_s h_f) + \sigma_2 \frac{N}{2}(1 - \Delta_s(1 - h_m))} \cdot \frac{(h_i - \Delta_s h_i)}{(1 - \Delta_s h_i)} + \\ &\quad \frac{\sigma_1 \frac{N}{2}(1 - \Delta_s h_f) + \sigma_2 \frac{N}{2}(1 - \Delta_s(1 - h_m))}{(1 - \Delta_s h_i) + \sigma_1 \frac{N}{2}(1 - \Delta_s h_f) + \sigma_2 \frac{N}{2}(1 - \Delta_s(1 - h_m))} \cdot \frac{\sigma_1(h_f - \Delta_s h_f) + \sigma_2(h_m)}{\sigma_1(1 - \Delta_s h_f) + \sigma_2(1 - \Delta_s(1 - h_m))} \end{aligned}$$

If a boy with the same observed ability  $h_i$  is assessed:

$$\begin{aligned} \hat{\pi}(i|m) &= \frac{(h_i) + \sigma_1 \frac{N}{2}(h_m) + \sigma_2 \frac{N}{2}(h_f - \Delta_s h_f)}{(1 - \Delta_s(1 - h_i)) + \sigma_1 \frac{N}{2}(1 - \Delta_s(1 - h_m)) + \sigma_2 \frac{N}{2}(1 - \Delta_s h_f)} \\ &= \frac{(1 - \Delta_s(1 - h_i))}{(1 - \Delta_s(1 - h_i)) + \sigma_1 \frac{N}{2}(1 - \Delta_s(1 - h_m)) + \sigma_2 \frac{N}{2}(1 - \Delta_s h_f)} \cdot \frac{h_i}{(1 - \Delta_s(1 - h_i))} + \\ &\quad \frac{\sigma_1 \frac{N}{2}(1 - \Delta_s(1 - h_m)) + \sigma_2 \frac{N}{2}(1 - \Delta_s h_f)}{(1 - \Delta_s(1 - h_i)) + \sigma_1 \frac{N}{2}(1 - \Delta_s(1 - h_m)) + \sigma_2 \frac{N}{2}(1 - \Delta_s h_f)} \cdot \frac{\sigma_1(h_m) + \sigma_2(1 - \Delta_s h_f)}{\sigma_1(1 - \Delta_s(1 - h_m)) + \sigma_2(1 - \Delta_s h_f)} \end{aligned}$$



Since  $h_i = 1/2$  and  $h_m = 1 - h_f$ , the weights are equal across gender:

$$\hat{\pi}(i|f) = \theta(N) \cdot \frac{(h_i - \Delta_s h_i)}{(1 - \Delta_s h_i)} + (1 - \theta(N)) \cdot \frac{\sigma_1(h_f - \Delta_s h_f) + \sigma_2(h_m)}{\sigma_1(1 - \Delta_s h_f) + \sigma_2(1 - \Delta_s(1 - h_m))}$$

If a boy with the same observed ability  $h_i$  is evaluated:

$$\hat{\pi}(i|m) = \theta(N) \cdot \frac{h_i}{(1 - \Delta_s(1 - h_i))} + (1 - \theta(N)) \cdot \frac{\sigma_1(h_m) + \sigma_2(h_f - \Delta_s h_f)}{\sigma_1(1 - \Delta_s(1 - h_m)) + \sigma_2(1 - \Delta_s h_f)}$$

Thus, if group stereotypes affect what decision-makers recall ( $\Delta_s > 0$ ), the decision-maker has a biased perception of signals favoring boys:  $\frac{h_i}{(1 - \Delta_s(1 - h_i))} > \frac{(h_i - \Delta_s h_i)}{(1 - \Delta_s h_i)}$ . If  $\Delta_s = 0$ , we go back to the standard models of discrimination where the perception of the signal  $h_i$  is the same across genders.

## A.2 More general similarity function (adding more features)

**Recall** Recall of experiences is governed by similarity and interference, as in [Bordalo et al. \(2023\)](#). Similarity captures the associative nature of memory. It is a symmetric function  $S: E \times E \rightarrow [0, 1]$ , and its logic works as follows. Every experience in the memory database is formed by a set of features. Two experiences are similar to each other if they share a higher number of features. For instance, a positive math experience with Susan is more similar to a positive math experience with Alice since they are both females, while it is less similar to math experiences with a male student.

We denote the similarity of any experience with the hypothesis "*Student i math ability/girl*" as  $S(e) \equiv S(e, \text{"Student i math ability/girl"})$ . Every experience  $e$  has three features: (i) whether experience is of student  $i$  or not, (ii) whether experience is from same gender student  $g$ , (iii) whether experience  $i$  is consistent or inconsistent with the stereotype "*girls bad at math, boys good at math*". We assume that similarity is 1 if two experiences share all the features, and decreases by  $\Delta_f$  for every feature that is not shared, with  $f \in \{i, g, s\}$  ((i)  $i$ =Susan/John, (ii)  $g$ =female/male, (iii)  $s$ =consistent/inconsistent). Thus, when thinking about Susan,  $S(e) = 1 - \Delta_s$  if  $e \in H_i$ ,  $S(e) = 1$  if  $e \in L_i$ ,  $S(e) = 1 - \Delta_i - \Delta_s$  if  $e \in H_f$ ,  $S(e) = 1 - \Delta_i - \Delta_g$  if  $e \in H_m$ .

The probability that experience  $e$  is recalled is:

$$r(e|f) = \frac{S(e|f)}{\sum_{u \in E} S(u|f)} \quad (14)$$

When thinking about "Student  $i$  math ability|girl",  $e$  is sampled more frequently when it is more similar to the hypothesis and when is consistent with the stereotype. The denominator captures interference (Jenkins and Dallenbach (1924), McGeoch (1932), Underwood (1957)): all experiences compete for retrieval and may inhibit the recall of  $e$ , especially experiences that are either more similar or more frequent in the database. If the teacher has many students to evaluate, experiences with other students interfere with the ones of student  $i$ , inhibiting their recall. If similarity is constant ( $\Delta_g = \Delta_i = \Delta_s = 0$ ), recall is frequency based, and every experience is recalled with probability  $r(e) = \frac{1}{10(S+1)}$ .

The probability of recalling positive math experiences when thinking about "Student  $i$  math ability|girl" is given by:

$$r(H_i|f) = \frac{\overbrace{h_i - \Delta_s h_i}^{\text{high math Susan}}}{\underbrace{1 - \Delta_s h_i}_{\text{high+low math Susan}} + \underbrace{\frac{N}{2}(1 - \Delta_i) - \frac{N}{2}\Delta_s h_f}_{\text{high+low girls}} + \underbrace{\frac{N}{2}(1 - \Delta_i - \Delta_g) - \frac{N}{2}\Delta_s(1 - h_m)}_{\text{high+low boys}}}$$

If student  $i$  is a boy, the probability of recalling positive math experiences is:

$$r(H_i|m) = \frac{\overbrace{h_i}^{\text{high math John}}}{\underbrace{1 - \Delta_s(1 - h_i)}_{\text{high+low math John}} + \underbrace{\frac{N}{2}(1 - \Delta_i) - \frac{N}{2}\Delta_s(1 - h_m)}_{\text{high+low boys}} + \underbrace{\frac{N}{2}(1 - \Delta_i - \Delta_g) - \frac{N}{2}\Delta_s(h_f)}_{\text{high+low girls}}}$$

**Belief Formation** If DM observes a girl, the average belief on individual ability is:

$$\hat{\pi}(i|f) = \left( \frac{1 - \Delta_s h_i}{1 - \Delta_s h_i + \sigma_1 \frac{N}{2}(1 - \Delta_i - \Delta_s h_f)} \right) \frac{h_i - \Delta_s h_i}{1 - \Delta_s h_i} + \left( \frac{\sigma_1 \frac{N}{2}(1 - \Delta_i - \Delta_s h_f)}{1 - \Delta_s h_i + \sigma_1 \frac{N}{2}(1 - \Delta_i - \Delta_s h_f)} \right) \frac{(1 - \Delta_i)h_f - \Delta_s h_f}{(1 - \Delta_i) - \Delta_s h_f}$$

If DM observes a boy, the average belief on individual ability is:

$$\hat{\pi}(i|m) = \left( \frac{1 - \Delta_s(1 - h_i)}{1 - \Delta_s(1 - h_i) + \sigma_1 \frac{N}{2}(1 - \Delta_i - \Delta_s(1 - h_m))} \right) \frac{h_i}{1 - \Delta_s(1 - h_i)} +$$

$$\left( \frac{\sigma_1 \frac{N}{2}(1 - \Delta_i - \Delta_s(1 - h_m))}{1 - \Delta_s(1 - h_i) + \sigma_1 \frac{N}{2}(1 - \Delta_i - \Delta_s(1 - h_m))} \right) \frac{(1 - \Delta_i)h_m}{(1 - \Delta_i) - \Delta_s(1 - h_m)}$$

If  $h_i=1/2$ , as in the Prolific experiment, and  $1 - h_m = h_f$  the weights are equal across gender. We proceed to compute  $D(N) = \hat{\pi}(i|m) - \hat{\pi}(i|f)$ , and we obtain the equation in the main text.

## B Additional Figures and Tables for First Result on Past Track Recommendations

### B.1 Summary Statistics

**Table A2:** Summary statistics of students and their math teachers in the main observational sample

| Variable   | Mean   | SD     | Min    | Max     | N      |
|--|--------|--------|--------|---------|--------|
| <b>Track Recommendation and Choice:</b>            |        |        |        |         |        |
| Top-tier scientific recommendation                 | 0.189  | 0.392  | 0.000  | 1.000   | 18,123 |
| Vocational track recommendation                    | 0.388  | 0.487  | 0.000  | 1.000   | 18,123 |
| Top-scientific choice                              | 0.276  | 0.447  | 0.000  | 1.000   | 15,361 |
| Top-classical choice                               | 0.051  | 0.221  | 0.000  | 1.000   | 15,361 |
| Technical choice                                   | 0.308  | 0.462  | 0.000  | 1.000   | 15,361 |
| Medium humanities choice                           | 0.209  | 0.406  | 0.000  | 1.000   | 15,361 |
| Vocational choice                                  | 0.157  | 0.364  | 0.000  | 1.000   | 15,361 |
| <b>Other Characteristics:</b>                      |        |        |        |         |        |
| Follow teacher recommendation: top-tier scientific | 0.733  | 0.443  | 0.000  | 1.000   | 18,123 |
| Std. test score Reading                            | 0.064  | 0.990  | -3.862 | 2.076   | 16,494 |
| Std. test score Math                               | 0.088  | 1.008  | -2.987 | 2.746   | 16,504 |
| High Education Mother (university)                 | 0.416  | 0.493  | 0.000  | 1.000   | 18,123 |
| High Occupation Mother                             | 0.172  | 0.377  | 0.000  | 1.000   | 18,123 |
| Immigrant  | 0.193  | 0.394  | 0.000  | 1.000   | 18,123 |
| Female   | 0.482  | 0.500  | 0.000  | 1.000   | 18,123 |
| Math grade 8                                       | 7.040  | 1.307  | 1.000  | 10.000  | 16,898 |
| Italian grade 8                                    | 7.085  | 1.147  | 1.000  | 10.000  | 16,903 |
| Number of Years with Math Teacher                  | 2.546  | 0.754  | 1.000  | 3.000   | 18,123 |
| Class size   | 22.246 | 2.670  | 12.000 | 29.000  | 18,123 |
| <b>Characteristics of Students' Math Teachers:</b> |        |        |        |         |        |
| Teacher: female                                    | 0.794  | 0.405  | 0.000  | 1.000   | 18,123 |
| Teacher: age                                       | 47.322 | 16.778 | 0.000  | 66.000  | 18,123 |
| Teacher: graduated with honors                     | 0.151  | 0.358  | 0.000  | 1.000   | 18,123 |
| Teacher: born in North                             | 0.600  | 0.490  | 0.000  | 1.000   | 18,123 |
| Teacher: permanent contract                        | 0.847  | 0.360  | 0.000  | 1.000   | 18,123 |
| Teacher: has children                              | 1.260  | 1.108  | 0.000  | 5.000   | 18,123 |
| Teacher: number of 8th grade students              | 23.545 | 5.632  | 12.000 | 48.000  | 18,123 |
| Teacher: number of students                        | 53.236 | 14.970 | 15.000 | 144.000 | 18,123 |

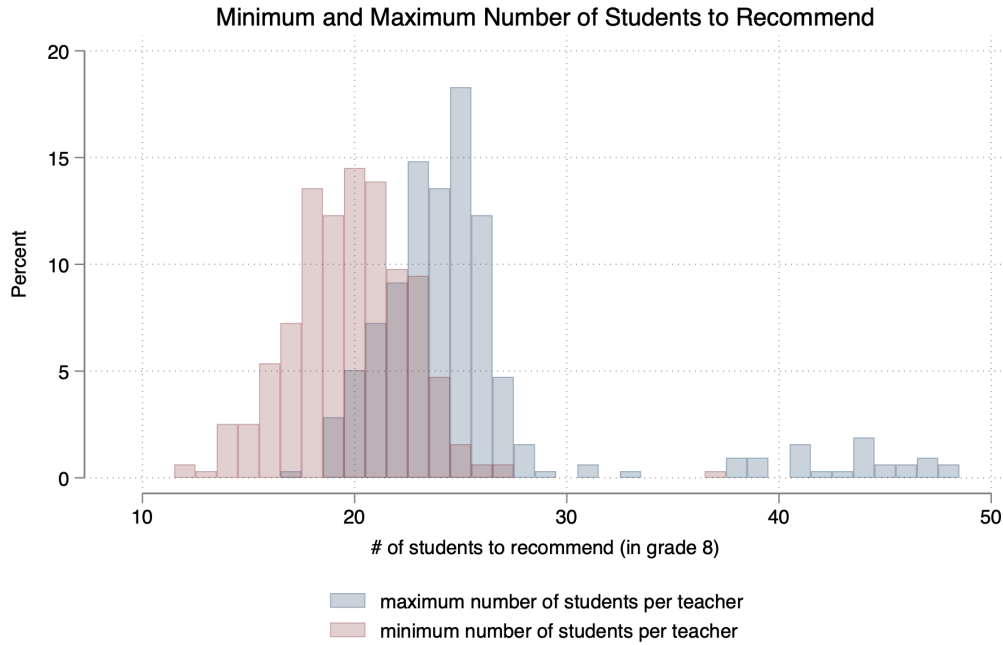
*Notes:* This table reports summary statistics about 8th-grade students and their assigned math teachers in the main administrative sample.

**Table A3:** Characteristics of students and students' teachers by student gender

| Variable   | (1)<br>Male | (2)<br>SD | (3)<br>Female | (4)<br>SD | (5)<br>Diff. | (6)<br>P-val. |
|--|-------------|-----------|---------------|-----------|--------------|---------------|
| <b>Students' Characteristics</b>                   |             |           |               |           |              |               |
| Top-tier scientific recommendation                 | 0.205       | (0.404)   | 0.172         | (0.377)   | -0.033       | (0.000)***    |
| Vocational track recommendation                    | 0.425       | (0.494)   | 0.349         | (0.477)   | -0.076       | (0.000)***    |
| Top-scientific choice                              | 0.321       | (0.467)   | 0.229         | (0.420)   | -0.093       | (0.000)***    |
| Top-classical choice                               | 0.034       | (0.181)   | 0.069         | (0.254)   | 0.035        | (0.000)***    |
| Technical choice                                   | 0.403       | (0.491)   | 0.209         | (0.406)   | -0.194       | (0.000)***    |
| Medium humanities choice                           | 0.091       | (0.287)   | 0.331         | (0.471)   | 0.240        | (0.000)***    |
| Vocational choice                                  | 0.151       | (0.358)   | 0.163         | (0.369)   | 0.012        | (0.046)**     |
| Follow teacher recommendation: top-tier scientific | 0.706       | (0.456)   | 0.762         | (0.426)   | 0.056        | (0.000)***    |
| Std. test score Reading                            | -0.029      | (1.001)   | 0.160         | (0.969)   | 0.189        | (0.000)***    |
| Std. test score Math                               | 0.182       | (1.018)   | -0.010        | (0.988)   | -0.192       | (0.000)***    |
| High Education Mother (university)                 | 0.412       | (0.492)   | 0.420         | (0.494)   | 0.008        | (0.283)       |
| High Occupation Mother                             | 0.168       | (0.374)   | 0.176         | (0.381)   | 0.007        | (0.201)       |
| Immigrant  | 0.194       | (0.395)   | 0.192         | (0.394)   | -0.002       | (0.767)       |
| Number of Years with Math Teacher                  | 2.532       | (0.761)   | 2.562         | (0.746)   | 0.030        | (0.007)***    |
| Mean math score classmates                         | 0.059       | (0.425)   | 0.056         | (0.426)   | -0.003       | (0.635)       |
| Mean reading score classmates                      | 0.041       | (0.406)   | 0.037         | (0.412)   | -0.004       | (0.489)       |
| Share immigrant classmates                         | 0.197       | (0.141)   | 0.195         | (0.142)   | -0.002       | (0.338)       |
| Share classmates with highly educated mother       | 0.410       | (0.229)   | 0.414         | (0.230)   | 0.003        | (0.367)       |
| Share classmates with high occupation father       | 0.169       | (0.160)   | 0.172         | (0.161)   | 0.003        | (0.179)       |
| <b>Students' Teachers' Characteristics</b>         |             |           |               |           |              |               |
| Teacher: female                                    | 0.793       | (0.405)   | 0.794         | (0.405)   | 0.001        | (0.928)       |
| Teacher: age                                       | 47.290      | (16.810)  | 47.357        | (16.744)  | 0.068        | (0.786)       |
| Teacher: graduated with honors                     | 0.155       | (0.362)   | 0.147         | (0.354)   | -0.008       | (0.134)       |
| Teacher: born in North                             | 0.601       | (0.490)   | 0.598         | (0.490)   | -0.003       | (0.659)       |
| Teacher: permanent contract                        | 0.845       | (0.362)   | 0.848         | (0.359)   | 0.003        | (0.549)       |
| Teacher: has children                              | 1.254       | (1.105)   | 1.267         | (1.111)   | 0.013        | (0.425)       |
| Teacher: number of 8th grade students              | 23.526      | (5.587)   | 23.564        | (5.680)   | 0.038        | (0.649)       |
| Teacher: number of students                        | 53.292      | (14.932)  | 53.176        | (15.010)  | -0.116       | (0.603)       |
| Observations                                       | 9,384       |           | 8,739         |           | 18,123       |               |

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

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

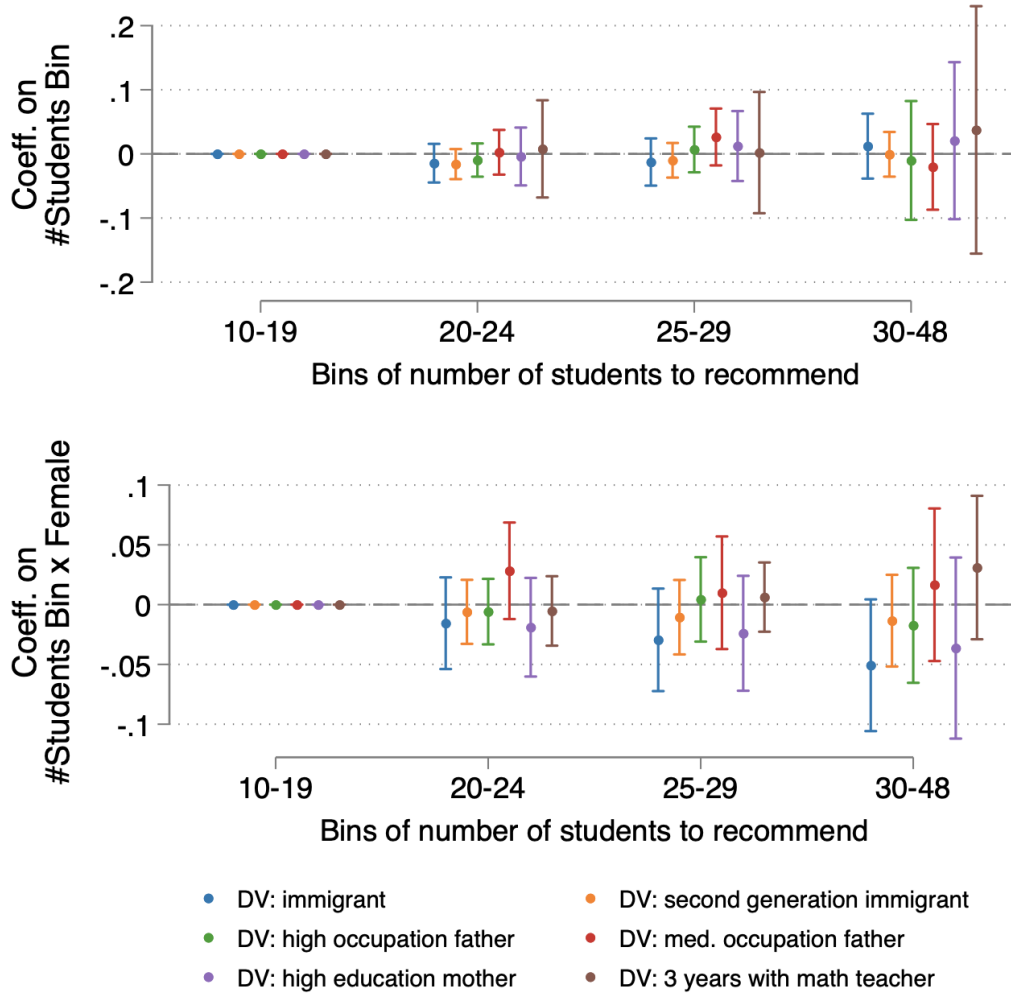


*Notes:* This figure shows the maximum and minimum numbers of 8th-grade students assigned to math teachers in the main observational sample. The average within-teachers range of students to be recommended by teachers across years is an additional 6 students.

