Motivated memory and favoritism*

Charlotte Cordes

February 19, 2026

The most recent version of the paper can be found here.

Abstract

Many important decisions that involve the judgment of others are based on recalled information. Because memory is noisy and potentially biased, it can influence how we judge people. This suggests that reliance on memory may interact with favoritism – a form of discrimination that remains widespread, even though it carries harmful economic consequences and raises ethical concerns. This paper investigates whether and how reliance on memory amplifies favoritism in personnel decisions. In an online experiment, managers make incentivized hiring decisions regarding workers of varying quality. Two experimental manipulations isolate the causal effect of memory on favoritism. First, I experimentally vary managers' preferences for workers. Second, I vary whether managers base their choices on workers' actual quality or on their recall of that quality. I find that when decisions rely on memory, the gap in the hiring probability between favored and non-favored workers more than doubles from 10 to 21 percentage points. I find that motivated memories mediate around one third of this effect: managers remember favored workers as being of higher quality than unfavored ones. The remainder of the effect arises because noise in memory leads managers to rely more on group identity when making decisions.

JEL-codes: C90, D83, D90, D91, J71
Keywords: beliefs, discrimination, favoritism, hiring decisions, memory, motivated cognition, personnel decisions

*__Contact:__ Charlotte Cordes, LMU Munich, charlotte.cordes@econ.lmu.de.
Acknowledgements: I am thankful to Raphael Brade, Valeria Burdea, Florian Englmaier, Tilman Fries, Svenja Friess, Katrin Gödker, Pedro Gonzalez-Fernandez, Yutaka Makabe, Marcel Quindt, Charlotte Saucet, Klaus Schmidt, Simeon Schudy, Lisa Spantig, Andrej Woerner, and audiences at LMU Munich, Bocconi University, RWTH Aachen, Ulm University, Maastricht University, the Women in Economics Brown Bag Seminar, the 27th Bavarian Micro Day in Regensburg, the 2025 ESA European meeting in Brno, the 1st Berlin Micro Theory and Behavioral Economics PhD Conference, M-BEES 2025 in Maastricht, IMEBESS 2025 in Valencia, the semi-annual CRC TRR 190 meetings, the Workshop on Narratives, Beliefs and Memory in Riederau 2024 and 2025, and the Summer School Cognitive Foundations of Decision Making in Rabat 2024. Moreover, I thank Svenja Friess, who kindly provided me with her knowledge-quiz material.
Funding: Support by the German Research Foundation through CRC TRR 190 "Rationality and Competition" (Project A04) is gratefully acknowledged.
Ethics approval: The experiment obtained ethics approval from LMU Munich (Project 2024-12).
Pre-registration: The experiment was pre-registered at `https://aspredicted.org/yrhb-s9kt.pdf`.

1 Introduction

Many important economic decisions rely on recalled information. For example, teachers base their evaluation of students' performance on their memory (Miserocchi, 2023), and judges rely on memory when deciding on the sentences of defendants.[1] Even in formalized hiring and promotion decisions, perceived performance is often subjective and dependent on decision makers' memory. In such decisions, memory imperfections can lead to the adoption of heuristics based on stereotypes (Miserocchi, 2023), associations (Radbruch and Schiprowski, 2025) and categorizations (Belot and Schröder, 2023). Memories are not only imperfect but also shaped by individuals' motives (see Amelio and Zimmermann, 2023, for a review). In hiring and promotion decisions, decision makers may therefore hold biased memories in favor of preferred individuals. This suggests that reliance on memory may amplify favoritism, defined as the *"unfair support shown to one person or group"*.[2]

This project provides causal evidence for the role of memory as an amplifier of favoritism by addressing two key questions: First, do people distort their memories in favor of their preferred group? Second, does reliance on memory amplify favoritism? Answering these questions, I find evidence for both motivated memory and 'memory-induced favoritism'. Thus, I identify reliance on memory as one important, novel driver of persistent favoritism in organizations. From a policy perspective, these findings underscore that efforts to promote fairness in hiring and evaluation decisions should not only target attitudinal changes but also the design of the decision environments, for instance, through systematic record-keeping and more frequent evaluations.