**Table A4:** Scientific track recommendation and choice

|   | DV: top-scientific track choice |                         |                         |                         |
|---|---------------------------------|-------------------------|-------------------------|-------------------------|
|   | (1)                             | (2)                     | (3)                     | (4)                     |
| 1(Scientific track recommendation)          | 0.695***<br>(0.0121)            | 0.695***<br>(0.0121)    | 0.576***<br>(0.0156)    | 0.572***<br>(0.0157)    |
| 1(Scientific track recommendation) × Female | -0.00538<br>(0.0159)            | -0.00538<br>(0.0159)    | 0.0150<br>(0.0157)      | 0.0128<br>(0.0156)      |
| Female                                      | -0.0596***<br>(0.00698)         | -0.0596***<br>(0.00698) | -0.0568***<br>(0.00734) | -0.0561***<br>(0.00730) |
| R-squared                                   | 0.421                           | 0.421                   | 0.473                   | 0.477                   |
| N. Obs                                      | 15361                           | 15361                   | 14419                   | 14419                   |
| Teacher FE                                  |                                 | ✓                       | ✓                       | ✓                       |
| Year FE                                     |                                 | ✓                       | ✓                       | ✓                       |
| Std. test score math decile FE              |                                 |                         | ✓                       | ✓                       |
| Std. test score ita deciles FE              |                                 |                         | ✓                       | ✓                       |
| Students' controls                          |                                 |                         |                         | ✓                       |
| Mean DV                                     | 0.276                           | 0.276                   | 0.288                   | 0.288                   |
| Sd Dependent Variable                       | 0.447                           | 0.447                   | 0.453                   | 0.453                   |

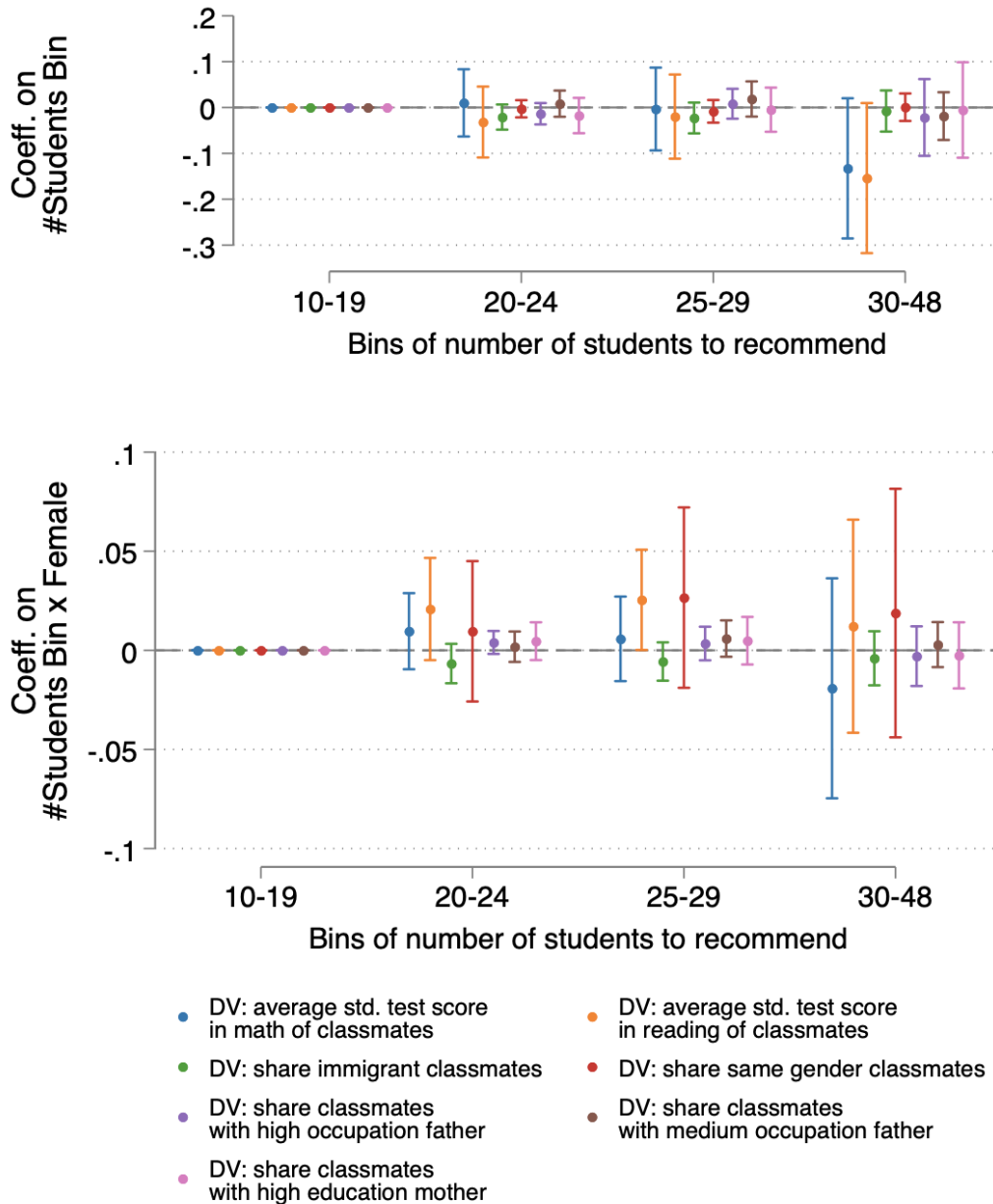
**Figure A2:** Observable characteristics of students and their teachers' number of students to recommend



*Notes:* This figure shows coefficients  $\beta_s$  (upper graph, the coefficients on  $1(\#Students Bin_{jt} = s)$ ) and  $\gamma_s$  (right graph, the coefficient on the interaction  $1(\#Students Bin_{jt} = s) \times Female_i$ ) from the estimation of equation 6 where the dependent variables are indicated on the y axis.

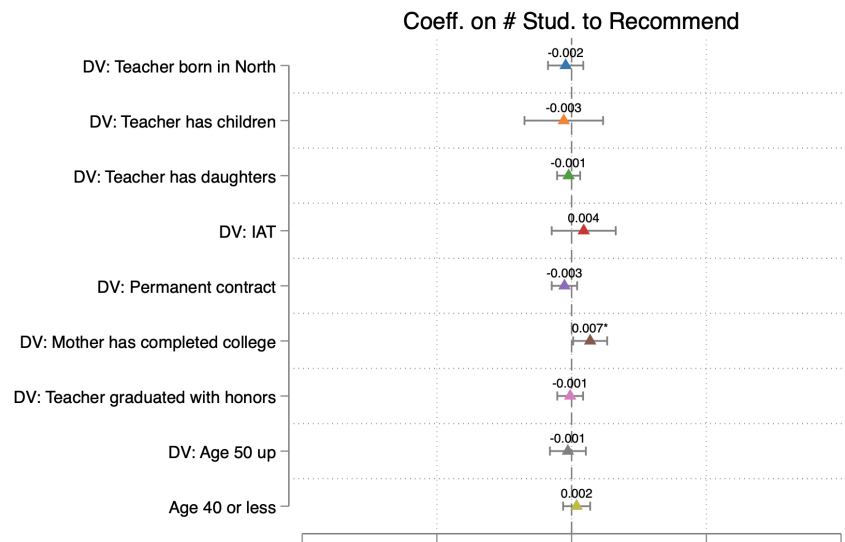


**Figure A3:** Observable characteristics of students' classmates and their teacher number of students to recommend



*Notes:* This figure shows coefficients  $\beta_s$  (upper graph, the coefficients on  $1(\#Students Bin_{jt} = s)$ ) and  $\gamma_s$  (right graph, the coefficient on the interaction  $1(\#Students Bin_{jt} = s) \times Female_i$ ) from the estimation of equation 6 where the dependent variables are indicated on the y axis.

**Figure A4:** Math teachers' characteristics and number of 8th grade students



*Notes:* The figure shows differences in teachers' characteristics in years of high and low evaluation load. It shows  $\beta_2$  coefficients from a regression where the dependent variable is math teacher's background characteristics and the independent variable is the number of 8th-grade students that the teacher needs to recommend. The sample includes math teachers in the main observational sample.

## B.2 Main Result with Linear Relationship

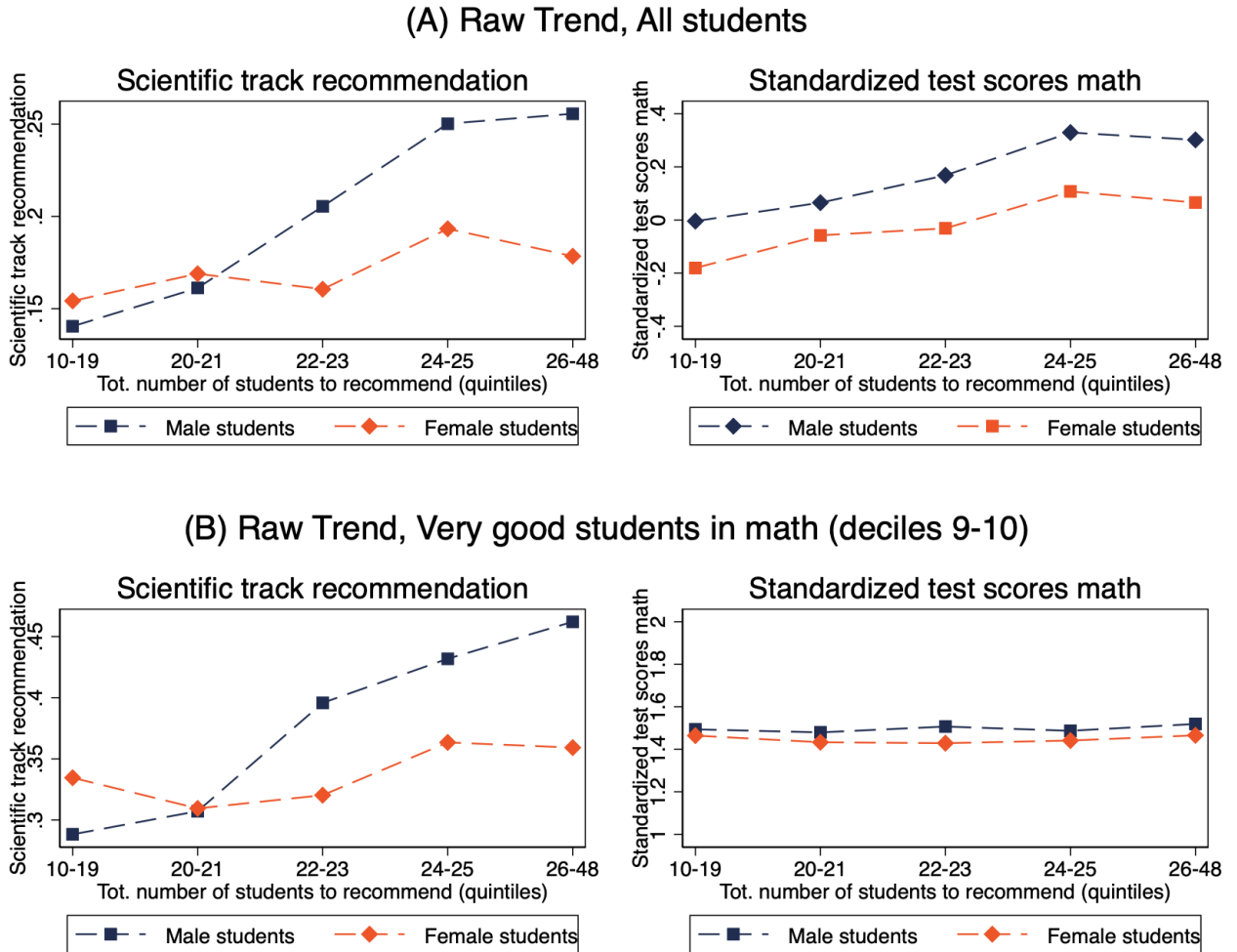
**Table A5:** Number of Students to Recommend and Top-Scientific Track Recommendation

|                                     | DV: top-scientific track recommendation |                         |                          |                          |                          |                          |                         |
|-------------------------------------|---|-------------------------|--------------------------|--------------------------|--------------------------|--------------------------|-------------------------|
|                                     | (1)                                     | (2)                     | (3)                      | (4)                      | (5)                      | (6)                      | (7)                     |
| Female                              | -0.0334***<br>(0.00656)                 | -0.0230***<br>(0.00628) | 0.0866***<br>(0.0286)    | 0.0835***<br>(0.0286)    | 0.0818***<br>(0.0286)    | 0.151***<br>(0.0533)     | 0.212***<br>(0.0668)    |
| # Students 8th                      |   | -0.000168<br>(0.000964) | 0.00220**<br>(0.00111)   | 0.00180<br>(0.00114)     | 0.00176<br>(0.00115)     | 0.00119<br>(0.00111)     | 0.00106<br>(0.00116)    |
| # Students 8th $\times$ Female      |   |                         | -0.00465***<br>(0.00124) | -0.00448***<br>(0.00123) | -0.00447***<br>(0.00123) | -0.00370***<br>(0.00124) | -0.00314**<br>(0.00134) |
| Mean DV Males                       | 0.205                                   | 0.224                   | 0.224                    | 0.224                    | 0.224                    | 0.224                    | 0.224                   |
| Female as %                         | -16.271%                                | -10.251%                |                          |                          |                          |                          |                         |
| # Students 8th $\times$ Female as % |   |                         | -2.073%                  | -1.996%                  | -1.992%                  | -1.650%                  | -1.401%                 |
| R-squared                           | 0.00181                                 | 0.321                   | 0.322                    | 0.329                    | 0.330                    | 0.334                    | 0.334                   |
| # Students                          | 18123                                   | 16486                   | 16486                    | 16486                    | 16486                    | 16486                    | 16486                   |
| # Teachers                          | 316                                     | 316                     | 316                      | 316                      | 316                      | 316                      | 316                     |
| Math Teacher FE                     |   | ✓                       | ✓                        | ✓                        | ✓                        | ✓                        | ✓                       |
| Year FE                             |   | ✓                       | ✓                        | ✓                        | ✓                        | ✓                        | ✓                       |
| Std.test scores                     |   | ✓                       | ✓                        | ✓                        | ✓                        | ✓                        | ✓                       |
| Stud. Controls                      |   |                         |                          | ✓                        | ✓                        | ✓                        | ✓                       |
| Quality Classmates                  |   |                         |                          |                          | ✓                        | ✓                        | ✓                       |
| Class size                          |   |                         |                          |                          | ✓                        | ✓                        | ✓                       |
| All controls $\times$ Female        |   |                         |                          |                          |                          | ✓                        | ✓                       |
| Class size $\times$ Female          |   |                         |                          |                          |                          |                          | ✓                       |

*Notes:* This table shows coefficients  $\alpha_2$ ,  $\beta$ , and  $\gamma$  from estimation of model 7. One observation is a student assigned to a math teacher in a given year. Stud. 8th Math measures the number of other students that the math teacher has to recommend in a given year. The sample includes the students' cohorts matched with their math teachers, from [Carlana \(2019\)](#). Stud.8th Math measures the total number of 8th-grade students assigned to a teacher (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, class size, 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).

### B.3 Raw Trends

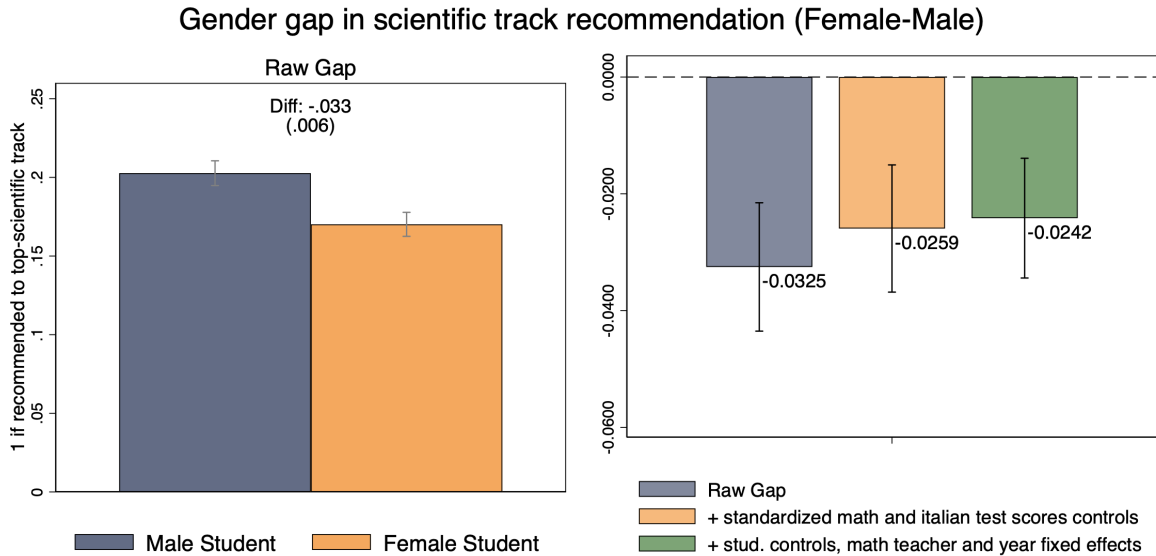
**Figure A5:** Raw trends in recommendations to the scientific track and number of students to recommend



*Notes:* This figure displays two sets of raw trends. The left graphs show the probability of being recommended to the scientific track as a function of the number of students that the math teacher has to recommend, for boys and girls. The right graphs show the raw trend of the standardized test scores in math as a function of the number of students to recommend, for boys and girls. Panel A includes all students in the sample, while Panel B only includes students in the 9th and 10th deciles of standardized math test scores.

## B.4 Gender gap in scientific track recommendations

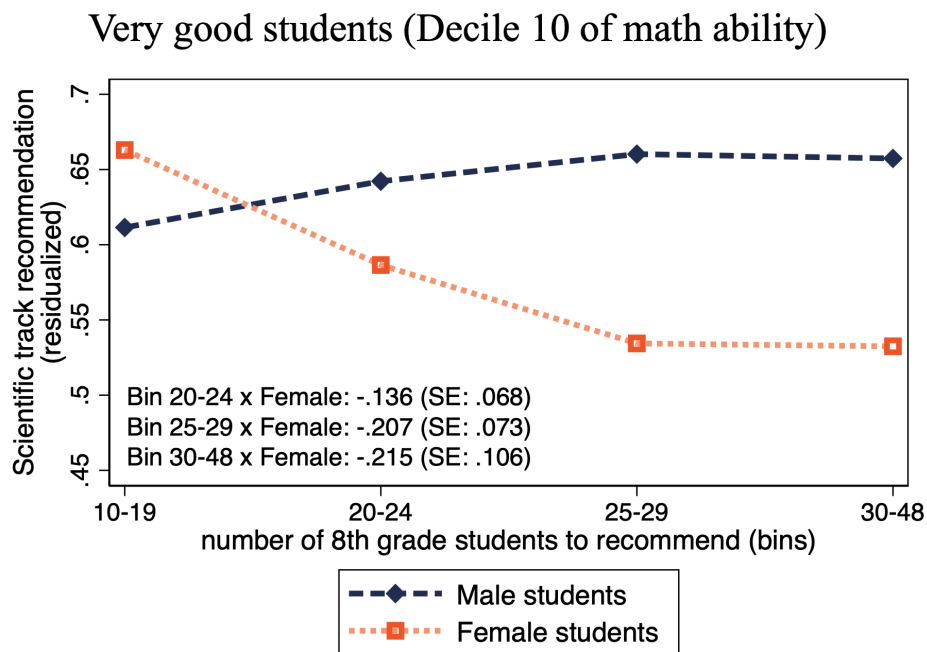
Figure A6: Top-scientific track recommendation



Notes: The left graph shows the fraction of boys and girls recommended to the scientific track, while the right graph shows the raw gender gap, the adjusted gender gap controlling for standardized test scores in math and reading, and the adjusted gender gap adding additional students' controls, math teacher and year fixed effects.

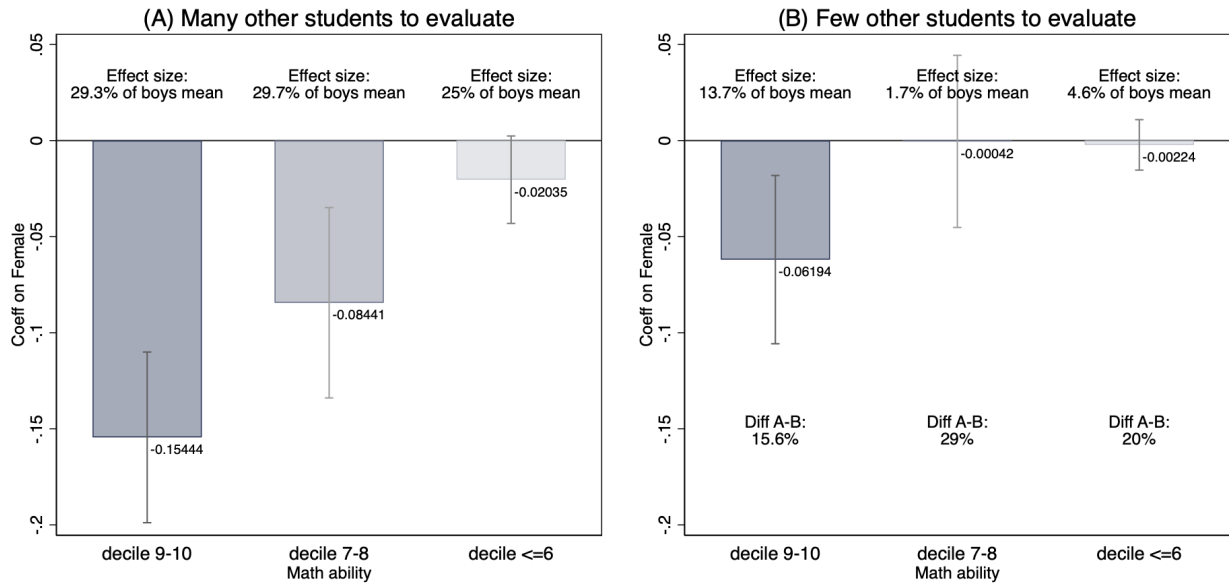
## B.5 Heterogeneity

**Figure A7:** Gender gaps for very good students in math (students in 10th decile of math ability)



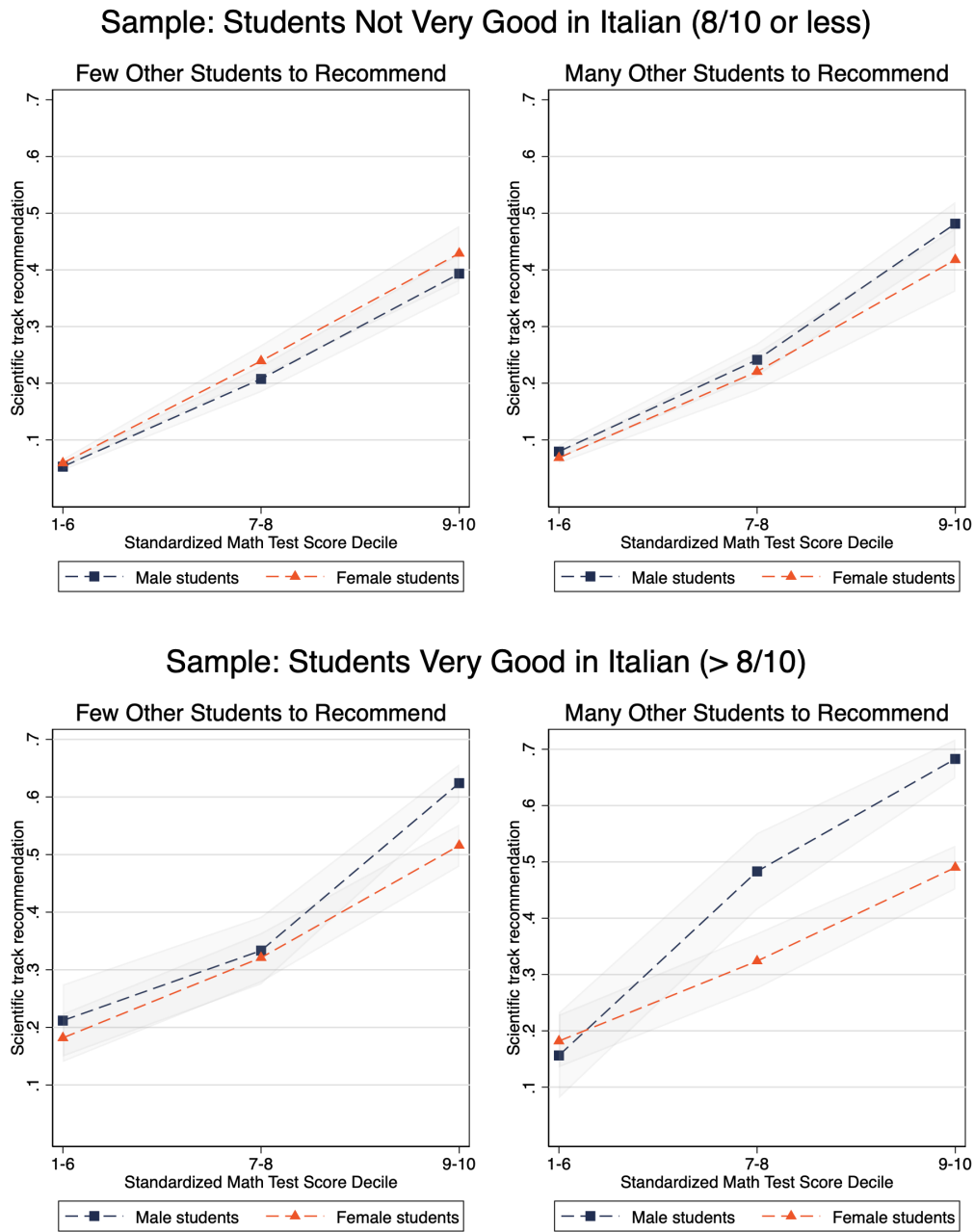
*Notes:* This figure shows the residualized trend in the probability of being recommended to the scientific track and the number of students that the math teacher needs to recommend. The figure is constructed by regressing the outcome variable on the students' controls, teacher, and year-fixed effects, and plotting the residuals by gender after adding back the mean of the dependent variable. The coefficients shown are the  $\gamma_q$  coefficients from specification 6, including the full set of students' level controls described in equation 6. The  $\gamma_q$  coefficients measure the relative gap (female-male) with respect to the gap in the first number of students bin (10-19).

**Figure A8:** Gender Gaps with High and Low Evaluation Load, by Students' Ability in Math



*Notes:* This figure shows the coefficients on the female dummy from regressing the probability of scientific track recommendation on student's gender, separately by students' math ability and by teachers' high or low evaluation load (high evaluation loads are years where the math teachers have more than 23 students to recommend, the median). As in the main specification, controls for teacher and year fixed effects, Italian math ability, students' characteristics, and quality of classmates are included.

**Figure A9:** Gender Gaps and Number of Students to Recommend, by Math and Italian Ability



*Notes:* This figure shows the fraction of students recommended to the scientific track separately by students' ability in math, students' ability in reading, and teachers' evaluation load. The two upper graphs show the fraction of boys and girls recommended to the scientific track by deciles of math ability, for students who are not very good in reading. The bottom graphs show the same statistics for students very good at reading (with reading test scores in the 9th or 10th deciles).



**Table A6:** Heterogeneity: Teachers with High and Low Implicit Stereotypes

|  | Teachers with Positive<br>Math-Boy Associations |                        |                        | Teachers with Absent or Negative<br>Math-Boy Associations |                        |                      |
|--|---|------------------------|------------------------|---|------------------------|----------------------|
|  | (1)   | (2)                    | (3)                    | (4)   | (5)                    | (6)                  |
| <b>DV: Scientific Track Recommendation</b> |   |                        |                        |   |                        |                      |
| Female                                     | 0.0306**<br>(0.0154)                            | 0.0213<br>(0.0152)     | -0.0279<br>(0.0551)    | 0.0138<br>(0.0218)  | 0.00787<br>(0.0212)    | 0.117<br>(0.0767)    |
| 1(20-24 Stud. to Recommend)                | 0.0431**<br>(0.0179)                            | 0.0352**<br>(0.0171)   | 0.0381**<br>(0.0170)   | 0.0116<br>(0.0219)  | 0.00357<br>(0.0217)    | -0.00399<br>(0.0215) |
| 1(25-29 Stud. to Recommend)                | 0.0310<br>(0.0237)                              | 0.0267<br>(0.0234)     | 0.0216<br>(0.0240)     | 0.0619*<br>(0.0332)                                       | 0.0539<br>(0.0327)     | 0.0340<br>(0.0325)   |
| 1(30-48 Stud. to Recommend)                | 0.0592*<br>(0.0350)                             | 0.0474<br>(0.0334)     | 0.0485<br>(0.0338)     | 0.0445<br>(0.0583)  | 0.0195<br>(0.0531)     | -0.00363<br>(0.0591) |
| 1(20-24 Stud. to Recommend) × Female       | -0.0484***<br>(0.0181)                          | -0.0414**<br>(0.0177)  | -0.0437**<br>(0.0176)  | -0.0161<br>(0.0262)                                       | -0.00612<br>(0.0256)   | 0.0119<br>(0.0255)   |
| 1(25-29 Stud. to Recommend) × Female       | -0.0858***<br>(0.0250)                          | -0.0836***<br>(0.0241) | -0.0778***<br>(0.0242) | -0.104***<br>(0.0328)                                     | -0.0940***<br>(0.0327) | -0.0458<br>(0.0331)  |
| 1(30-48 Stud. to Recommend) × Female       | -0.120***<br>(0.0362)                           | -0.107***<br>(0.0352)  | -0.114***<br>(0.0323)  | -0.0582<br>(0.0573)                                       | -0.0424<br>(0.0530)    | 0.0106<br>(0.0623)   |
| R-squared                                  | 0.322   | 0.353                  | 0.341                  | 0.326   | 0.359                  | 0.351                |
| N. Obs                                     | 10507   | 10507                  | 10507                  | 5979  | 5979                   | 5979                 |
| Mean DV Males                              | 0.216   | 0.216                  | 0.216                  | 0.239   | 0.239                  | 0.239                |
| Math Teacher FE                            | ✓   | ✓                      | ✓                      | ✓   | ✓                      | ✓                    |
| Year FE                                    | ✓   | ✓                      | ✓                      | ✓   | ✓                      | ✓                    |
| Std.test scores                            | ✓   | ✓                      | ✓                      | ✓   | ✓                      | ✓                    |
| Stud. Controls                             |   | ✓                      | ✓                      |   | ✓                      | ✓                    |
| Quality Classmates                         |   | ✓                      | ✓                      |   | ✓                      | ✓                    |
| Squared std. test scores                   |   | ✓                      | ✓                      |   | ✓                      | ✓                    |
| All controls × Female                      |   |                        | ✓                      |   |                        | ✓                    |

*Notes:* The table shows the  $\beta_s$  and  $\gamma_s$  coefficients from estimation of model 6. One observation is a student and the dependent variable is a dummy equal to one if the student is recommended to the top scientific high school track. Columns 1 to 3 report the results considering math teachers with positive math-boy implicit associations, while columns 4 to 6 report results for math teachers with zero or negative math-boy implicit associations. Implicit associations are measured through the Implicit Association Test (IAT) (see [Carlana \(2019\)](#), [Glover et al. \(2017\)](#)). Standard errors are clustered at the teacher level. \*p<0.1; \*\*p<0.05, \*\*\*p<0.01.

## B.6 Sensitivity and Robustness

**Table A7:** Counterfactual Decision Rules

|   | True Recommendation      | Very Good<br>in Math     | Top 30% students<br>in class | Teacher specific<br>fraction by gender |
|---|--------------------------|--------------------------|------------------------------|--|
|   | (1)                      | (2)                      | (3)                          | (4)                                    |
| # Students to Recommend                 | 0.00153<br>(0.000981)    | -0.000531<br>(0.000549)  | 0.000504<br>(0.000625)       | 0.00104<br>(0.000739)                  |
| # Students to Recommend $\times$ Female | -0.00350***<br>(0.00120) | -0.0000315<br>(0.000501) | -0.0000345<br>(0.000762)     | -0.00115<br>(0.000951)                 |
| R-squared                               | 0.357                    | 0.746                    | 0.539                        | 0.363                                  |
| N. Obs                                  | 16486                    | 16486                    | 16486                        | 16486                                  |
| Teacher FE                              | ✓                        | ✓                        | ✓                            | ✓                                      |
| Year FE                                 | ✓                        | ✓                        | ✓                            | ✓                                      |
| Std. test scores                        | ✓                        | ✓                        | ✓                            | ✓                                      |
| Controls                                | ✓                        | ✓                        | ✓                            | ✓                                      |
| Quality classmates                      | ✓                        | ✓                        | ✓                            | ✓                                      |
| Squared std. test scores                | ✓                        | ✓                        | ✓                            | ✓                                      |
| Controls $\times$ Female                | ✓                        | ✓                        | ✓                            | ✓                                      |
| Mean DV                                 | 0.204                    | 0.214                    | 0.163                        | 0.105                                  |
| Sd Dependent Variable                   | 0.403                    | 0.410                    | 0.369                        | 0.307                                  |

*Notes:* This table shows coefficient  $\gamma$  from estimation of model 7. One observation is a student assigned to a math teacher in a given year. The dependent variable of column (1) is the actual scientific recommendation, while the dependent variables in columns (2) to (4) are counterfactual recommendations if teachers followed a set of pre-determined decision rules. The decision rule in column (2) is: the student is recommended to the scientific track if she is very good at math (is in decile of math ability 9-10). The decision rule in column (3) is: the student is recommended to the scientific track if she is one of the 30% top students in math of her class. The decision rule in column (4) is: each teacher has a gender-specific fraction of students that she recommends to the scientific track (derived from the data). A student is recommended to scientific if she is one of the best X% of students according to her teacher-specific threshold. Students' controls include students' standardized math and Italian test scores, students' mother education dummies, students' father occupation dummies, immigrant status, class size, 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), and their interaction with the total number of 8th-grade students. Standard errors are clustered at the teacher level. \* $p < 0.1$ ; \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

## B.7 Alternative Explanations

**Table A8:** Only Using Variation from Number of Students (to recommend) in Other 8th Grade Classes

|   | <b>DV: top-scientific track recommendation</b> |            |            |           |
|---|--|------------|------------|-----------|
|   | (1)  | (2)        | (3)        | (4)       |
| Female                                      | -0.0268***                                     | -0.0198*** | -0.0219*** | 0.224***  |
|   | (0.00628)                                      | (0.00636)  | (0.00660)  | (0.0666)  |
| 1(14-19 Students in Other Classes)          | -0.0163  | -0.00338   | -0.0123    | 0.000460  |
|   | (0.0537)                                       | (0.0602)   | (0.0535)   | (0.0553)  |
| 1(20-26 Students in Other Classes)          | 0.00390  | 0.0382     | 0.0258     | 0.0304    |
|   | (0.0287)                                       | (0.0265)   | (0.0253)   | (0.0250)  |
| 1(14-19 Students in Other Classes) × Female | -0.0203  | 0.0363     | 0.0272     | -0.00327  |
|   | (0.0338)                                       | (0.0444)   | (0.0464)   | (0.0534)  |
| 1(20-26 Students in Other Classes) × Female | -0.0782**                                      | -0.0803**  | -0.0692**  | -0.0761** |
|   | (0.0307)                                       | (0.0311)   | (0.0309)   | (0.0315)  |
| R-squared                                   | 0.162  | 0.322      | 0.352      | 0.357     |
| N. Obs                                      | 18123  | 16486      | 16486      | 16486     |
| Mean DV Males                               | 0.205  | 0.224      | 0.224      | 0.224     |
| Math Teacher FE                             | ✓  | ✓          | ✓          | ✓         |
| Year FE                                     | ✓  | ✓          | ✓          | ✓         |
| Std.test scores                             |  | ✓          | ✓          | ✓         |
| Stud. Controls                              |  |            | ✓          | ✓         |
| Quality Classmates                          |  |            | ✓          | ✓         |
| Class size                                  |  |            | ✓          | ✓         |
| Squared std. test scores                    |  |            | ✓          | ✓         |
| Std.test scores × Female                    |  |            |            | ✓         |
| Stud. Controls × Female                     |  |            |            | ✓         |
| Quality Classmates × Female                 |  |            |            | ✓         |
| Class size × Female                         |  |            |            | ✓         |
| Squared std. test scores × Female           |  |            |            | ✓         |

*Notes:* This table shows coefficients  $\gamma_s$  and  $\beta_s$  from estimation of model 6. One observation is a student assigned to a math teacher in a given year. The dependent variable is a dummy equal to 1 if the student is recommended for the scientific track. The indicator variables measure bins of the total number of students that the math teacher needs to recommend in a given year and are in 8th-grade classes different than the student's own class. The reference groups are teachers in years in which they need to recommend no students from other 8th-grade classes (because they are only assigned to one 8th-grade class). The sample includes the 8th-grade students matched with their math teachers from the main observational sample. Standard errors are clustered at the teacher level. \*p<0.1; \*\*p<0.05, \*\*\*p<0.01.

**Table A9:** Teachers' perception of limited spots in scientific track and girls better in Italian

|  | DV: top-scientific track recommendation |                      |                       |                       |
|--|---|----------------------|-----------------------|-----------------------|
|  | (1)                                     | (2)                  | (3)                   | (4)                   |
| $\mathbb{1}(20-24 \text{ Stud. to Recommend})$                                     | 0.0124<br>(0.0131)                      | 0.0126<br>(0.0133)   | 0.00746<br>(0.0122)   | 0.00975<br>(0.0123)   |
| $\mathbb{1}(25-29 \text{ Stud. to Recommend})$                                     | -0.00480<br>(0.0171)                    | -0.00502<br>(0.0169) | -0.00949<br>(0.0157)  | -0.00999<br>(0.0160)  |
| $\mathbb{1}(30-48 \text{ Stud. to Recommend})$                                     | 0.00816<br>(0.0252)                     | 0.00858<br>(0.0252)  | -0.00468<br>(0.0229)  | -0.00560<br>(0.0244)  |
| $\mathbb{1}(20-24 \text{ Students to Recommend}) \times \text{Italian Test Score}$ |   | 0.00304<br>(0.00875) | -0.00484<br>(0.00828) | -0.00280<br>(0.00845) |
| $\mathbb{1}(25-29 \text{ Students to Recommend}) \times \text{Italian Test Score}$ |   | 0.00507<br>(0.0122)  | -0.00955<br>(0.0114)  | -0.00144<br>(0.0117)  |
| $\mathbb{1}(30-48 \text{ Students to Recommend}) \times \text{Italian Test Score}$ |   | 0.00277<br>(0.0163)  | -0.00662<br>(0.0148)  | 0.0000537<br>(0.0159) |
| R-squared  | 0.322                                   | 0.322                | 0.352                 | 0.342                 |
| N. Obs   | 16486                                   | 16486                | 16486                 | 16486                 |
| Mean DV Males  | 0.204                                   | 0.204                | 0.204                 | 0.204                 |
| Math Teacher FE  | ✓                                       | ✓                    | ✓                     | ✓                     |
| Year FE  | ✓                                       | ✓                    | ✓                     | ✓                     |
| Std.test scores  | ✓                                       | ✓                    | ✓                     | ✓                     |
| Stud. Controls   |   |                      | ✓                     | ✓                     |
| Quality classmates   |   |                      | ✓                     | ✓                     |
| Squared std. test scores   |   |                      | ✓                     | ✓                     |
| All controls $\times$ Female   |   |                      |                       | ✓                     |

*Notes:* This table shows coefficients  $\gamma_s$  and  $\beta_s$  from the estimation of a similar model as 6, but with the interaction of number of students to recommend and reading test score. One observation is a student assigned to a math teacher in a given year. The dependent variable is a dummy equal to 1 if the student is recommended for the scientific track. The indicator variables measure bins of the total number of students that the math teachers need to recommend in a given year, alone and interacted with the student's standardized test scores in reading. The reference groups are teachers in years when they need to recommend 10-19 students (the first number of students bin). The sample includes the 8th-grade students matched with their math teachers from the main observational sample. Standard errors are clustered at the teacher level. \*p<0.1; \*\*p<0.05, \*\*\*p<0.01.

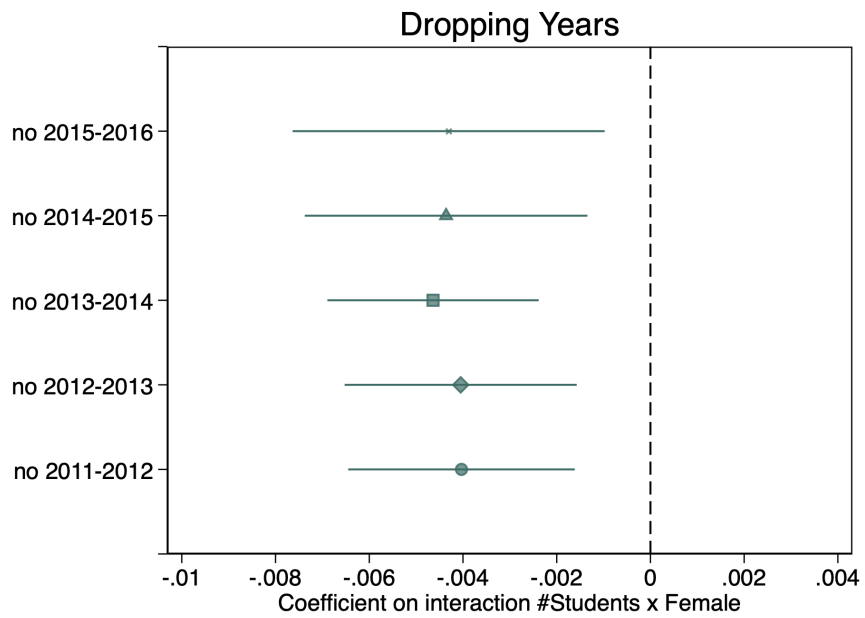
**Table A10:** Total number of students assigned versus Number Students in 8th Grade (who need to receive a recommendation)

|   | DV: <b>top-scientific track recommendation</b> |                          |                          |                          |
|---|--|--------------------------|--------------------------|--------------------------|
|   | (1)  | (2)                      | (3)                      | (4)                      |
| # Students assigned to Teacher          | 0.000672<br>(0.000576)                         | 0.000447<br>(0.000626)   | 0.000627<br>(0.000615)   | 0.000506<br>(0.000615)   |
| # Students assigned to Teacher × Female | -0.000657<br>(0.000452)                        | -0.000145<br>(0.000451)  | -0.000140<br>(0.000440)  | 0.0000524<br>(0.000445)  |
| # Students to Recommend                 |  | 0.00193<br>(0.00120)     | 0.00154<br>(0.00109)     | 0.00126<br>(0.00106)     |
| # Students to Recommend × Female        |  | -0.00452***<br>(0.00128) | -0.00410***<br>(0.00128) | -0.00353***<br>(0.00125) |
| R-squared                               | 0.322  | 0.322                    | 0.353                    | 0.357                    |
| N. Obs                                  | 16486  | 16486                    | 16486                    | 16486                    |
| Mean DV Males                           | 0.204  | 0.204                    | 0.204                    | 0.204                    |
| Math Teacher FE                         | ✓  | ✓                        | ✓                        | ✓                        |
| Year FE                                 | ✓  | ✓                        | ✓                        | ✓                        |
| Std.test scores                         | ✓  | ✓                        | ✓                        | ✓                        |
| Stud. controls                          |  |                          | ✓                        | ✓                        |
| Quality classmates                      |  |                          | ✓                        | ✓                        |
| Squared std. test scores                |  |                          | ✓                        | ✓                        |
| All controls × Female                   |  |                          |                          | ✓                        |

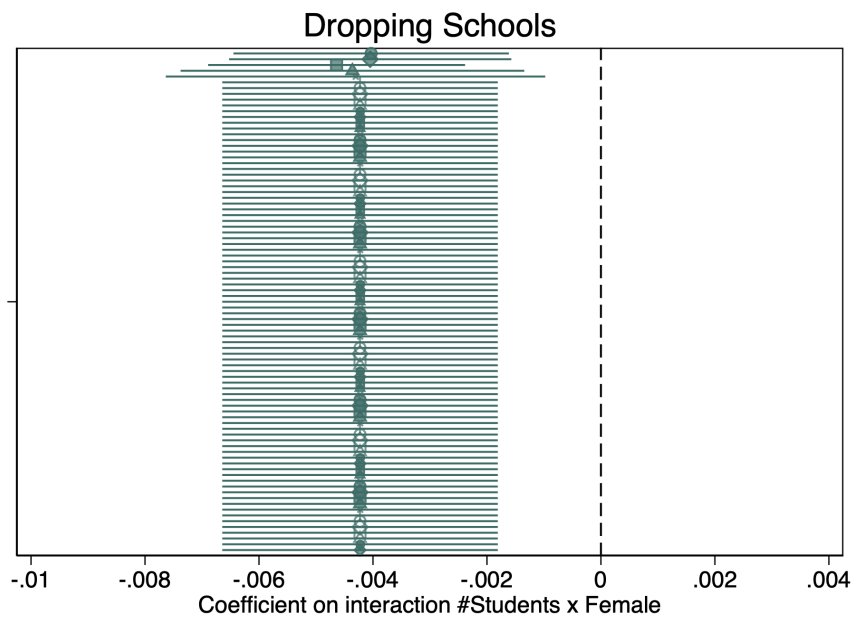
*Notes:* This table shows results from the estimation of model 7, showing both the coefficients on the number of students to recommend and on the total number of students assigned to the teacher (in 6th and 7th grade as well). One observation is a student assigned to a math teacher in a given year. The dependent variable is a dummy equal to 1 if the student is recommended to the scientific track. The sample includes the 8th-grade students matched with their math teachers from the main observational sample. Standard errors are clustered at the teacher level. \*p<0.1; \*\*p<0.05, \*\*\*p<0.01.

Figure A10: Dropping years and schools

Panel (a): Dropping years



Panel (b): Dropping schools



Notes:

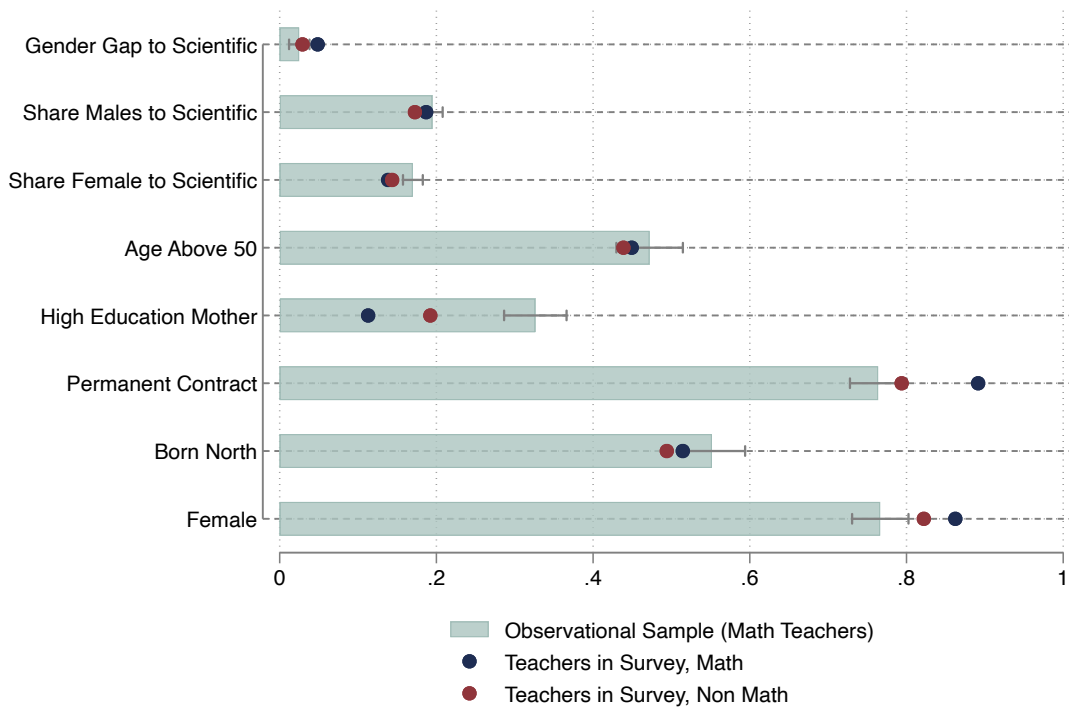
**Table A11:** Restricting range of students to recommend

|   | DV: top-scientific track recommendation |                      |                                |
|---|---|----------------------|--------------------------------|
|   | (1)                                     | (2)                  | (3)                            |
| Female                                  | 0.0585*                                 | 0.213***             | 0.0923**                       |
|   | (0.0335)                                | (0.0512)             | (0.0420)                       |
| # Students to Recommend                 | 0.00177                                 | 0.00382*             | 0.00174                        |
|   | (0.00113)                               | (0.00224)            | (0.00140)                      |
| # Students to Recommend $\times$ Female | -0.00365***                             | -0.0105***           | -0.00493***                    |
|   | (0.00137)                               | (0.00234)            | (0.00185)                      |
| Mean DV Males                           | 0.212                                   | 0.204                | 0.210                          |
| # Students 8th $\times$ Female as %     | -1.545%                                 | -4.740%              | -2.165%                        |
| R-squared                               | 0.346                                   | 0.352                | 0.361                          |
| # Students                              | 14148                                   | 15488                | 13355                          |
| # Teachers                              | 306                                     | 314                  | 240                            |
| Math Teacher FE                         | ✓                                       | ✓                    | ✓                              |
| Year FE                                 | ✓                                       | ✓                    | ✓                              |
| Std.test scores                         | ✓                                       | ✓                    | ✓                              |
| Stud. Controls                          | ✓                                       | ✓                    | ✓                              |
| Quality Classmates                      | ✓                                       | ✓                    | ✓                              |
| Squared std. test scores                | ✓                                       | ✓                    | ✓                              |
| Sample                                  | # students $\geq$ 20                    | # students $\leq$ 30 | teacher observed for 4/5 years |