Personnel decisions provide a natural environment in which the proposed mechanism is particularly relevant. First, favoritism is prevalent in personnel decisions. In many modern workplaces, objective performance measures are noisy, unavailable, or difficult to retrieve. Therefore, subjective evaluations – often grounded in memory – play a central role in personnel decisions (see, e.g., Frederiksen et al., 2017). Such discretion can reinforce unequal treatment: for example, managers tend to reward those with whom they have stronger social ties (Breuer et al., 2013; Cullen and Perez-Truglia, 2023), and women are systematically rated lower in subjective assessments even when objective performance is comparable (Benson et al., 2024). Second, favoritism in hiring decisions distorts efficient talent allocation in organizations, resulting in detrimental performance (Bandiera et al., 2009; Riaño, 2023).[3] Third, favoritism in personnel decisions undermines the principle of merit-based decision-making in the work-

[1] In many European countries laws prohibit recording most criminal cases, see, for example, the case of Germany (§169 GVG), where even written protocols only include specific statements by request (§160 ZPO).

[2] Note that favoritism is a specific form of discrimination, defined as *"treating a person or particular group of people differently"*. Favoritism is not the same as positive discrimination, *"the act of giving advantage to those groups in society that are often treated unfairly because of their race, sex, etc."*, as positive discrimination is aimed at establishing equality and not based on personal preferences (all definitions come from the Cambridge Dictionary: `https://dictionary.cambridge.org/`).

[3] Hiring based on social networks can also have beneficial effects to the firm by lowering recruitment cost and turnover (Burks et al., 2015; Friebel et al., 2023).

place. Experiencing favoritism reduces labor supply (Gagnon et al., 2025), lowers workers' morale (Ruebeck, 2025), and triggers antisocial behavior (Dickinson et al., 2018). In contrast, meritocratic promotion systems have been shown to increase workers' effort (Deserranno et al., 2025).

Studying the link between memory and favoritism in naturally occurring settings is challenging as any relationship between memory reliance and favoritism in hiring outcomes is likely endogenous. For instance, managers with strong preferences in favor of a specific group may invest less effort in double-checking records and, therefore, rely more heavily on memory. Moreover, data on decision makers' memories are typically unavailable. To overcome challenges of causal identification and data limitations, I design an online experiment. The experiment mimics a real-world labor market setting in which individuals with managerial experience make incentivized decisions. The experimental approach enables me to precisely control the information environment and cleanly elicit memories. It also allows me to causally identify the effect of memory reliance on favoritism through exogenous variation in decision makers' memory and preferences.

In the experiment, participants with managerial experience decide whether to hire workers to perform a task. Each manager evaluates the profiles of eight workers. Every profile includes a picture and a name that signal the worker's group identity (gender), and an objective quality attribute (quiz performance). Managers are monetarily incentivized to hire workers of high quality. To allow imperfect memory to influence decisions, the experiment comprises a screening stage and a hiring stage, separated by a distraction task. In the screening stage, managers learn about their incentives and see the workers' profiles for a first impression. In the hiring stage, managers decide separately whether to hire each given worker.

The experiment features two sources of exogenous variation. The first induces preferences in favor of a specific group of workers. Managers receive a monetary bonus if they hire a worker from a specific group. As such, the experimental setting models the trade-off between favoritism and merit-based hires. I vary between participants whether male or female workers constitute a manager's 'preferred group'. Thus, each worker can belong either to the PREF or NOPREF group. That is, I exogenously induce taste-based favoritism, i.e., a preference for workers that is independent from worker quality. This treatment variation is orthogonal to managers' pre-existing gender preferences. Second, to identify the causal effect of memory on favoritism, I manipulate reliance on memory in a between-subject design. I do so by varying whether managers make their hiring decisions based on recalled quality attributes (RECALL) or after receiving a reminder of the true quality attributes (REMINDER).

To formalize the trade-off between favoritism and merit-based hires and build intuition about my hypotheses, I conceptualize the hiring decision in a simple theoretical framework. The manager decides whether or not to hire a worker based on two attributes: the worker's quality and group identity. These two attributes are independent. The manager's material utility is affected

by both attributes, and she favors workers from one of the groups. In this setup, engaging in favoritism is defined as conditioning the decision on group identity. Managers enjoy hiring favored workers but are also self-image concerned and therefore incur psychological costs when engaging in favoritism. Faced with a trade-off between gaining utility through favoritism and reducing psychological costs by avoiding it, managers will refrain from favoritism when those psychological costs are sufficiently high. Following the experimental setup and based on the idea that more salient attributes are recalled with a higher likelihood than less salient attributes, I assume that when managers decide based on their memories, they always remember the group identity but may forget the quality attribute. The noisy memories of the quality attribute may be biased. Additionally, the absence of tangible information may reduce psychological costs of engaging in favoritism. As a result, reliance on memory can lead to increased favoritism. Accordingly, I formulate and test two main hypotheses. First, managers remember the quality attributes of preferred workers more favorably. Second, reliance on memory amplifies favoritism in hiring decisions.

The experimental results support both hypotheses. To test the first hypothesis about motivated memory formation, I analyze the incentivized recall of the quality attributes in the RECALL treatment. I find that managers recall the quality attributes of PREF-workers significantly more favorably than those of NOPREF-workers. They recall PREF-workers of having a 12% standard deviation higher quiz score than NOPREF workers ($p = 0.002$). These empirical results provide clear evidence for motivated memory formation.

Next, I investigate how reliance on memory affects favoritism by comparing hiring decisions between the experimental treatments RECALL and REMINDER. In REMINDER, managers hire 40.6% of the NOPREF-workers and 50.4% of the PREF-workers. Thus, the baseline hiring gap in REMINDER, driven by the incentives to hire PREF-workers, is 9.8 pp ($p < 0.001$). In line with the second hypothesis – reliance on memory amplifies favoritism –, the gap in the hiring probability between PREF and NOPREF workers more than doubles to 20.6 pp when decisions rely on memory in RECALL, driven by a significant increase in the hiring probability of PREF-workers ($p < 0.001$). Thus, I document strong evidence for 'memory-induced favoritism'.

The experimental design also allows me to quantify the relative importance of two mechanisms in driving the increase in the hiring gap from REMINDER to RECALL: one operates through motivated memory distortions, and the other stems from reliance on memory itself. The first mechanism builds on the idea that even when a manager applies the same quality standards when deciding based on memories as when quality signals are directly observable, biased memories can lead them to recall preferred workers' quality as better, increasing observed favoritism. This mechanism accounts for 34% of the total treatment effect. The remaining 64% reflect a shift in how evaluators weigh group identity relative to quality. This is the second mechanism. It is based on the idea that the noise in memory creates a 'moral wriggle room', allowing managers to rely on the group identity without feeling that they are violating quality standards.

This project contributes to three different strands of the literature. First, I contribute to the active research on discrimination. Specifically, the paper adds to the literature that explores why inaccurate belief-based discrimination persists even when information should correct such beliefs (Bohren et al., 2025). The literature has shown that rational attention allocation (Bartoš et al., 2016; Cavounidis et al., 2024) and information acquisition (Coffman et al., 2025) can lead to such inaccurate beliefs.[4] At the information processing stage, non-Bayesian updating patterns – such as conservatism (Campos-Mercade and Mengel, 2024), contrast effects (Esponda et al., 2023) or asymmetric updating based on stereotypes (Sarsons, 2022; Rackstraw, 2025) – can contribute to the persistence of inaccurate beliefs. Most closely related to my project, Eyting (2022) provides experimental evidence that people discriminate based on motivated beliefs formed through selective information acquisition and processing. In a similar vein, people tend to avoid altruistic actions by distorting their beliefs about others' altruism (Di Tella et al., 2015; Grunewald et al., 2023). In the context of discrimination, Stötzer and Zimmermann (2024) show that the opportunity to benefit financially at the expense of an out-group leads people to justify their selfish behavior by forming discrediting beliefs about that out-group. I provide the first evidence that links motivated memory formation to 'motivated discrimination'.

Beyond motivated memory distortions, I also provide evidence for an increase in favoritism due to reliance on memory per se. 'Memory-induced favoritism' is therefore a form of 'subtle discrimination' – biased acts that cannot be objectively ascertained as discriminatory (Pikulina and Ferreira, 2025). This relates to the idea of 'implicit discrimination' (Cunningham and De Quidt, 2022; Barron et al., 2025), which describes when decision makers do not engage in discrimination when this would reveal their bias, but do discriminate when their discriminatory motive is obscured. Barron et al. (2025) empirically identify such 'implicit discriminators' who only discriminate against women in binary choices when female candidates have different qualifications than their male competitor. The presence of the direct effect suggests noisy memory as a complementing source of 'implicit discrimination'.