# C Additional Figures and Tables for Second Result on Past Track Recommendations

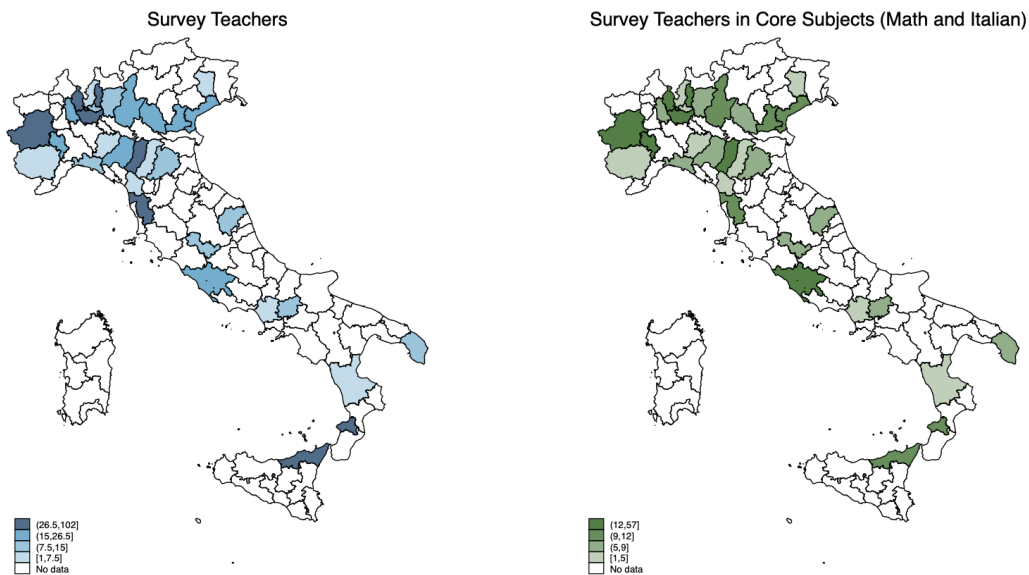
## C.1 Summary Statistics

Figure A11: Teachers in Survey and Observational Sample





**Figure A12:** Survey Teachers (all teachers and teachers in math and Italian)

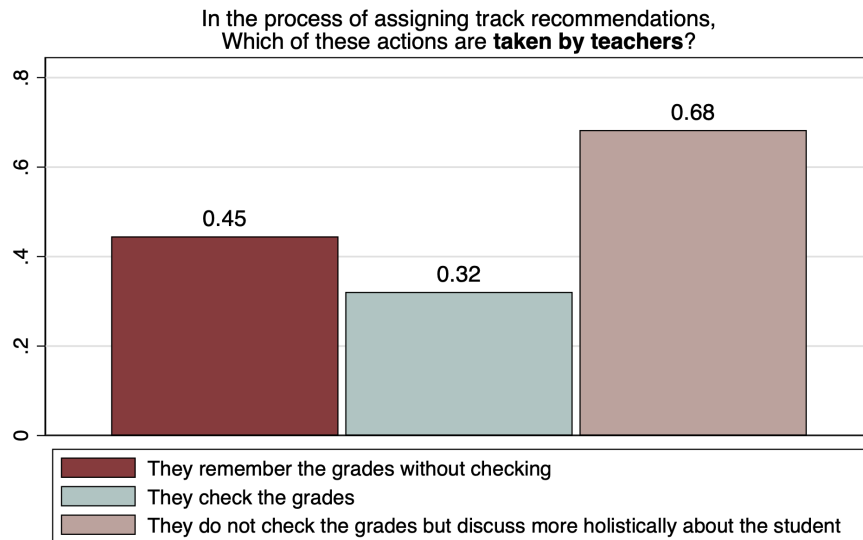


**Table A12:** Characteristics of teachers in the survey and in administrative sample

| Variable                                       | (1)<br>Math Teachers in Admin. Sample | (2)<br>SD | (3)<br>Teachers in Survey | (4)<br>SD | (5)<br>Diff. | (6)<br>P-val. |
|--|---------------------------------------|-----------|---------------------------|-----------|--------------|---------------|
| Female   | 0.767                                 | (0.424)   | 0.828                     | (0.378)   | 0.061        | (0.017)**     |
| Born in Northern Italy                         | 0.552                                 | (0.498)   | 0.493                     | (0.500)   | -0.059       | (0.070)*      |
| Permanent contract                             | 0.764                                 | (0.425)   | 0.814                     | (0.389)   | 0.050        | (0.056)*      |
| Mother is HE                                   | 0.326                                 | (0.469)   | 0.175                     | (0.381)   | -0.151       | (0.000)***    |
| Age  | 45.016                                | (18.116)  | 47.273                    | (9.740)   | 2.257        | (0.010)***    |
| IAT  | -0.401                                | (0.892)   | 0.452                     | (0.740)   | 0.853        | (0.000)***    |
| Gender gap in past scientific track recomm.    | 2.495                                 | (15.370)  | 3.344                     | (12.738)  | 0.848        | (0.392)       |
| Fraction of girls recommended to scient. track | 0.170                                 | (0.149)   | 0.142                     | (0.098)   | -0.028       | (0.002)***    |
| Fraction of boys recommended to scient. track  | 0.195                                 | (0.153)   | 0.176                     | (0.111)   | -0.019       | (0.041)**     |
| Observations                                   | 377                                   |           | 651                       |           | 1,028        |               |

## C.2 Recommendation Process

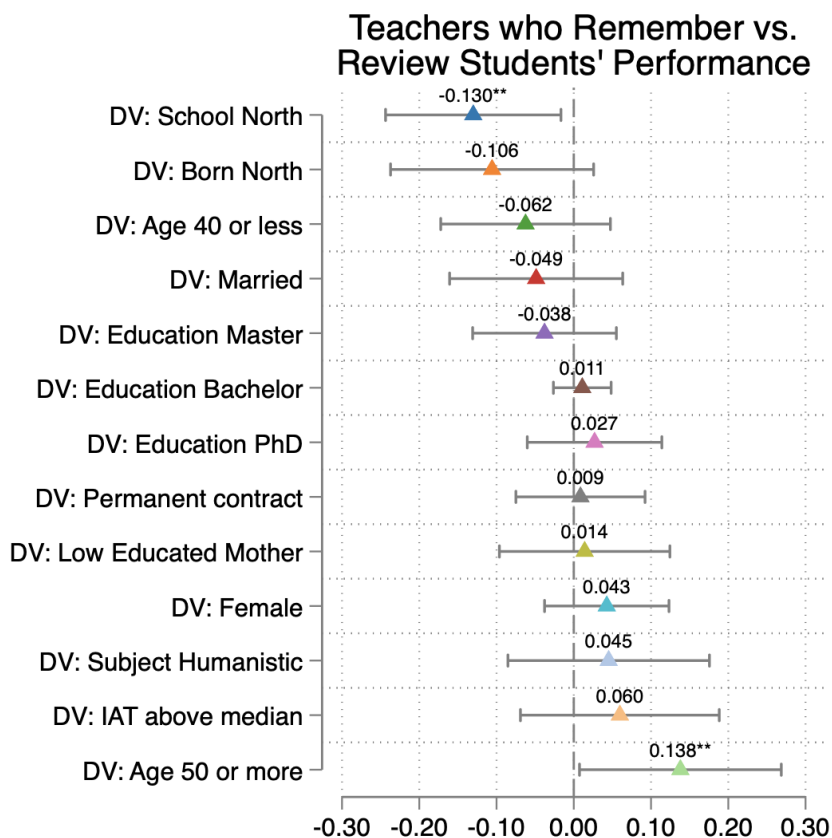
**Figure A13:** Survey to teachers: actions taken when assigning track recommendation



*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 around 500 teachers from a survey run in February 2023 in 70 schools.

### C.3 Characteristics of Teachers who Check vs. Rely on Memory

**Figure A14:** Characteristics of teachers who rely on memory vs. review student performance



*Notes:* This figure shows coefficients from regressions where the dependent variables are the teacher's characteristics and the independent variable is a dummy equal to one if the teacher reported that when assigning track recommendations she does not check but remembers student performance. The sample includes teachers in core subjects matched with their past students.

## C.4 Alternative Comparison Groups

**Table A13:** Alternative comparison group: teachers who report they rely on memory vs. teachers who report that either check, someone else checks, or other

|                          | DV: Scientific Track Recommendation |                           |                         |                        |
|--------------------------|-------------------------------------|---------------------------|-------------------------|------------------------|
|                          | (1)                                 | (2)                       | (3)                     | (4)                    |
| Female                   | -0.0275**<br>(0.0128)               | -0.0361<br>(0.0578)       | -0.0283***<br>(0.00899) | -0.0167<br>(0.0937)    |
| Memory                   | 0.0199**<br>(0.00992)               | 0.0205**<br>(0.00901)     | 0.0130*<br>(0.00693)    | 0.0156**<br>(0.00646)  |
| Memory × Female          | -0.0387**<br>(0.0193)               | -0.0415**<br>(0.0180)     | -0.0265*<br>(0.0141)    | -0.0319**<br>(0.0132)  |
| Mean control             | 0.203                               | 0.203                     | 0.194                   | 0.194                  |
| Memory × Female as a %   | -19.065%                            | -20.456%                  | -13.653%                | -16.432%               |
| Observations             | 9995                                | 9995                      | 22551                   | 22551                  |
| # teachers               | 200                                 | 200                       | 347                     | 347                    |
| R <sup>2</sup>           | 0.263                               | 0.271                     | 0.261                   | 0.280                  |
| Year FE                  | ✓                                   | ✓                         | ✓                       | ✓                      |
| std. Test Scores         | ✓                                   | ✓                         | ✓                       | ✓                      |
| Stud. Controls           | ✓                                   | ✓                         | ✓                       | ✓                      |
| Teacher Controls         | ✓                                   | ✓                         | ✓                       | ✓                      |
| Class FE                 | ✓                                   | ✓                         | ✓                       | ✓                      |
| IAT                      | ✓                                   | ✓                         | ✓                       | ✓                      |
| Squared Std. Test Scores |                                     | ✓                         |                         | ✓                      |
| All Controls × Female    |                                     | ✓                         |                         | ✓                      |
| Sample                   | Teachers<br>core subjects           | Teachers<br>core subjects | All survey<br>teachers  | All survey<br>teachers |

## C.5 Sensitivity and Robustness

## C.6 Where are the missing "scientific" girls sent?

**Table A14:** Number of Students to Recommend and Gender Gaps, Supplemental Sample of Survey Teachers

|                          | Scientific                | Classical                | Medium Humanities       | Med. Tech             | Vocational             |
|--------------------------|---------------------------|--------------------------|-------------------------|-----------------------|------------------------|
|                          | (1)                       | (2)                      | (3)                     | (4)                   | (5)                    |
| Stud.8th $\times$ Female | -0.00279***<br>(0.000854) | 0.00125***<br>(0.000427) | 0.0000506<br>(0.000774) | 0.00113<br>(0.000863) | 0.000697<br>(0.000832) |
| Mean control             | 0.176                     | 0.046                    | 0.204                   | 0.266                 | 0.299                  |
| Observations             | 12460                     | 12460                    | 12619                   | 12619                 | 12619                  |
| N. teachers              | 248                       | 248                      | 248                     | 248                   | 248                    |
| R <sup>2</sup>           | 0.213                     | 0.109                    | 0.152                   | 0.077                 | 0.331                  |
| Std. Test Scores         | ✓                         | ✓                        | ✓                       | ✓                     | ✓                      |
| Stud. Controls           | ✓                         | ✓                        | ✓                       | ✓                     | ✓                      |
| Year FE                  | ✓                         | ✓                        | ✓                       | ✓                     | ✓                      |
| Teacher FE               | ✓                         | ✓                        | ✓                       | ✓                     | ✓                      |

# D Additional Figures and Tables for Teachers Experiment

## D.1 Experiment Design

Figure A15: Profiles' Characteristics

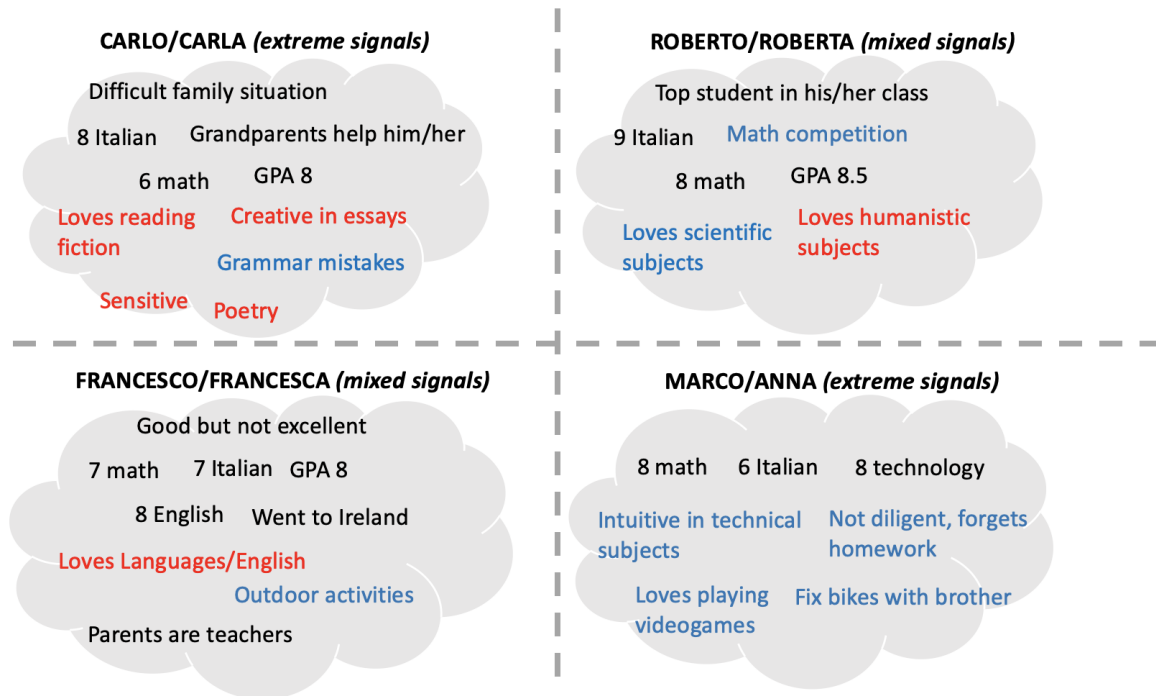


Figure A16: Baseline Treatment

Student: [Roberto/Roberta]: [...]  
Student: [Carlo/Carla] [...]  
Student: [Francesco/ Francesca] [...]  
Student: [Marco/ Anna] [...]

---

*In this section, we ask you to report the characteristics that you have read previously about the 4 students, and to provide track recommendations.*

**Student: Roberto/a**

|  |   |                                    |                                  |
|--|---|------------------------------------|----------------------------------|
| <b>Academic ability:</b>   | Grade Math:<br>(drop-down list)               | Grade Italian:<br>(drop-down list) | Grade Other:<br>(drop-down list) |
| <b>Interests:</b>  | <input type="text"/>                          |                                    |                                  |
| <b>Other Characteristics:</b>                                    | <input type="text"/>                          |                                    |                                  |
| <b>Which high school track would you recommend to Roberto/a?</b> | <input type="text" value="(drop-down list)"/> |                                    |                                  |

Figure A17: Memory Treatment

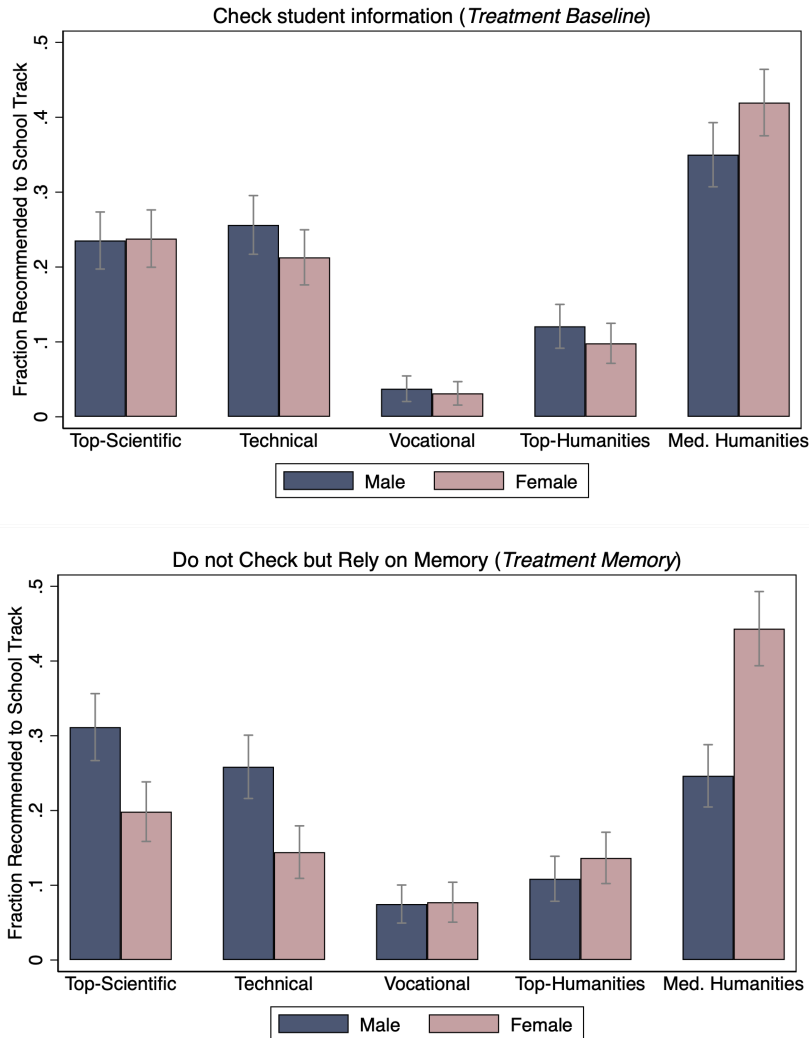
*In this section, we ask you to report the characteristics that you have read previously about the 4 students, and to provide track recommendations. Please write down the characteristics that you remember.*

**Student: Roberto/a**

|  |   |                                    |                                  |
|--|---|------------------------------------|----------------------------------|
| <b>Academic ability:</b>   | Grade Math:<br>(drop-down list)               | Grade Italian:<br>(drop-down list) | Grade Other:<br>(drop-down list) |
| <b>Interests:</b>  | <input type="text"/>                          |                                    |                                  |
| <b>Other Characteristics:</b>                                    | <input type="text"/>                          |                                    |                                  |
| <b>Which high school track would you recommend to Roberto/a?</b> | <input type="text" value="(drop-down list)"/> |                                    |                                  |

## D.2 Recommendations to all tracks

**Figure A18:** Experiment with Teachers: Limited Memory and Teachers' Track recommendations



*Notes:* This figure shows the fraction of students assigned to humanistic and scientific tracks, for female and male students and for teachers in the memory and in the baseline condition. Both teachers in the baseline and memory conditions observed the same students' profiles. Teachers in the memory condition need to retrieve students' characteristics from their memory (they do not have the profiles in front of them when they make evaluations), while teachers in the baseline condition can review students' characteristics before providing recommendations. The baseline sample included 443 teachers from 68 middle schools who completed the teachers' experiment.



**Table A15:** Teachers limited memory and recommendation gaps

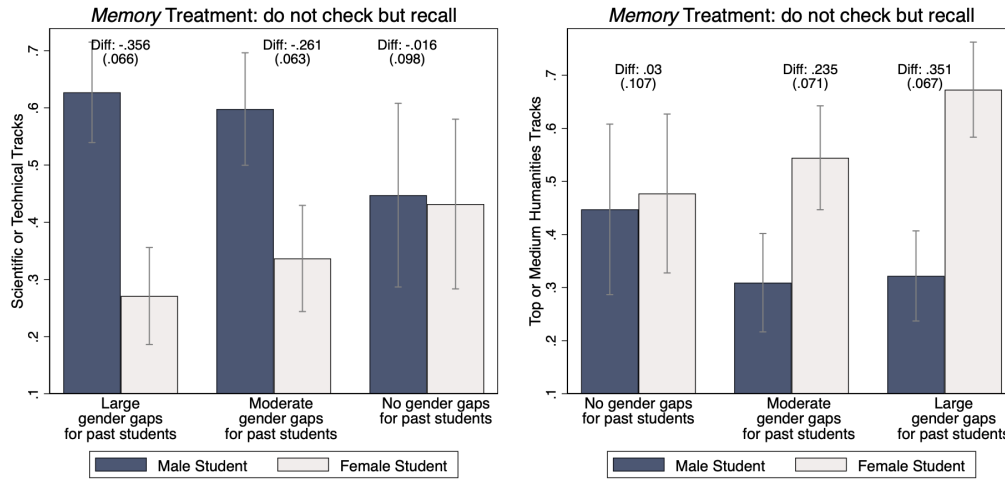
|  | (1)                   | (2)                   | (3)                 | (4)                   | (5)                 |
|--|-----------------------|-----------------------|---------------------|-----------------------|---------------------|
|  | Top-Scientific        | Med-Tech              | Top-Hum             | Med-Hum               | Low-Vocational      |
| <b>Panel A: baseline sample, no teacher FE</b> |                       |                       |                     |                       |                     |
| Female   | 0.001<br>(0.016)      | -0.043**<br>(0.017)   | -0.020<br>(0.018)   | 0.070***<br>(0.023)   | -0.008<br>(0.013)   |
| Female $\times$ Memory                         | -0.122***<br>(0.033)  | -0.063**<br>(0.032)   | 0.044<br>(0.028)    | 0.125***<br>(0.040)   | 0.016<br>(0.023)    |
| Observations                                   | 1761                  | 1761                  | 1761                | 1761                  | 1761                |
| N. teachers                                    | 448                   | 448                   | 448                 | 448                   | 448                 |
| R <sup>2</sup>                                 | 0.384                 | 0.344                 | 0.099               | 0.305                 | 0.058               |
| Teacher FE                                     | No                    | No                    | No                  | No                    | No                  |
| <b>Panel B: baseline sample, teacher FE</b>    |                       |                       |                     |                       |                     |
| Female   | -0.0124<br>(0.0197)   | -0.0388**<br>(0.0191) | -0.0171<br>(0.0236) | 0.0836***<br>(0.0275) | -0.0153<br>(0.0140) |
| Female $\times$ Memory                         | -0.130***<br>(0.0385) | -0.0897**<br>(0.0348) | 0.0369<br>(0.0353)  | 0.157***<br>(0.0481)  | 0.0259<br>(0.0250)  |
| Observations                                   | 1757                  | 1757                  | 1757                | 1757                  | 1757                |
| N. teachers                                    | 444                   | 444                   | 444                 | 444                   | 444                 |
| R <sup>2</sup>                                 | 0.535                 | 0.480                 | 0.275               | 0.435                 | 0.326               |
| Teacher FE                                     | Yes                   | Yes                   | Yes                 | Yes                   | Yes                 |
| Control mean                                   | 0.235                 | 0.256                 | 0.121               | 0.350                 | 0.037               |
| Student FE                                     | Yes                   | Yes                   | Yes                 | Yes                   | Yes                 |
| Controls                                       | Yes                   | Yes                   | Yes                 | Yes                   | Yes                 |

*Notes:* This table shows coefficients  $\beta_2$  and  $\beta_3$  from estimation of equations 9 and 10 where the depend variables are each track recommendation. In Panel A, we do not include teacher fixed effects, while in Panel B teacher fixed effects are included. Teachers in the baseline sample are included. 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. Columns (3) and (6) include teacher fixed effects.

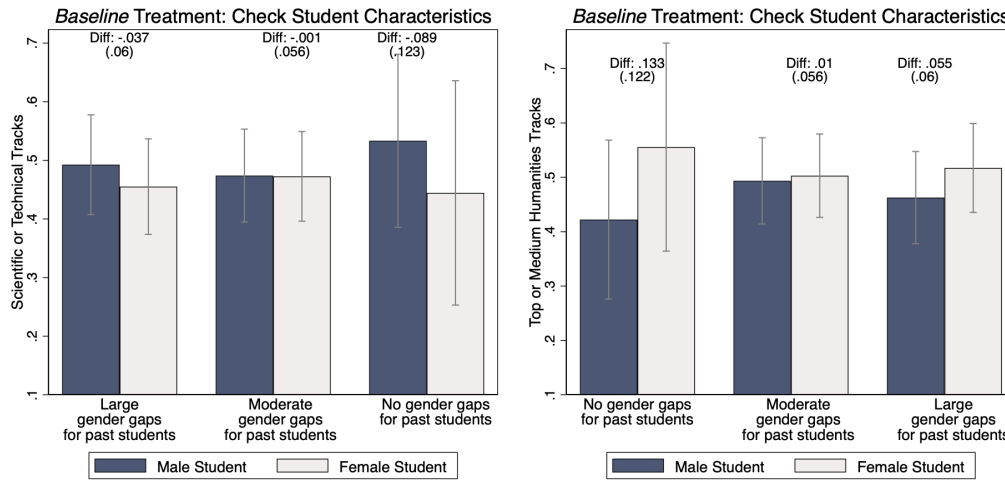
### D.3 Heterogeneity

**Figure A19:** Gaps for hypothetical students and gaps for past students

**Panel (a):** Gaps for hypothetical students and past gaps for teachers in *Memory* group



**Panel (b):** Gaps for hypothetical students and past gaps for teachers in *Baseline* group



*Notes:* This figure shows gender gaps in recommendations for the scientific and technical tracks (Panel (a)) and the top and medium humanities tracks (Panel (b)) for teachers in the baseline and memory conditions as a function of the past gaps in track recommendations in scientific or technical tracks provided to their past students.

Figure A20: Heterogeneity of treatment effect by IAT

Male-typed track recommendations (top-scientific and medium-technical)  
by teachers implicit associations (IAT)

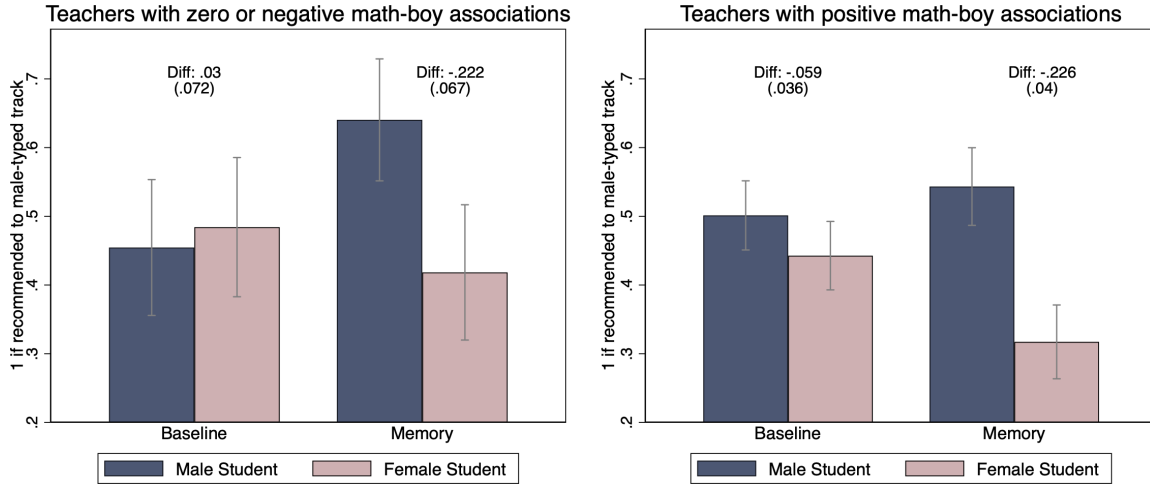
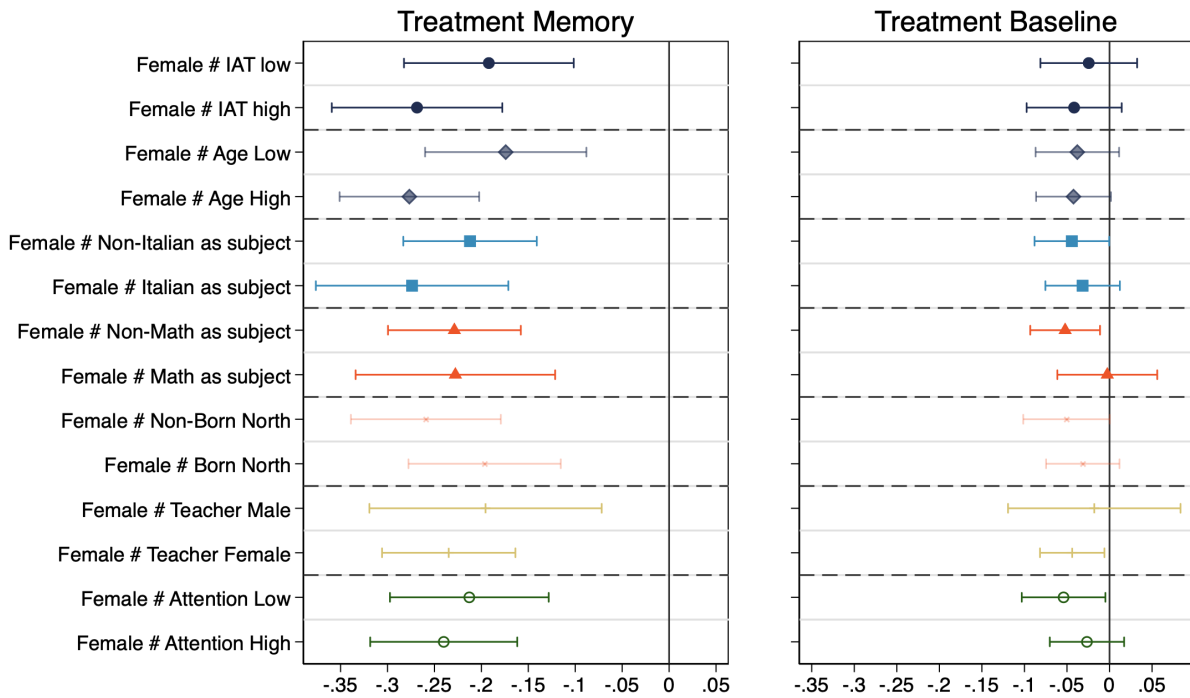
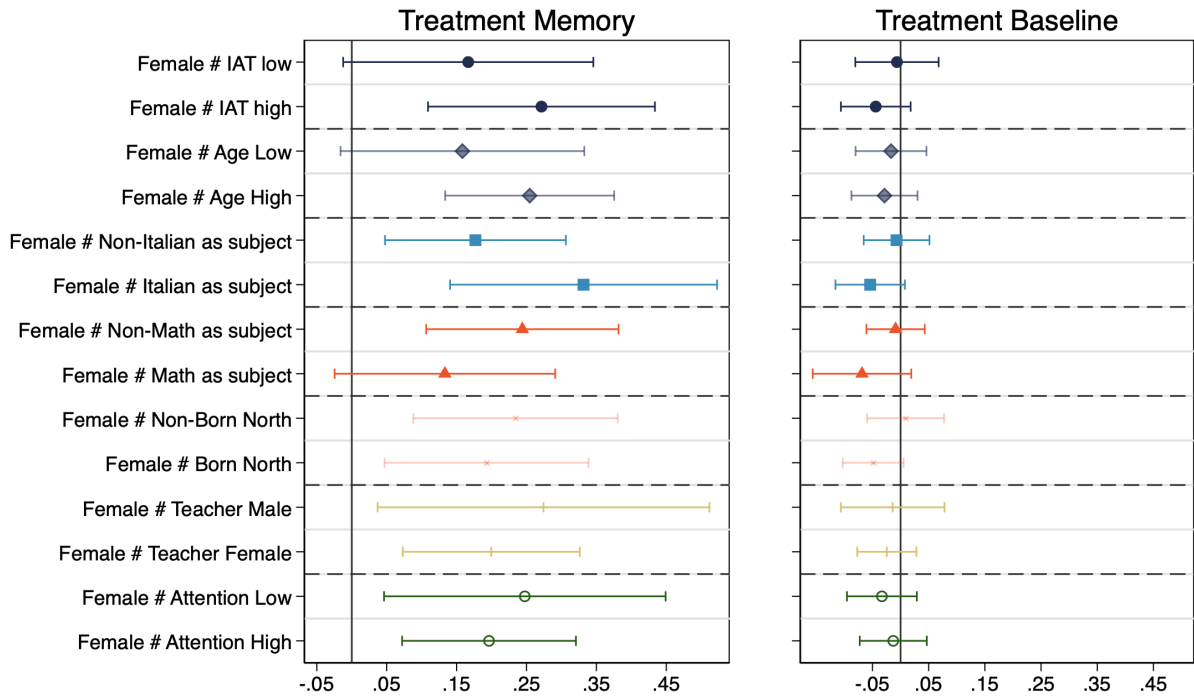


Figure A21: Treatment Effect Heterogeneity. The dependent variable is 1 if the student is recommended to scientific or technical tracks

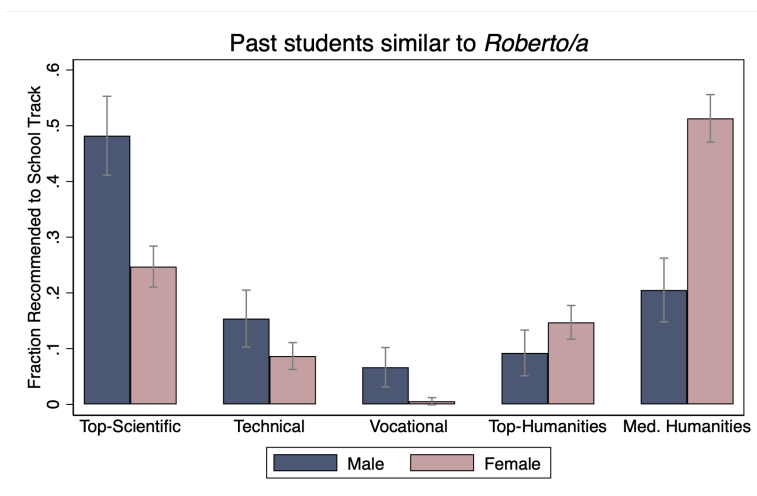


**Figure A22:** Treatment Effect Heterogeneity. The dependent variable is the recalled share of female-typed minus male-typed signals

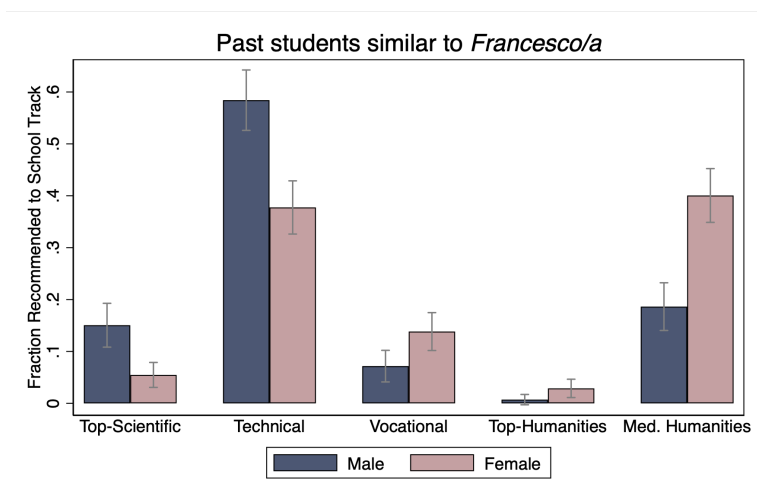


## D.4 Recommendations for past students with same grades as student profiles

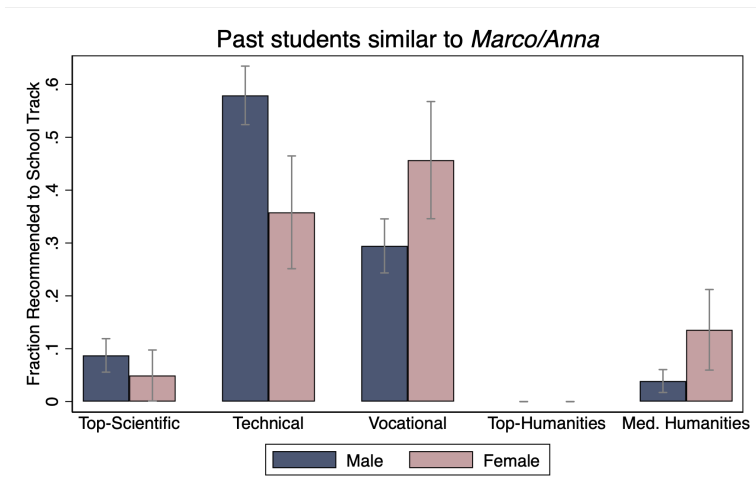
**Figure A23:** Past Students with 8/10 in math and 9/10 in Italian, GPA around 8.5/10 (same grades as *Roberto/Roberta*). Additional characteristics of student profile with respect to past students with the same grades: participated in a math competition.



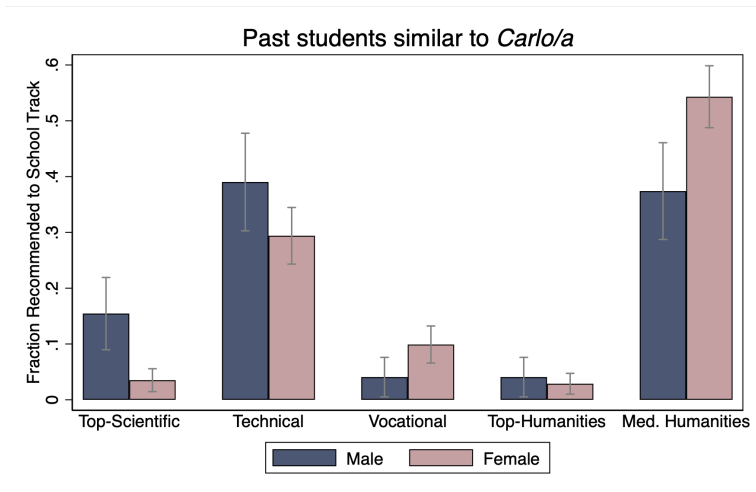
**Figure A24:** Past Students with 7/10 in math and Italian, GPA around 8, English grade 8 (same grades as *Francesco/Francesca*). Additional characteristics of student profile with respect to past students with the same grades: very passionate about languages and spent summer in Ireland to learn English.



**Figure A25:** 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.



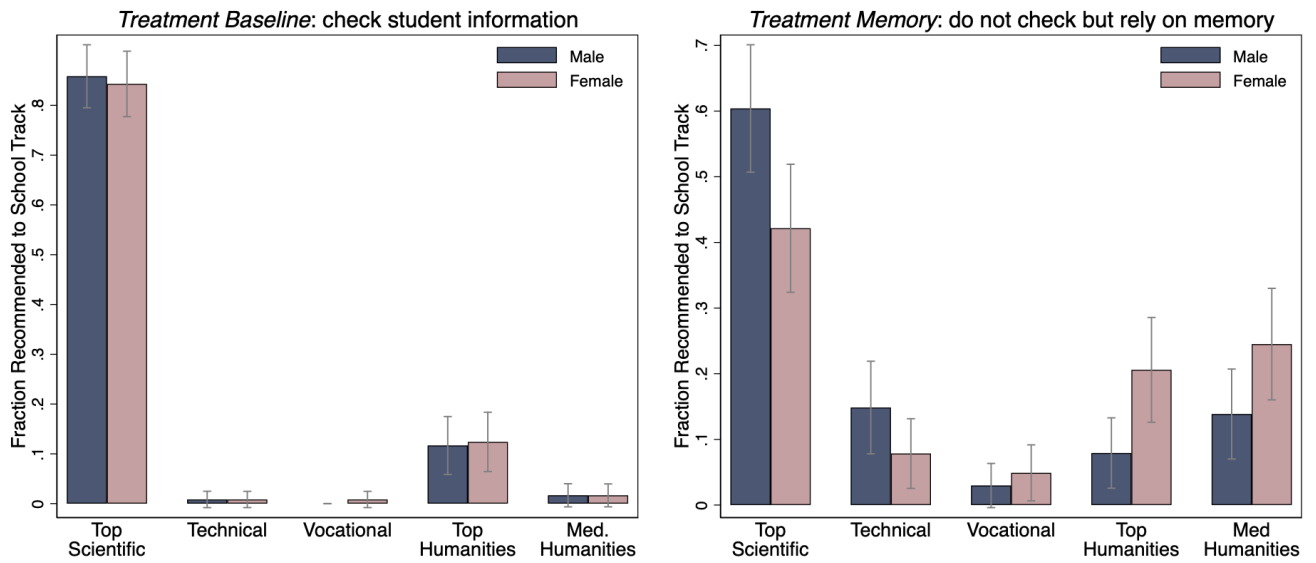
**Figure A26:** Past Students with 6/10 in math and 8/10 Italian, GPA around 8 (same grades as *Carlo/Carla*). Additional characteristics of student profile with respect to past students with the same grades: loves reading fiction and poetry, creative in essays.



## D.5 Results by Student Profile

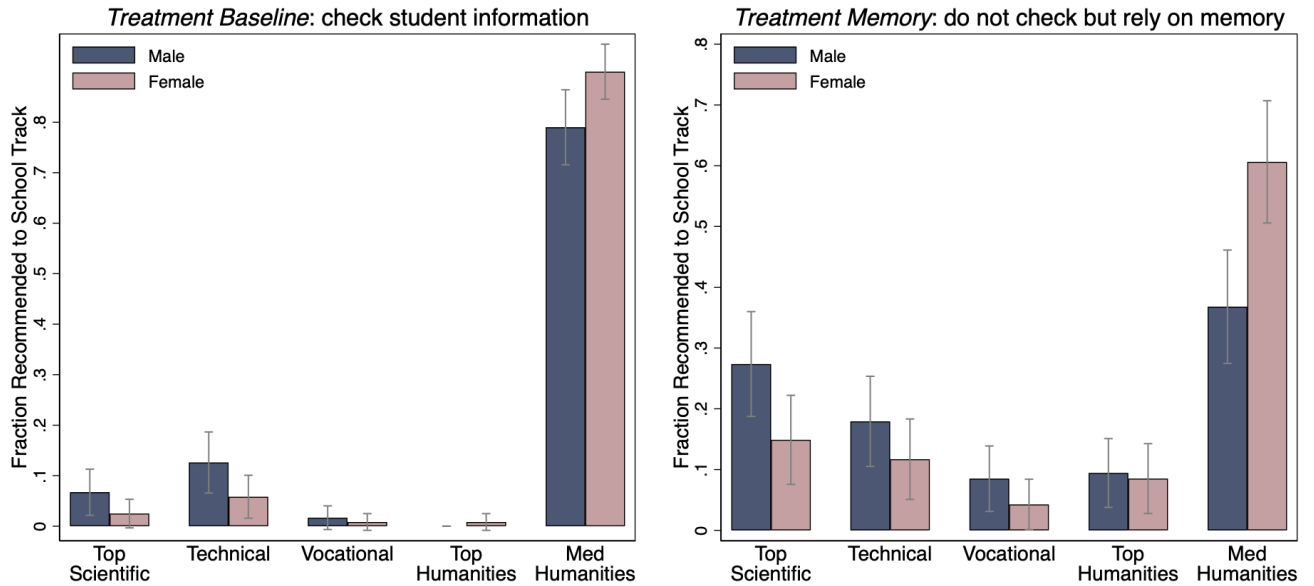
**Figure A27:** *Roberto/a*: Excellent Student in Math and Italian

**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 A28:** *Francesco/a*: Average student passionate about languages

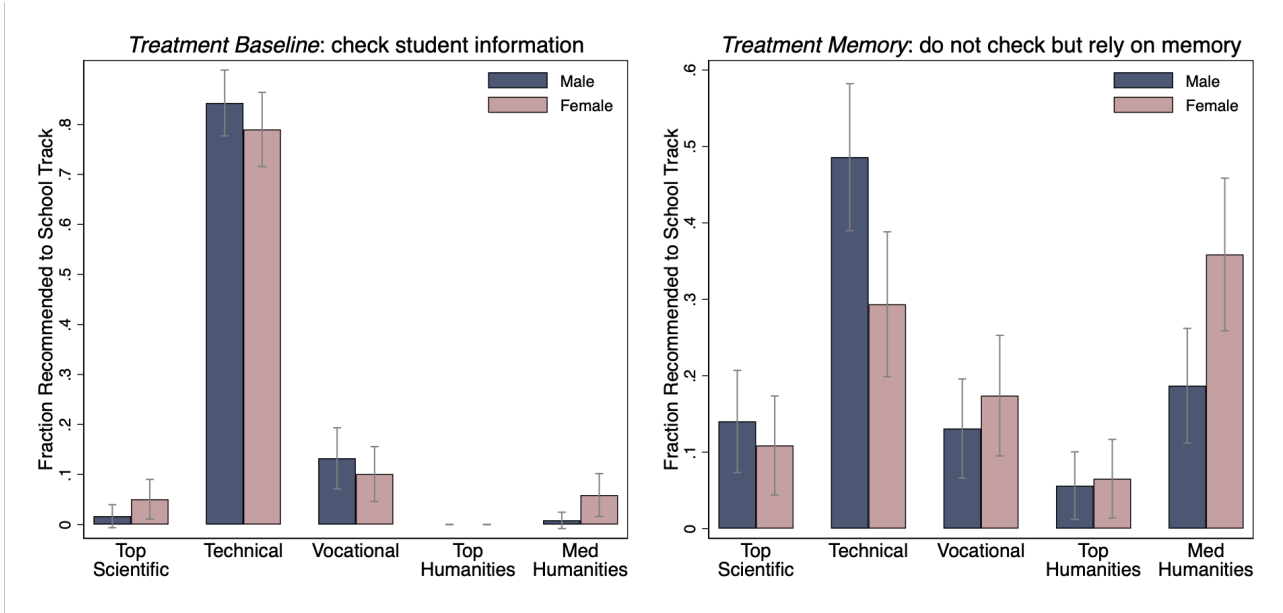
**Vignette:** *Francesco/a* is a good student but not excellent. He got 7 in math and Italian, and his GPA is around 8. He is also very passionate about languages, and he got an 8 in English. He spent 3 weeks in Ireland in the summer where he substantially improved his ability to speak English, which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





**Figure A29:** *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 A30:** *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".