Second, this project links to the literature on motivated memory more broadly. Empirical evidence has documented motivated memory across various domains, including prosocial behavior (Li, 2013; Saucet and Villeval, 2019; Carlson et al., 2020; Galeotti et al., 2020), financial decisions (Gödker et al., 2025), own performance (Chew et al., 2020; Zimmermann, 2020; Huffman et al., 2022; Roy-Chowdhury, 2024; Hagenbach et al., 2025), work organization (Cordes et al., 2025), life decisions (Müller, 2022), or gym attendance (Sial et al., 2023). I expand this literature by studying hiring decisions as a novel domain in which motivated memory arises. My project differs from previous work on motivated memory in the moral domain (Li, 2013; Saucet and Villeval, 2019; Carlson et al., 2020; Galeotti et al., 2020) as I focus on memory as a tool to establish a 'moral wiggle room' (Dana et al., 2007) before making a choice rather than ex-post

[4] Also institutional and structural factors (Bohren et al., 2025), network formation (Evsyukova et al., 2025), and organizational design (Badalyan et al., 2023; Hirmas and Hausfeld, 2024) can lead to discrimination through distorted and selective available information.

memory distortion of past choices.[5] The anticipated hiring decisions provide an instrumental motive for memory distortions, which may be a necessary condition for the formation of motivated memory (Schwardmann and Van der Weele, 2019; Galeotti et al., 2020; Schwardmann et al., 2022). In addition, by eliciting incentivized choices, I provide direct evidence of how such motivated memories influence behavior, which remains rare in this literature.[6]

Finally, I add to the nascent literature on the interaction between memory and personnel decisions. So far, this literature has focused on specific cognitive memory patterns. These memory patterns include associations (Radbruch and Schiprowski, 2025) and categorizations (Belot and Schröder, 2023). Bordalo et al. (2016) develop a model in which stereotypes arise because individuals overweight characteristics that are most representative, i.e., most diagnostic relative to a comparison group, leading to context-dependent and stereotyped assessments of groups. Most closely linked to my project, Miserocchi (2023) identifies stereotype-reinforcing memories as a driver for gender discrimination in school track recommendations. Thus, associations and stereotypes influence memory and thereby assessments of other people. My project assesses the importance of memory biases that are not directly tied to cognitive factors but are instead based on affective factors. The results highlight that when managers favor one group over another, reliance on memory will increase favoritism with consequences for talent allocation in organizations and equality of opportunity.

The rest of the paper is structured as follows. Section 2 covers the experimental design, as well as information about the implementation. Section 3 derives the main hypotheses using a simple theoretical framework. Section 4 presents the main results on motivated memory and favoritism. Section 5 contains additional results including insights on metacognition, welfare, and confidence as well as heterogeneity analyses and an overview of conducted robustness checks. Finally, Section 6 concludes.

2 Experimental design

In this section, I first provide an overview of the experimental design. Then, I explain the main design features in more detail, followed by information on the implementation. I provide instructions in Appendix D.

[5] In particular, I study memories of choice-underlying attributes. Previous studies have instead considered (signals about) own performance or investments (Zimmermann, 2020; Huffman et al., 2022; Roy-Chowdhury, 2024; Gödker et al., 2025), noise (Hagenbach et al., 2025), past choices (Li, 2013; Saucet and Villeval, 2019; Carlson et al., 2020; Galeotti et al., 2020; Sial et al., 2023), and past preferences (Müller, 2022) as items to be recalled.

[6] Exceptions include re-investment decisions in Gödker et al. (2025) and work organization choices in Cordes et al. (2025).

2.1 Overview

I conduct an artefactual field experiment with managers on Prolific.[7] The experiment features a labor market with two groups of workers, as well as a multi-period setup that enables retrieval from a memory database. The labor market mimics a real-world hiring scenario in which participants act as managers who hire male and female workers to perform a task. For their decisions, managers have access to a worker's picture, name, and an objective quality attribute. Managers face incentives to hire well-performing workers. To allow imperfect memory to influence decisions, the experiment comprises two stages, separated by a distraction task. In the first stage, the screening stage, managers learn about their incentives and build their memory database by seeing each of eight workers for a first impression. The early announcement of the incentives allows me to capture distortions arising throughout the whole memory process. Then, after engaging in a distraction task, managers proceed to the second stage, the hiring stage, where they face a series of personnel decisions. The experiment ends with a questionnaire. The structure of the experiment is illustrated in Figure 1.