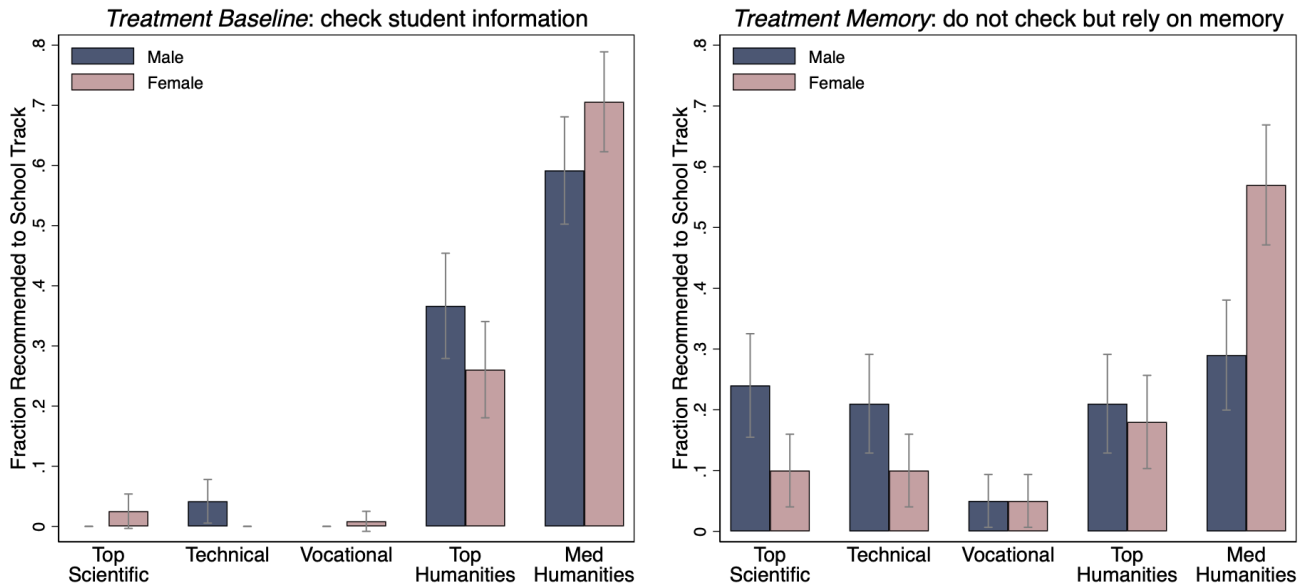
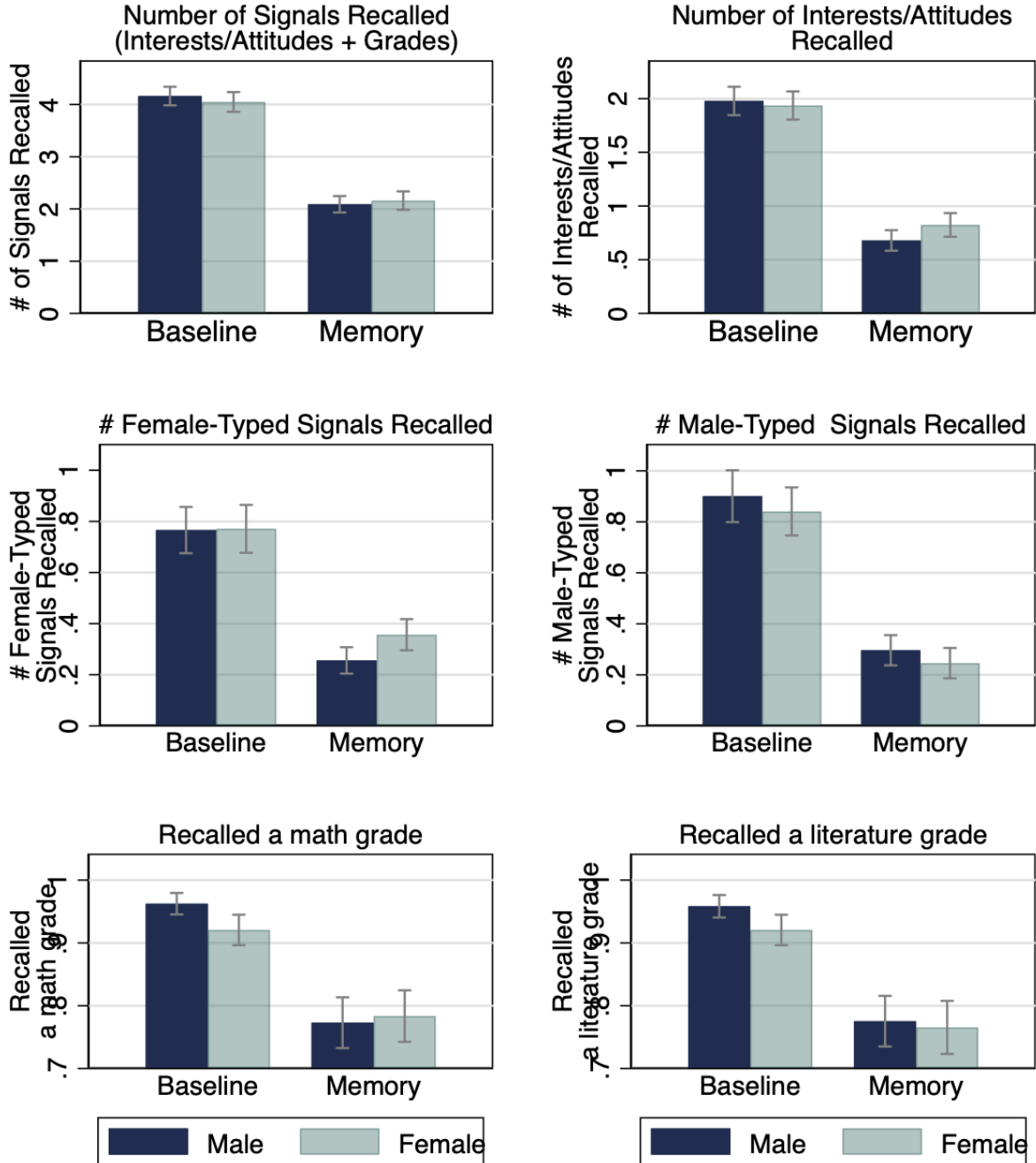


Figure A31: Average recall for all profiles

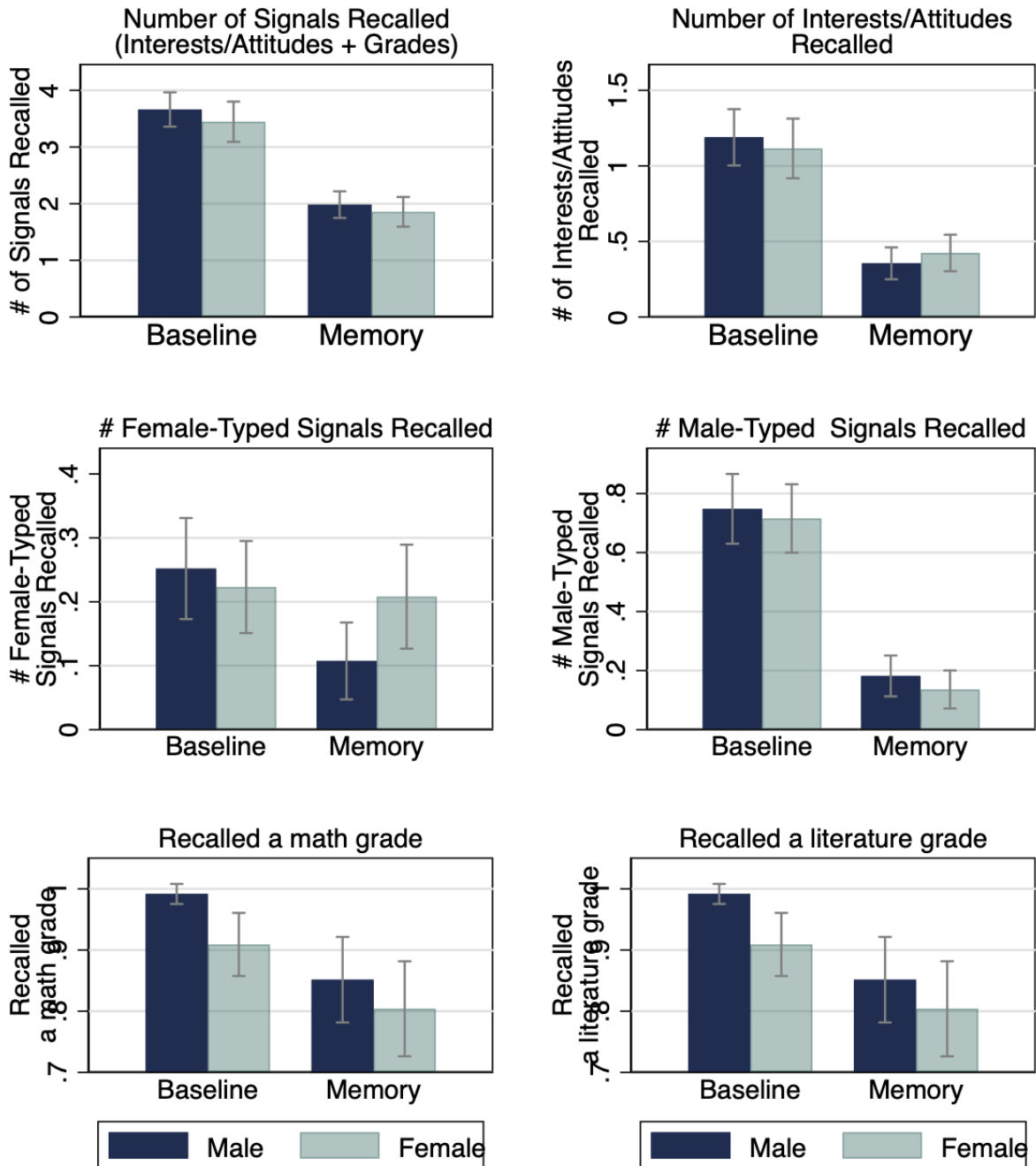
Average recall for all student profiles



**Figure A32:** Recall for Excellent Student in Math and Italian (mixed signals)

**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.

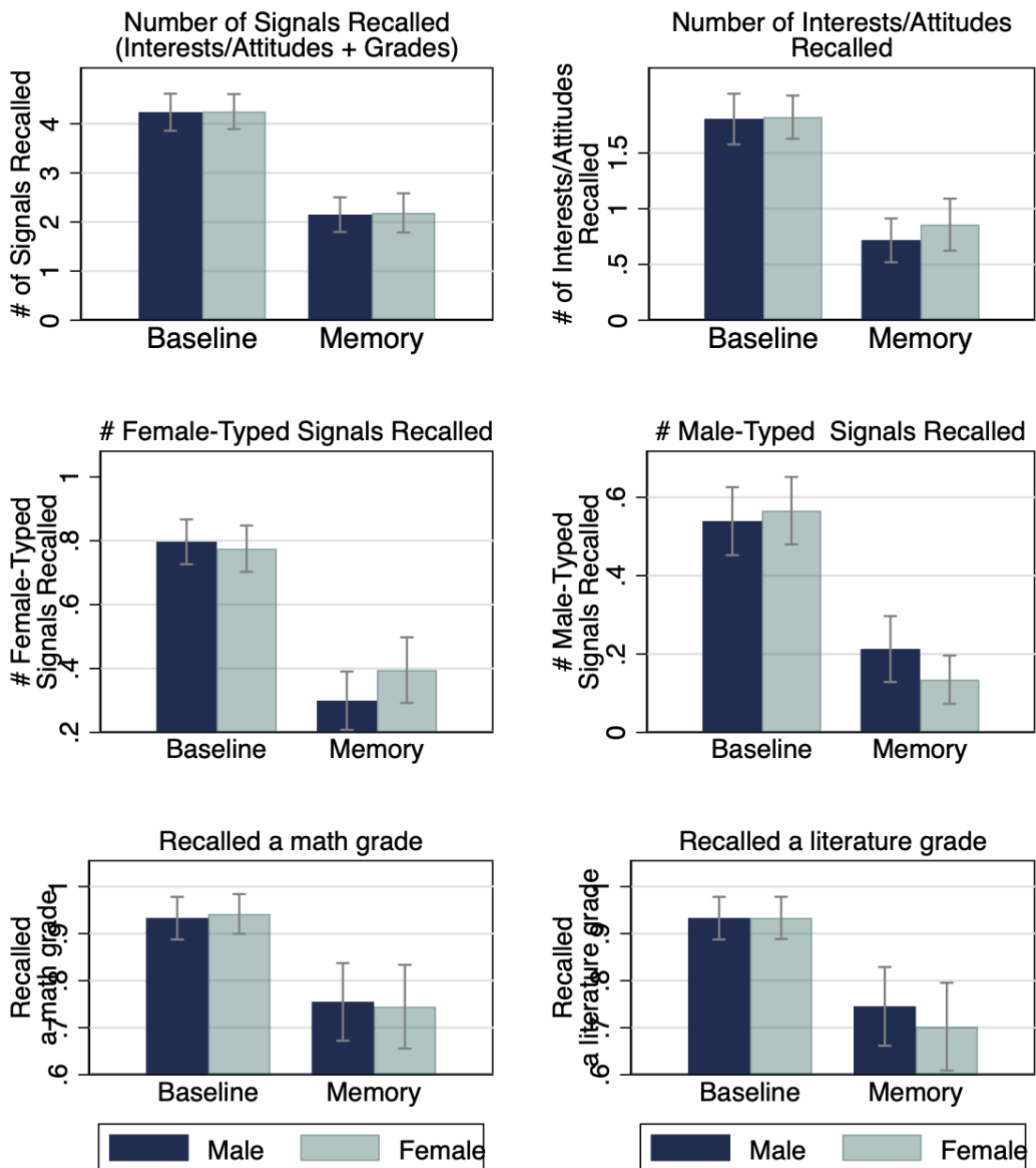
**Roberto/a: Excellent Student in Math and Italian**



**Figure A33:** Average student in math and literature passionate about languages (mixed signals)

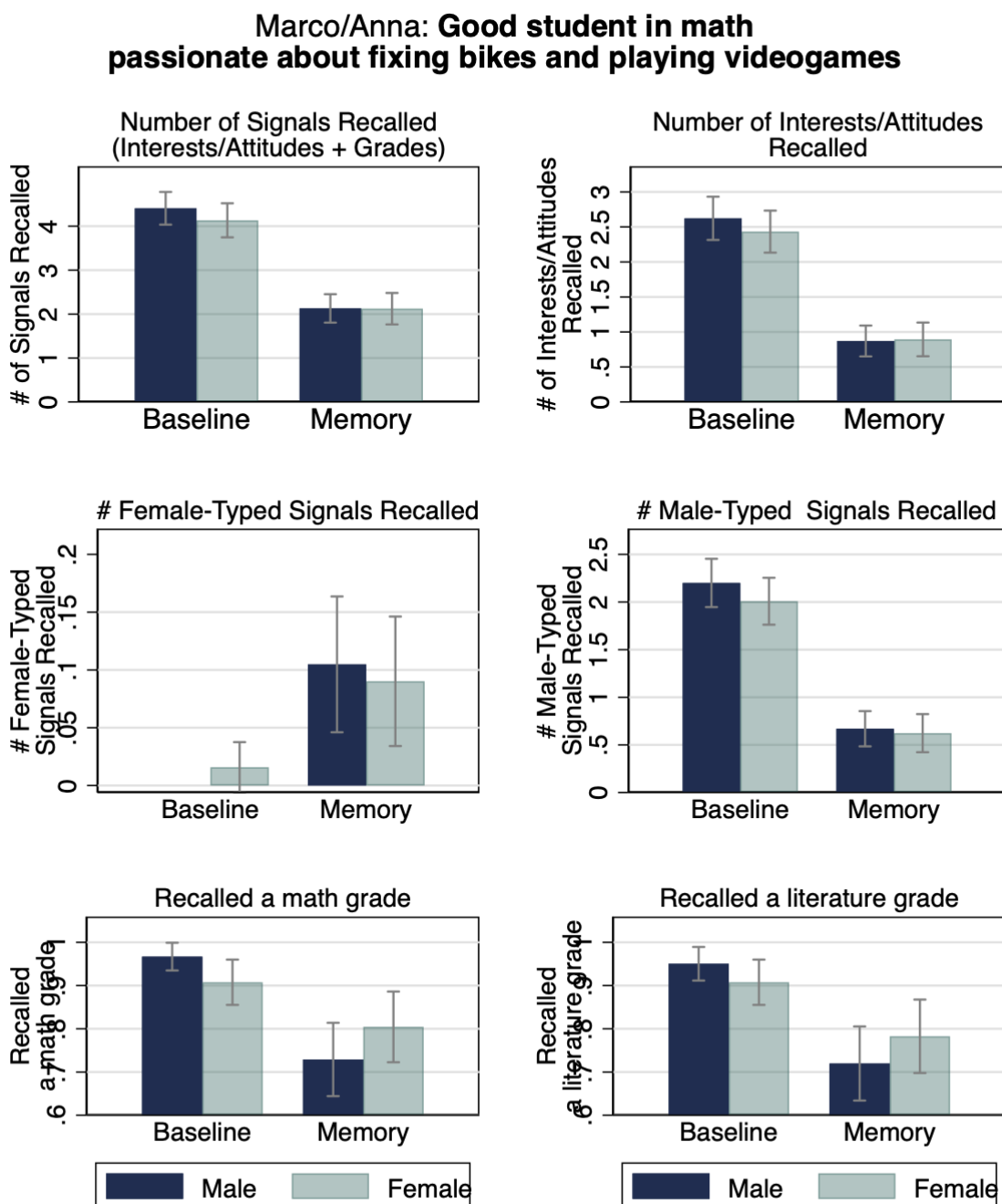
**Vignette:** *Francesco/a is a good student but not excellent. He got 7 in math and Italian, and his GPA is around 8. He is also very passionate about languages, and he got an 8 in English. He spent 3 weeks in Ireland in the summer where he substantially improved his ability to speak English, which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

**Francesco/a: Average student good at languages**



**Figure A34:** Good student in math passionate about fixing bikes and playing video games (stereotypically male)

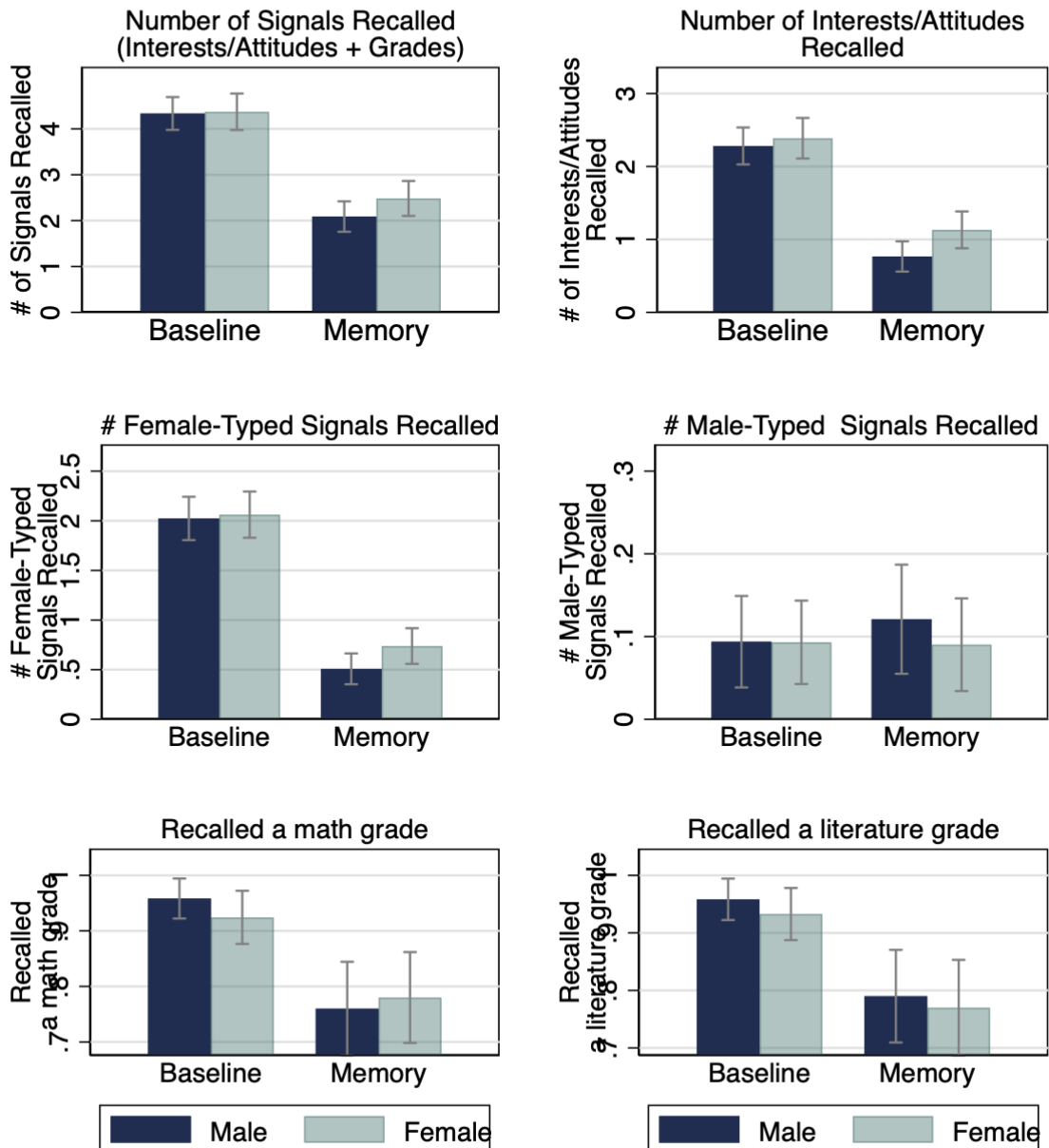
**Vignette:** *Marco/Anna is a very extroverted and social boy. He is not very diligent at school and he often forgets to do his homework. He often disrupts lectures by chatting with his friends. He is very intuitive and talented in math, where he got 8, while he got 6 in Italian. He is passionate about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



**Figure A35:** Good student in literature passionate about poetry (stereotypically female)

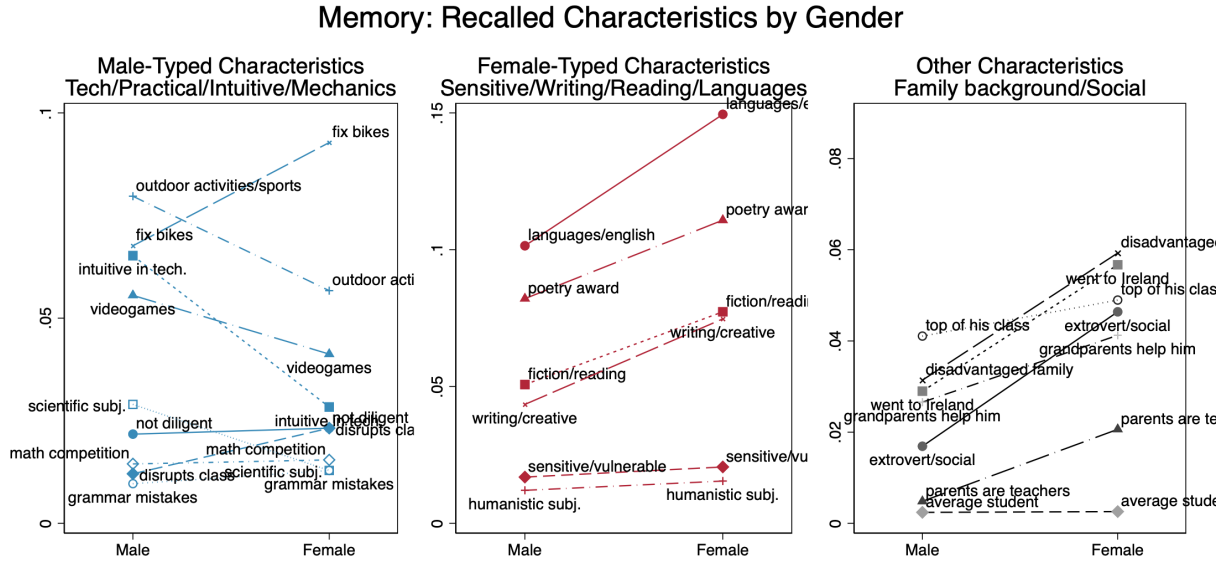
**Vignette:** Carlo/a comes from a disadvantaged family background. His father left when he was 5, his mother had some health issues and he mainly lives with his grandparents. However, his grandparents support him a lot in his education, and he manages to do quite well at school. He got a 6 in math and 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".

**Carlo/a: Good student in Literature passionate about poetry**



## D.6 Biases in Recall

Figure A36: Recalled Characteristics, by Student's Gender

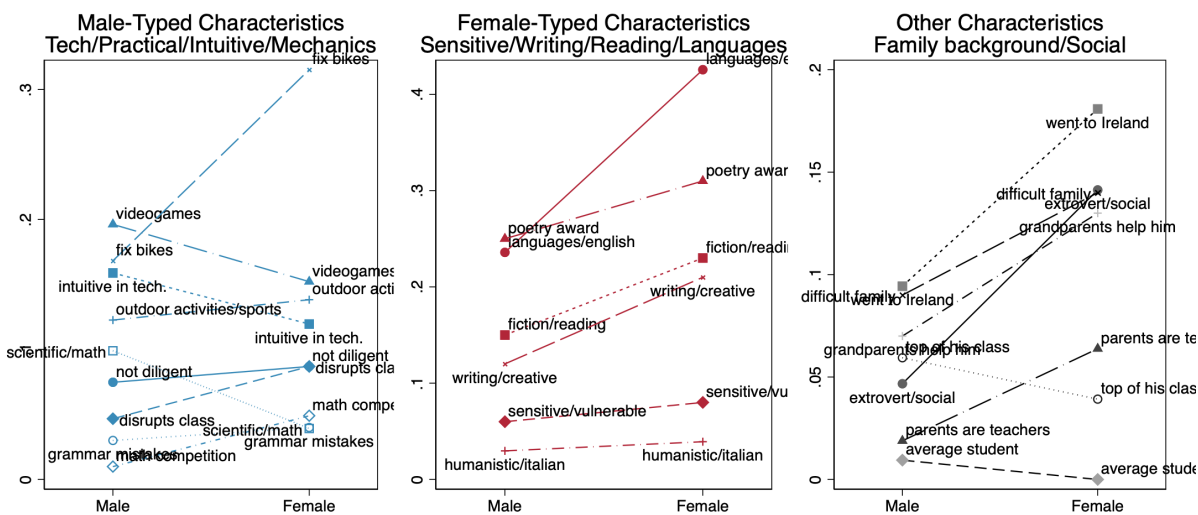


Notes: The figures show the probability that each characteristic is recalled when prompted to think about a student profile of a different gender. The left graph shows male-typed characteristics, the graph in the center shows female-typed characteristics and the right graph shows other characteristics. The sample includes teachers in the memory condition of the experiment.



Figure A37: Selective Memories

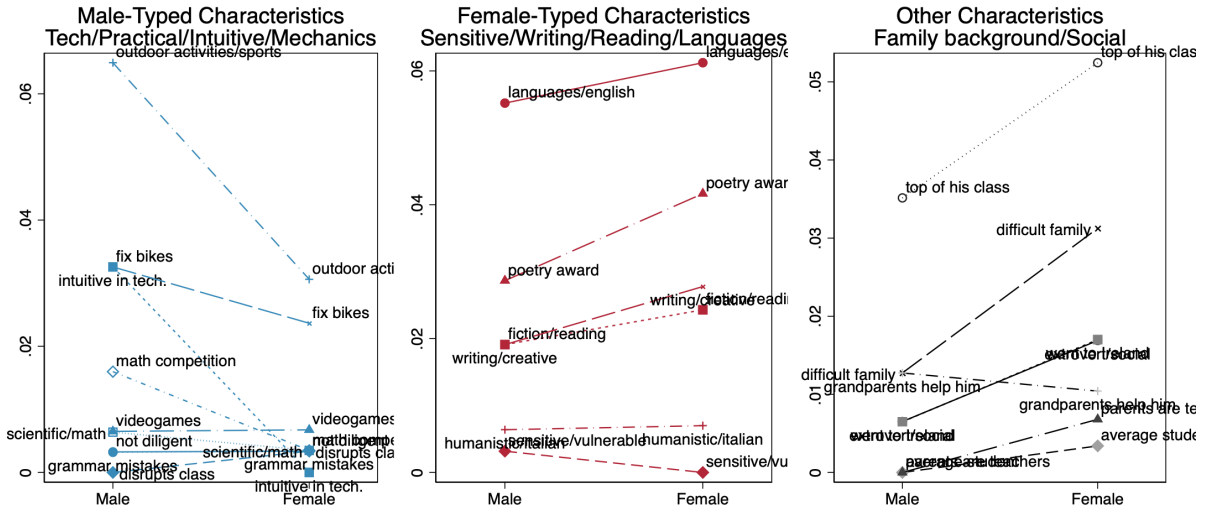
Selective Memories of **Correct Student**



Notes: The figures show the probability that each characteristic is recalled when prompted to think about the student profile originally associated with the characteristic. The left graph shows male-typed characteristics, the graph in the center shows female-typed characteristics and the right graph shows other characteristics. The sample includes teachers in the memory condition of the experiment.

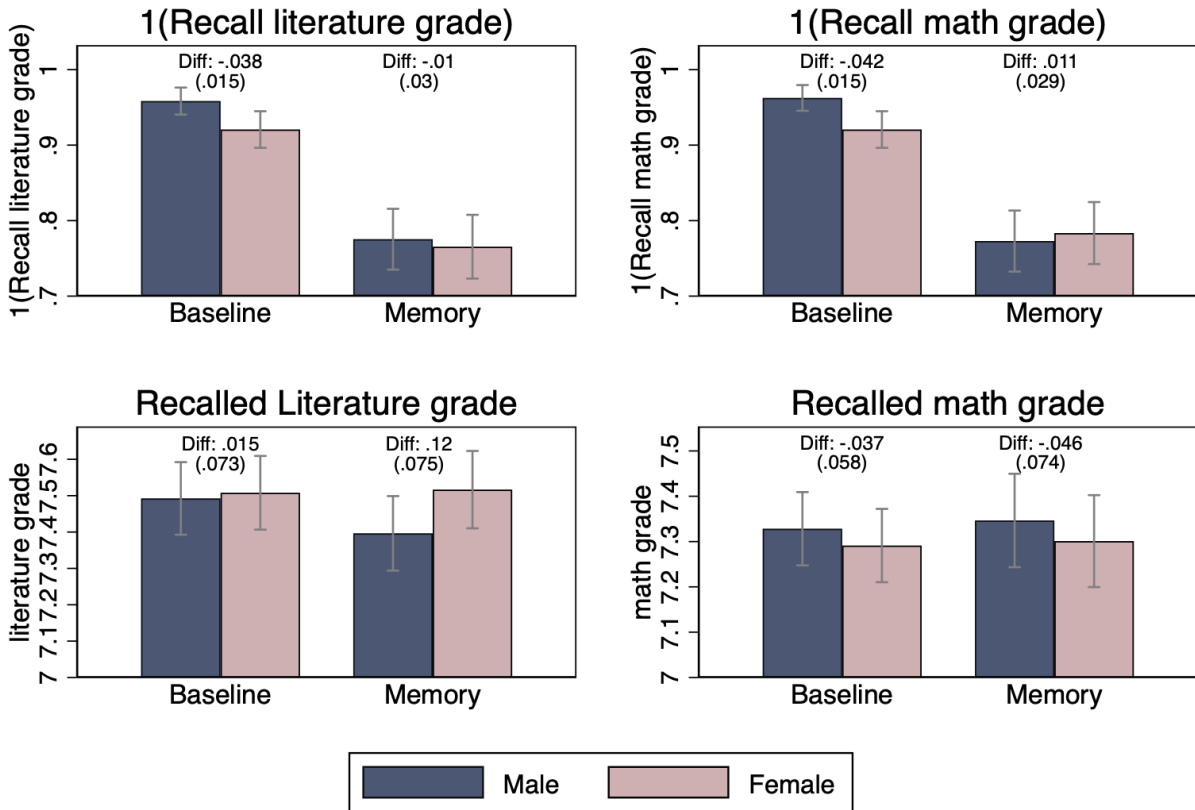
Figure A38: False Memories

False Memories - Characteristics of **Other Students**



Notes: The figures show the average probability that each characteristic is recalled when prompted to think about the three student profiles originally not associated with the characteristic. The left graph shows male-typed characteristics, the graph in the center shows female-typed characteristics and the right graph shows other characteristics. The sample includes teachers in the memory condition of the experiment.

## Limited and selective memory of grades



**Figure A39:** Experiment with Teachers: Memory of grades

*Notes:* This figure shows the probability that teachers recall literature and math grades (upper graphs), and the actual literature and math grades recalled (bottom graphs), for teachers in the memory and baseline conditions. Both teachers in the baseline and memory conditions observed the same students' profiles. Teachers in the memory condition need to retrieve students' characteristics from their memory (they do not have the profiles in front of them when they make evaluations), while teachers in the baseline condition can review students' characteristics before providing recommendations. The baseline sample included 443 teachers from 68 schools who completed the teachers' experiment.

**Table A16:** Limited recall of grades

|                        | Dv: 1(Recall literature grade) |                      | 1(Recall math grade) |                      |
|------------------------|--------------------------------|----------------------|----------------------|----------------------|
|                        | (1)                            | (2)                  | (3)                  | (4)                  |
| Female                 | -0.038**<br>(0.019)            | -0.039**<br>(0.020)  | -0.042**<br>(0.018)  | -0.043**<br>(0.019)  |
| Female $\times$ Memory | 0.026<br>(0.037)               | 0.030<br>(0.037)     | 0.052<br>(0.036)     | 0.057<br>(0.035)     |
| Memory                 | -0.182***<br>(0.032)           | -0.179***<br>(0.033) | -0.189***<br>(0.031) | -0.187***<br>(0.032) |
| Mean control           | 0.940                          | 0.940                | 0.942                | 0.942                |
| Observations           | 1761                           | 1761                 | 1761                 | 1761                 |
| N. teachers            | 448                            | 448                  | 448                  | 448                  |
| R <sup>2</sup>         | 0.066                          | 0.089                | 0.062                | 0.088                |
| Student FE             | Yes                            | Yes                  | Yes                  | Yes                  |
| Controls               | No                             | Yes                  | No                   | Yes                  |

*Notes:* This table shows coefficients  $\beta_2$  and  $\beta_3$  from estimation of equations 9 where the depend variables are dummies indicating whether the teacher recalled math grades (columns 1-2) and literature grades (columns 3-4). Teachers in the baseline sample are included. 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 A17:** Selective recall of grades

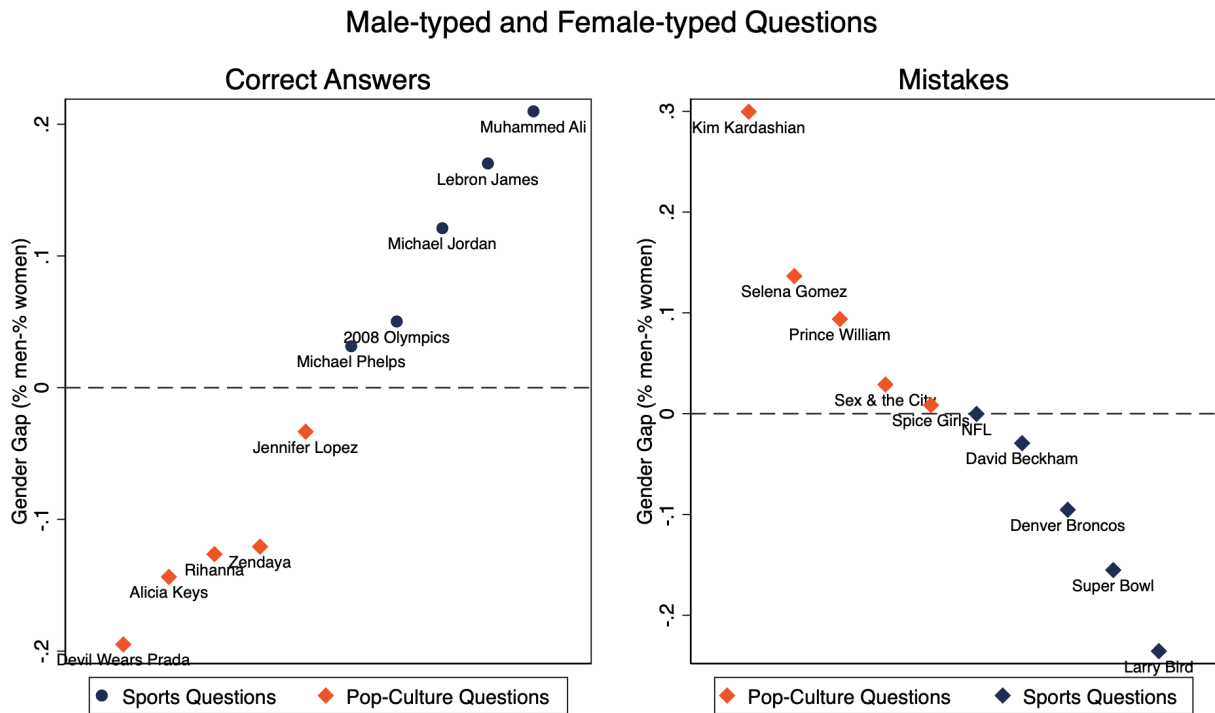
|                        | Recalled literature grade |                  |                  | Recalled math grade |                   |                   | Recalled gap lit-math |                  |                  |
|------------------------|---------------------------|------------------|------------------|---------------------|-------------------|-------------------|-----------------------|------------------|------------------|
|                        | (1)                       | (2)              | (3)              | (4)                 | (5)               | (6)               | (7)                   | (8)              | (9)              |
| Female                 | 0.023<br>(0.032)          | 0.016<br>(0.032) | 0.021<br>(0.038) | -0.022<br>(0.025)   | -0.019<br>(0.026) | -0.011<br>(0.030) | 0.042<br>(0.050)      | 0.034<br>(0.050) | 0.034<br>(0.059) |
| Female $\times$ Memory | 0.105<br>(0.076)          | 0.108<br>(0.076) | 0.106<br>(0.088) | -0.000<br>(0.061)   | -0.009<br>(0.061) | -0.093<br>(0.078) | 0.088<br>(0.102)      | 0.101<br>(0.103) | 0.181<br>(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 <sup>2</sup>         | 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  $\beta_2$  and  $\beta_3$  from the estimation of equations 9 where the dependent variables are the recalled literature grades (columns 1-2), the recalled math grades (columns 3-4), and the gap between the recalled literature-math grades (columns 7-9). Columns (3), (6), (9) include teachers fixed effects. Teachers in the baseline sample are included. 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.

# E Additional Figures and Tables for Prolific Experiment

## E.1 Design

Figure A40: Prolific Experiment: Male-typed and female-typed signals



*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.

**SPORTS: LeBron James**

Which of the following is a nickname for LeBron James?


- a. The answer
- b. The truth
- c. Mr. big shot
- d. The king

**SPORTS: LeBron James**

Which of the following is a nickname for LeBron James?

- a. The answer
- b. The truth
- c. Mr. big shot
- d. The king**

**John's answer: The king**

 **John gave the right answer!**

**SPORTS: LeBron James**

Which of the following is a nickname for LeBron James?

- a. The answer
- b. The truth
- c. Mr. big shot
- d. The king

**SPORTS: LeBron James**

Which of the following is a nickname for LeBron James?

- a. The answer
- b. The truth
- c. Mr. big shot
- d. The king**

**Susan's answer: The king**

 **Susan gave the right answer!**

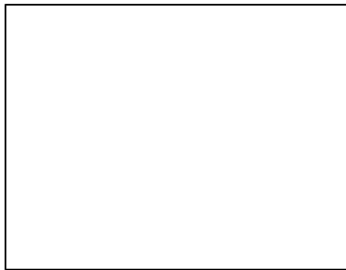
**Figure A41:** More abstract experiment, example of a question

**Part C**

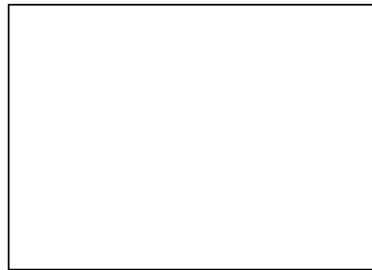
In this section, you are asked to recall all the questions that **John** answered correctly and incorrectly in sports and celebrity-pop culture.

*If this part is selected for payment, you will increase your likelihood to get the bonus payment by writing down a higher number of questions in the appropriate box. Each question can be ONLY written in one box (either in the correct-answer box or in the incorrect-answer box); otherwise, it is not accounted for. There is no penalty for wrong guesses, but only correct questions count for the bonus payment. You will earn 10 points for each question recalled.*

**John answered the following questions correctly:**  
(you may report either the question's title, the question's answer, or a clearly identifiable question topic):



**John answered the following questions incorrectly:**  
(you may report either the question's title, the question's answer, or a clearly identifiable question topic):

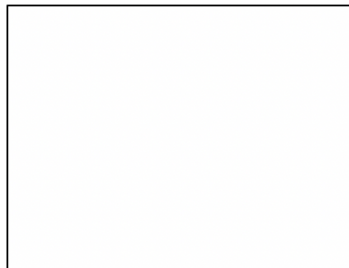


**Part C**

In this section, you are asked to recall all the questions that **John** answered correctly and incorrectly in sports and celebrity-pop culture.

**You can review the list of questions**, together with John's answers, by clicking this **link** (please, open this link in a new tab).

**John answered the following questions correctly:**  
(you may report either the question's title, the question's answer, or a clearly identifiable question topic):



**John answered the following questions incorrectly:**  
(you may report either the question's title, the question's answer, or a clearly identifiable question topic):



**Figure A42:** More abstract experiment, free recall task



## E.2 Balance

**Table A18:** Balance Prolific Experiment

| Variable                             | (1)<br>Baseline | (2)<br>SD | (3)<br>Memory | (4)<br>SD | (5)<br>Diff. | (6)<br>P-val. |
|--------------------------------------|-----------------|-----------|---------------|-----------|--------------|---------------|
| <b>Baseline vs. Memory treatment</b> |                 |           |               |           |              |               |
| Age                                  | 34.569          | (12.171)  | 35.262        | (12.989)  | 0.693        | (0.334)       |
| Gender (1=male)                      | 0.514           | (0.500)   | 0.494         | (0.500)   | -0.021       | (0.470)       |
| Republican                           | 0.119           | (0.324)   | 0.118         | (0.323)   | -0.000       | (0.983)       |
| More than high school                | 0.868           | (0.339)   | 0.843         | (0.364)   | -0.024       | (0.222)       |
| Number rejections on Prolific        | 2.472           | (3.353)   | 2.588         | (3.129)   | 0.116        | (0.530)       |
| Number of approvals on Prolific      | 826.489         | (609.559) | 786.031       | (593.218) | -40.458      | (0.237)       |
| Full time work                       | 0.430           | (0.496)   | 0.471         | (0.500)   | 0.041        | (0.270)       |
| <b>Male vs. Female treatment</b>     |                 |           |               |           |              |               |
| Variable                             | (1)<br>Male     | (2)<br>SD | (3)<br>Female | (4)<br>SD | (5)<br>Diff. | (6)<br>P-val. |
| Age                                  | 34.883          | (12.358)  | 34.982        | (12.864)  | 0.099        | (0.890)       |
| Gender (1=male)                      | 0.528           | (0.500)   | 0.479         | (0.500)   | -0.049       | (0.084)*      |
| Republican                           | 0.128           | (0.334)   | 0.109         | (0.312)   | -0.019       | (0.305)       |
| More than high school                | 0.856           | (0.351)   | 0.853         | (0.354)   | -0.003       | (0.898)       |
| Number rejections on Prolific        | 2.674           | (3.152)   | 2.389         | (3.317)   | -0.284       | (0.122)       |
| Number of approvals on Prolific      | 820.485         | (595.647) | 789.770       | (606.770) | -30.714      | (0.369)       |
| Full time work                       | 0.437           | (0.497)   | 0.465         | (0.499)   | 0.028        | (0.457)       |
| Observations                         | 625             |           | 614           |           | 1,239        |               |

### E.3 Sensitivity

**Table A19:** Sensitivity to including only very attentive respondents, and respondents who took more or less time than the median to complete the survey

|                                       | DV: Estimated share correct new task |                        |                       | DV: Recalled share correct in old task |                        |                       | DV: Total number recalled questions |                       |                       |
|---------------------------------------|--------------------------------------|------------------------|-----------------------|--|------------------------|-----------------------|-------------------------------------|-----------------------|-----------------------|
|                                       | (1)                                  | (2)                    | (3)                   | (4)                                    | (5)                    | (6)                   | (7)                                 | (8)                   | (9)                   |
| Stereotype-Consistent $\times$ Memory | 0.0319***<br>(0.00941)               | 0.0329***<br>(0.0114)  | 0.0290**<br>(0.0115)  | 0.0407**<br>(0.0206)                   | 0.0760***<br>(0.0281)  | 0.0285<br>(0.0279)    | -0.0351<br>(0.121)                  | -0.212<br>(0.142)     | -0.0789<br>(0.145)    |
| Stereotype-Consistent                 | 0.0160***<br>(0.00547)               | 0.0201***<br>(0.00709) | 0.0145**<br>(0.00702) | -0.00826<br>(0.00518)                  | -0.0208**<br>(0.00982) | 0.00188<br>(0.0122)   | 0.0140<br>(0.0771)                  | 0.00172<br>(0.0942)   | 0.112<br>(0.0966)     |
| Memory                                | -0.00857<br>(0.00755)                | -0.00713<br>(0.00953)  | -0.00790<br>(0.00961) | -0.0405***<br>(0.0127)                 | -0.0526***<br>(0.0161) | -0.0334*<br>(0.0172)  | -4.591***<br>(0.189)                | -3.880***<br>(0.247)  | -4.741***<br>(0.236)  |
| R-squared                             | 0.0908                               | 0.0906                 | 0.0984                | 0.0258                                 | 0.0300                 | 0.0307                | 0.471                               | 0.420                 | 0.519                 |
| N. Obs                                | 1696                                 | 1240                   | 1238                  | 1562                                   | 1123                   | 1119                  | 1696                                | 1240                  | 1238                  |
| Mean DV                               | 0.507                                | 0.519                  | 0.505                 | 0.463                                  | 0.462                  | 0.467                 | 0.463                               | 0.462                 | 0.467                 |
| domain FE                             | ✓                                    | ✓                      | ✓                     | ✓                                      | ✓                      | ✓                     | ✓                                   | ✓                     | ✓                     |
| controls                              | ✓                                    | ✓                      | ✓                     | ✓                                      | ✓                      | ✓                     | ✓                                   | ✓                     | ✓                     |
| Sample                                | Attentive                            | Duration above median  | Duration below median | Attentive                              | Duration above median  | Duration below median | Attentive                           | Duration above median | Duration below median |

**Table A20:** Assessment of Candidate Ability in Sports and pop-culture

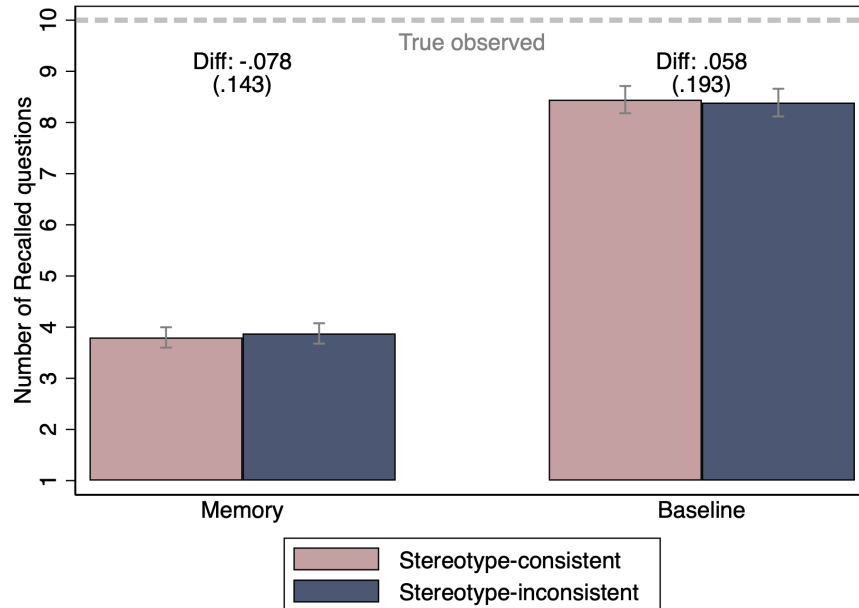
|                        | DV: Sports              |                         | Pop-culture          |                      |
|------------------------|-------------------------|-------------------------|----------------------|----------------------|
|                        | (1)                     | (2)                     | (3)                  | (4)                  |
| Female $\times$ Memory | -0.0357***<br>(0.0129)  | -0.0375***<br>(0.0129)  | 0.0249**<br>(0.0127) | 0.0255**<br>(0.0128) |
| Female                 | -0.0289***<br>(0.00811) | -0.0249***<br>(0.00814) | 0.00555<br>(0.00835) | 0.00604<br>(0.00847) |
| Memory                 | 0.0115<br>(0.00903)     | 0.0119<br>(0.00913)     | 0.00795<br>(0.00906) | 0.00881<br>(0.00928) |
| R-squared              | 0.0473                  | 0.0928                  | 0.0179               | 0.0464               |
| N. Obs                 | 1239                    | 1239                    | 1239                 | 1239                 |
| Mean DV                | 0.513                   | 0.513                   | 0.550                | 0.550                |
| Sd Dependent Variable  | 0.118                   | 0.118                   | 0.113                | 0.113                |
| domain FE              | ✓                       | ✓                       | ✓                    | ✓                    |
| controls               |                         | ✓                       |                      | ✓                    |

**Table A21:** Recalled share correct questions in Sports and pop culture

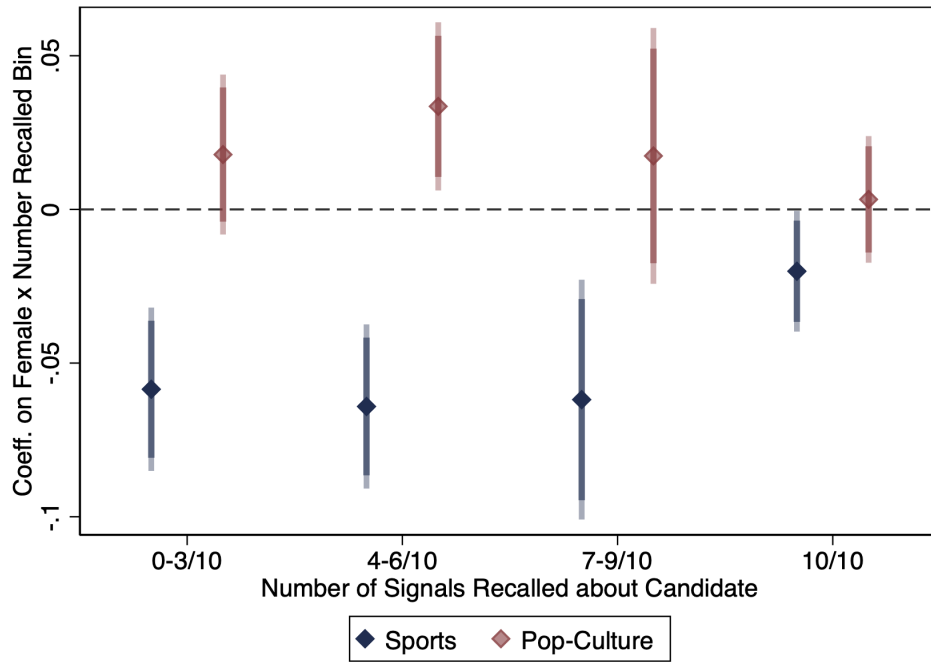
|                 | DV: Sports            |                      | Pop-culture           |                       |
|-----------------|-----------------------|----------------------|-----------------------|-----------------------|
|                 | (1)                   | (2)                  | (3)                   | (4)                   |
| Female × Memory | -0.0501**<br>(0.0227) | -0.0429*<br>(0.0232) | 0.0538**<br>(0.0248)  | 0.0526**<br>(0.0249)  |
| Female          | 0.00885<br>(0.00797)  | 0.00280<br>(0.00863) | -0.0113<br>(0.00952)  | -0.00794<br>(0.0104)  |
| Memory          | 0.00624<br>(0.0158)   | 0.00350<br>(0.0166)  | -0.0381**<br>(0.0166) | -0.0381**<br>(0.0171) |
| R-squared       | 0.00809               | 0.0369               | 0.00634               | 0.0360                |
| N. Obs          | 1130                  | 1130                 | 1112                  | 1112                  |
| Mean DV         | 0.502                 | 0.502                | 0.468                 | 0.468                 |
| domain FE       | ✓                     | ✓                    | ✓                     | ✓                     |
| controls        |                       | ✓                    |                       | ✓                     |

## E.4 Biases in Recall

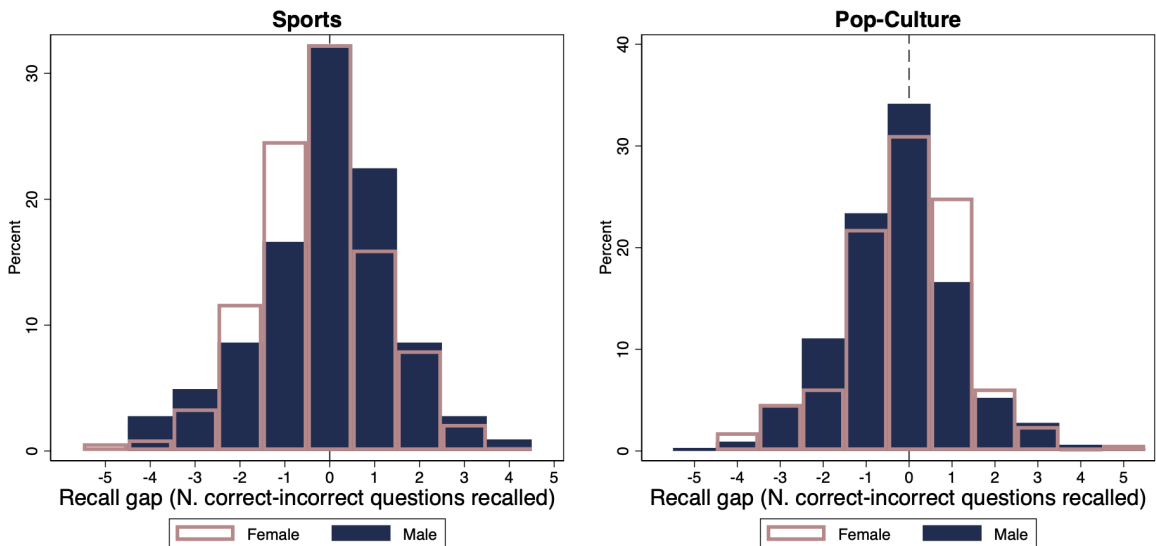
**Figure A43:** Number of questions recalled for stereotype-consistent vs. inconsistent candidate in a domain (male is stereotype-consistent in sports, female in pop culture)



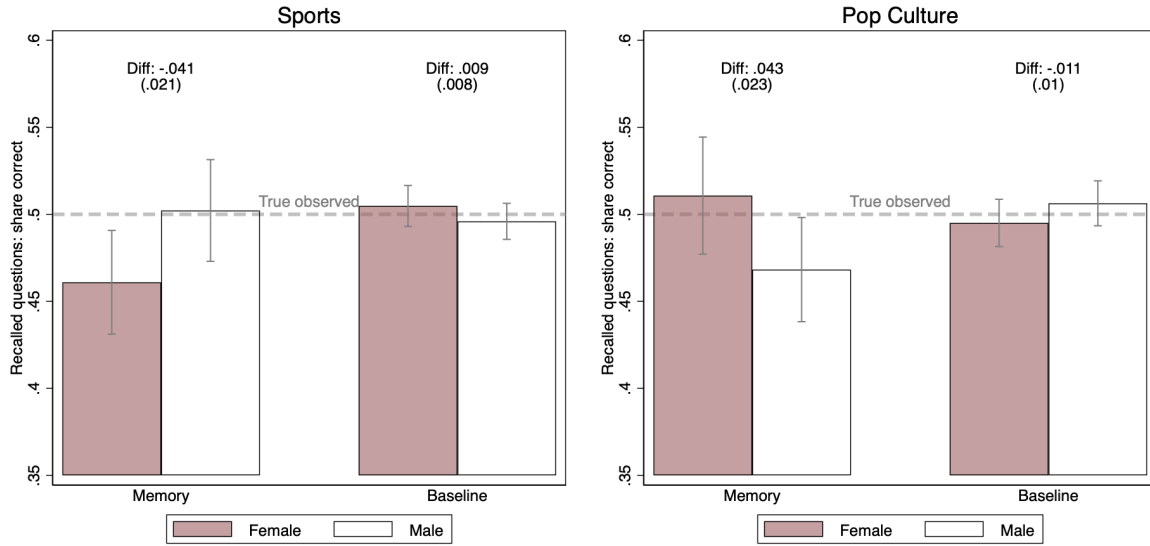
**Figure A44:** Sanity Check 2: Assessment of Ability and Number of Signals Recalled



**Figure A45:** Memory Treatment: Number of Positive minus Negative Signals Recalled about Candidate



**Figure A46:** Share correct questions among recalled signals by domain and memory vs. baseline treatment



**Table A22:** Experiment 2 (More abstract experiment): Selective memory of Questions

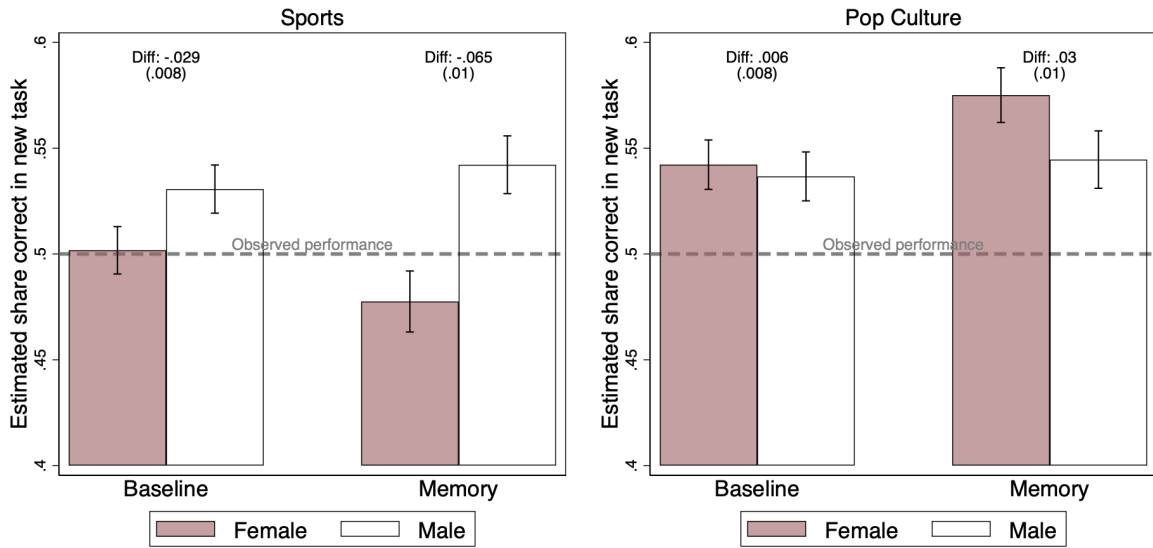
|                       | Recalled Success Ratio |                        | N. Correct Questions  |                       | N. Incorrect Questions |                       |
|-----------------------|------------------------|------------------------|-----------------------|-----------------------|------------------------|-----------------------|
|                       | (1)                    | (2)                    | (3)                   | (4)                   | (5)                    | (6)                   |
| Congruent             | -0.0101<br>(0.00759)   | -0.0101<br>(0.00764)   | 0.0153<br>(0.0402)    | 0.0123<br>(0.0399)    | 0.0424<br>(0.0365)     | 0.0402<br>(0.0367)    |
| Memory                | -0.0409***<br>(0.0116) | -0.0410***<br>(0.0117) | -2.393***<br>(0.0892) | -2.396***<br>(0.0879) | -2.119***<br>(0.0926)  | -2.124***<br>(0.0903) |
| Congruent × Memory    | 0.0517***<br>(0.0194)  | 0.0520***<br>(0.0196)  | 0.0355<br>(0.0685)    | 0.0385<br>(0.0683)    | -0.172***<br>(0.0662)  | -0.169**<br>(0.0663)  |
| R-squared             | 0.00703                | 0.0187                 | 0.367                 | 0.417                 | 0.319                  | 0.369                 |
| N. Obs                | 2242                   | 2242                   | 2478                  | 2478                  | 2478                   | 2478                  |
| controls              |                        | ✓                      |                       | ✓                     |                        | ✓                     |
| Mean DV               | 0.493                  | 0.493                  | 2.967                 | 2.967                 | 3.048                  | 3.048                 |
| Sd Dependent Variable | 0.205                  | 0.205                  | 1.959                 | 1.959                 | 1.951                  | 1.951                 |

Standard errors in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

## E.5 Bias in Assessment of Ability by Domain

**Figure A47:** Share correct questions among recalled signals by domain and memory vs. baseline treatment



# F Prior Beliefs in Teacher and Prolific Experiments

## F.1 Prolific Experiment

Figure A48: Prior Belief on Number of Correct Questions

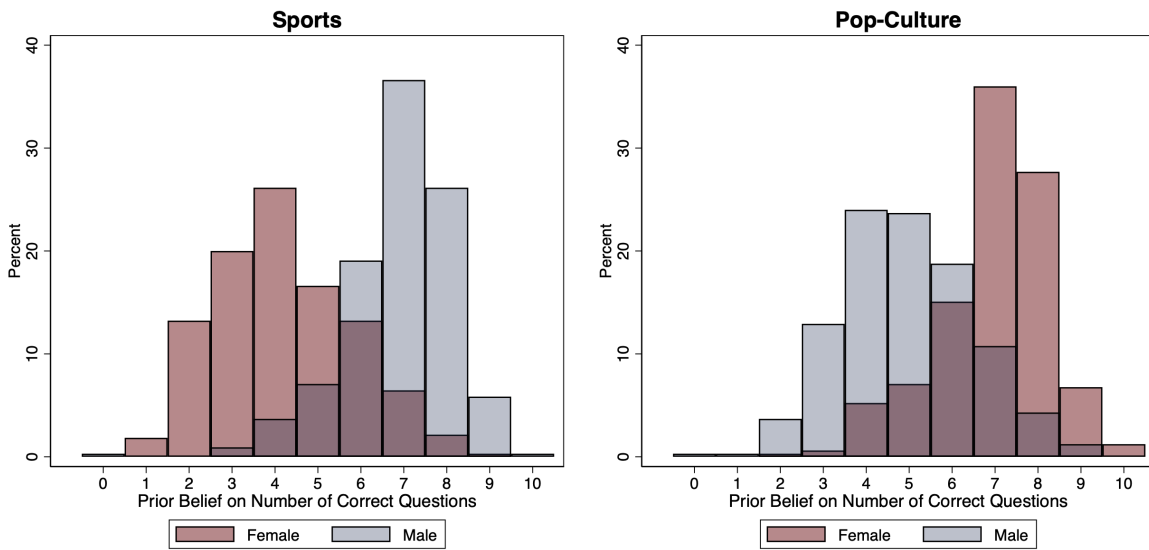
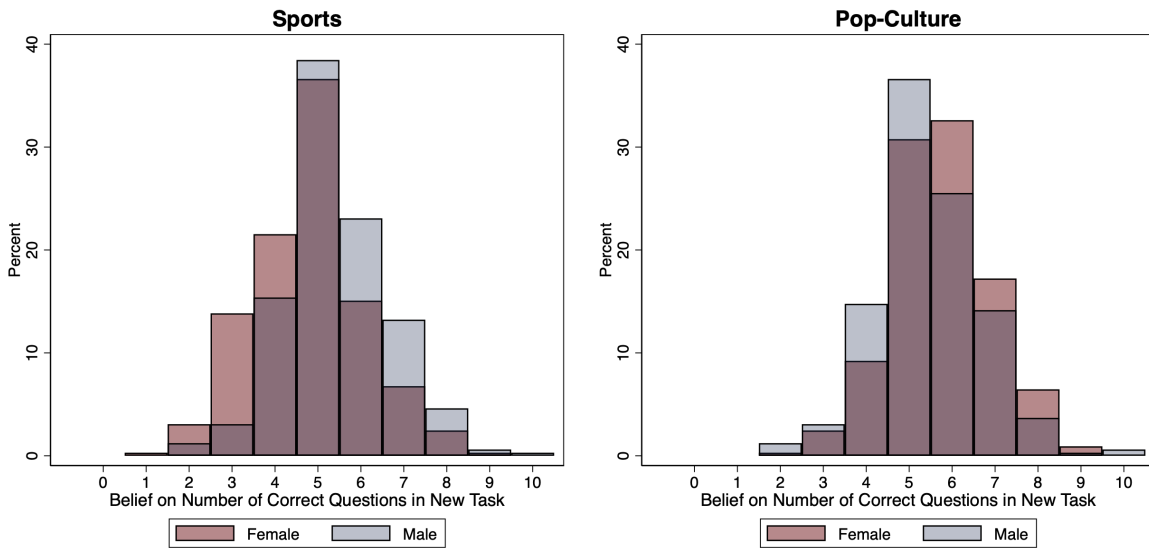
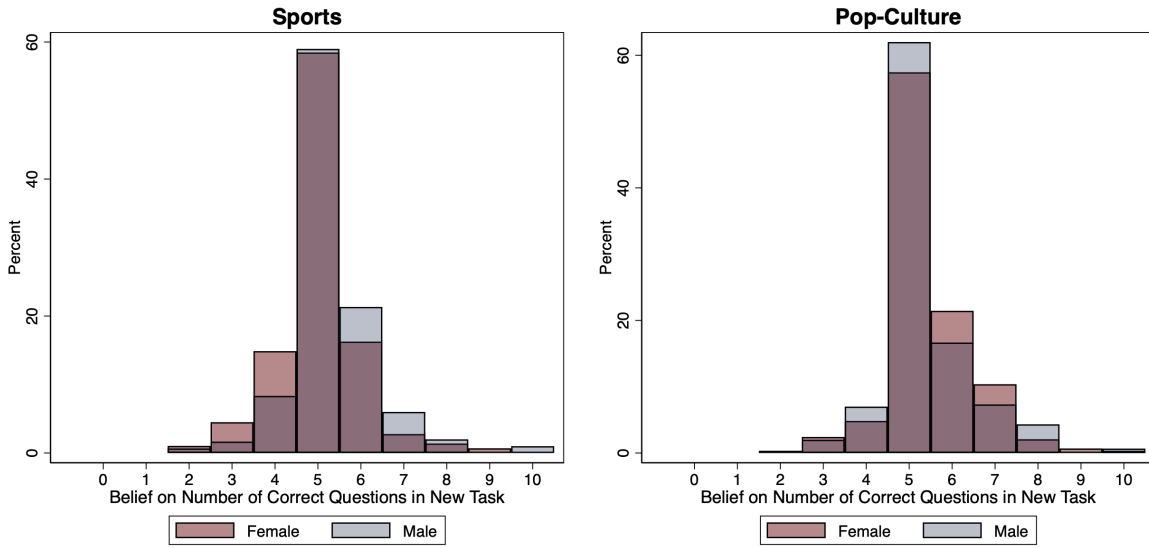


Figure A49: Memory Treatment: Estimated Number of Correct Questions in New Task

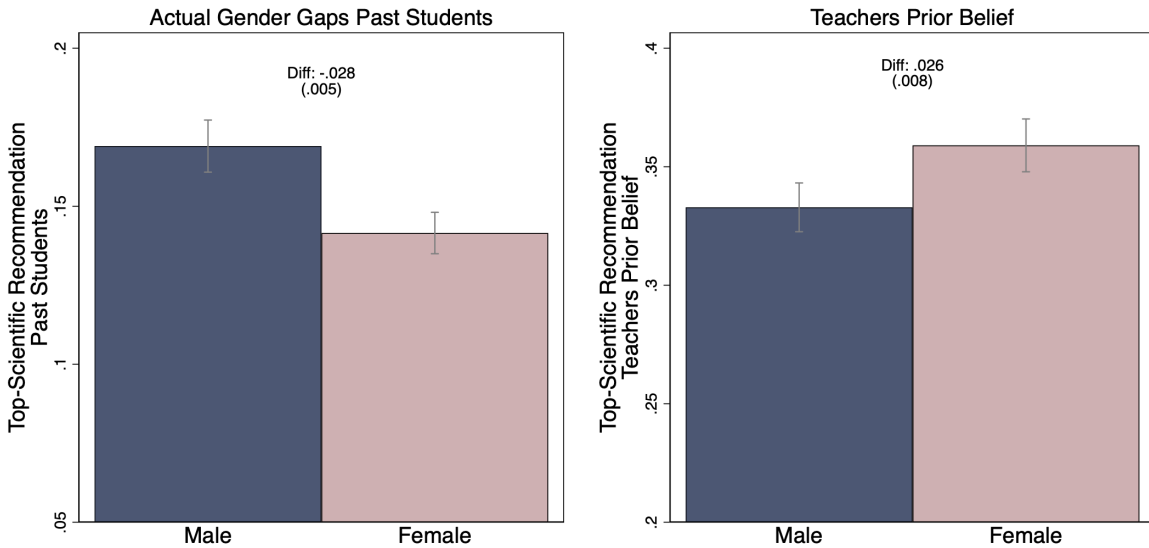


**Figure A50:** Baseline Treatment: Estimated Number of Correct Questions in New Task



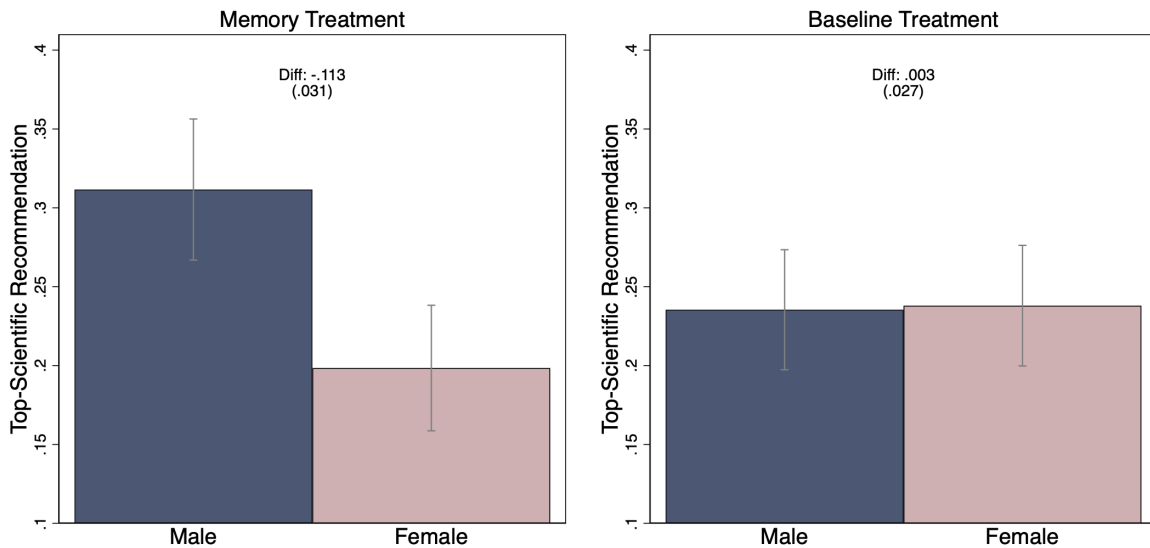
## F.2 Teachers Experiment

**Figure A51:** Teachers' Actual Behaviors with Past Students (left) and Elicited Prior Belief (right)





**Figure A52:** Teachers' Behavior in Memory Treatment (left) and Baseline Treatment (right)



## G Data Appendix

### G.1 Teacher Survey and Experiment

**Teacher Survey Questions.** The survey for middle school teachers was implemented in February 2023. The teachers who participated in the survey were recruited in November 2022 to be part of a long-term collaboration aimed at providing teachers with information on the academic performance and school choices of their past students once they finished middle school.

**Classification of Students' Characteristics in Teacher Experiment.** Below I outline all the characteristics of the students' profiles, as well as the keywords used to identify them in the text.

| Roberto/a |  |   |
|-----------|--|---|
| Features: |  |   |
| 1         | <i>very smart/among the best students of his/her class</i> | brillante, eccell*, brav*, miglior*, tra i primi, intelligente, eccelle in tutte le materie |
| 2         | <i>last rounds in math competition</i>                     | olimpiadi, giochi mat*, gare, finale, competitiv*, competizione                             |
| 3         | <i>talented in scientific subjects</i>                     | scientific*   |
| 4         | <i>talented in humanistic subjects</i>                     | umanistic*  |
| Grades:   |  |   |
| 5         | <i>9 in Italian</i>  |   |
| 6         | <i>8 in math</i>   |   |
| 7         | <i>8.5 GPA</i>   |   |

| Carlo/a   |                                   |   |
|-----------|-----------------------------------|---|
| Features: |                                   |   |
| 1         | <i>difficult family situation</i> | genitor*, famil*, famiglia, situazione difficile, background diff*, background complicato |
| 2         | <i>grandparents help him/her</i>  | nonni   |
| 3         | <i>loves reading fiction</i>      | lettura, lettrici, lettore, romanzi, leggere, libri                                       |
| 4         | <i>grammar mistakes</i>           | errori grammatica, diffic* ortografiche   |
| 5         | <i>poetry</i>                     | poesi*, poeta   |
| 6         | <i>sensitive</i>                  | fragile, sensibile, emotiv*   |
| 7         | <i>creative in essays</i>         | temi, creativ*, scrittura, scrivere   |
| Grades    |                                   |   |
| 8         | <i>6 in math</i>                  |   |
| 9         | <i>8 in Italian</i>               |   |
| 10        | <i>8 GPA</i>                      |   |

| Francesco/a |   |                                  |
|-------------|---|----------------------------------|
| Features    |   |                                  |
| 1           | <i>good but not excellent</i>           | non eccelle                      |
| 2           | <i>passionate about languages</i>       | lingue, inglese, lingu*          |
| 3           | <i>went to Ireland</i>                  | Irlanda, vacanza studio, viaggi* |
| 4           | <i>outdoor activities with friends</i>  | amici, apert*, aria, sport       |
| 5           | <i>parents are high school teachers</i> | insegnanti                       |
| Grades:     |   |                                  |
| 6           | <i>7 in math</i>                        |                                  |
| 7           | <i>7 in italian</i>                     |                                  |
| 8           | <i>GPA 8, 8 in English</i>              |                                  |

| Marco/Anna |  |   |
|------------|--|---|
| Features   |  |   |
| 1          | <i>extrovert and social</i>                                  | estrovers*, social*, socievole*, esuberante   |
| 2          | <i>not diligent, doesn't do homework</i>                     | non fa i compiti, non diligente, poco diligente, indisciplinato, poco impegno, non si impegna |
| 3          | <i>disrupts lectures</i>                                     | vivace, distratt*, chiaccher*, disturb*   |
| 4          | <i>intuitive and good and talented in technical subjects</i> | intuitiv*, tecnolog*, meccanica, svegli*, tecnica, logica                                     |
| 5          | <i>loves playing videogames with friends</i>                 | pc, computer, video, informatica  |
| 6          | <i>in free time loves fixing bikes with brother</i>          | ripara*, bici, costruire, meccanic*, manual*, pratic*   |
| Grades:    |  |   |
| 7          | <i>8 in math</i>   |   |
| 8          | <i>6 Italian</i>   |   |
| 9          | <i>8 technology</i>  |   |