To causally identify the role of memory for favoritism, I exogenously vary both reliance on memory and group specific preferences for workers. First, in a between-subject design, managers make their decisions either based on a reminder of the actual quality signal (REMINDER) or their memory of that signal (RECALL). Second, I induce favoritism for a specific group of workers. That is, in addition to incentives to hire well-performing workers, managers receive a bonus if they hire a worker from a specific group. Conceptually, I thereby create taste-based discrimination. The favored group (PREF) either consists of female or male workers, while the non-incentivized group (NOPREF) consists of the respective other group of workers. The treatment variation is orthogonal to any additional, existing preferences for a specific gender. In contrast to naturally occurring preferences, induced preferences for specific worker groups cannot be correlated with memory quality nor the tendency to distort memories. Consequently, I can causally assess how memories about individuals from favored groups differ from those about non-favored groups, without taking stances on the source of discriminatory preferences.

2.2 Design details

Worker profiles During the experiment, managers hire workers based on short profiles. I collected data on workers' performance in a quiz prior to the main data collection.[8] For each worker, I used an AI tool to generate pictures that match the real worker's physical

[7] The experimental sample consists of Prolific participants who report to *"have experience being in a management position"*. This distinguishes the experiment from a conventional online experiment according to the taxonomy by Harrison and List (2004).

[8] For a more detailed description of this additional worker experiment, see Appendix C.1.

characteristics.[9] To balance workers' quality across groups, I then selected a final pool of eight workers, four of each gender.[10] For each gender, one of these workers achieved a quiz score of 3, 4, 6, and 7, respectively. Thus, the quality distribution is the same across both groups. Any differences that may nevertheless exist between males and females (for example, due to different memorability of the pictures) are orthogonal to the treatment variation. Finally, I created a short profile for each worker in the final worker pool, containing i) the AI-generated picture, ii) a name (chosen from a list of common names and surnames), and iii) the quiz score of the respective worker. Figure 1 contains an example of such a profile. The name and picture serve as a device for recognition and a signal for gender identity. The quiz score serves as a quality attribute and ranges from 0 to 10. The managers know about the format of the quiz (10 multiple choice questions with limited time to answer), but not the content to avoid stereotypical beliefs.

Incentives The structure of the bonus payments is designed such that managers have an incentive to hire well-scoring workers. They also induce preferences in favor of one gender, which experimentally creates taste-based favoritism. If a manager decides to hire a worker, she receives a performance bonus of 2 GBP if the project is successful. The project is successful with probability 10% times the quiz score. If the manager decides not to hire a worker, she receives the 2 GBP bonus with 50% probability. Additionally, managers receive a bonus payment of 0.50 GBP if they hire a worker from PREF. This favoritism bonus is received irrespective of a worker's quiz score and whether the project is successful.[11] Managers decide on each of the workers separately. None of the hiring decisions mechanically affect subsequent hiring decisions. Thus, the hiring task consists of several simple, nonstrategic decisions. One of the eight decisions is then randomly selected to be payout-relevant. For this decision, the performance bonus of 2 GBP and the favoritism bonus of 0.50 GBP are paid out to the manager (if applicable).[12]

Screening stage After reading the instructions and providing the correct answers to three comprehension questions, managers proceed to the screening stage where they are presented

[9] In particular, the prompts I used contained the workers' gender, eye color, skin type, hair color, hair style and approximate age. Appendix C.2 contains the exact wording of these prompts and more information on the procedure.

[10] Appendix C.3 contains a detailed description of the selection procedure. Note that the selection procedure involved a rating of the pictures along several dimensions and excluded pictures that were rated very high or low.

[11] If managers were maximizing expected monetary payoff, this incentive structure could result in the following strategy: Hire all PREF-workers and hire NOPREF-workers if and only if their quiz score is above 5. By framing the screening as a first impression, keeping some ambiguity about future tasks, and letting managers engage with the information on worker profiles, I nudge managers to pay attention to all workers. Still, the incentive scheme may give rise to the concern that incentives induce managers to pay less attention to the quiz score of PREF-workers. However, two pieces of evidence render this concern highly unlikely. First, the average fraction of correct recalls does not significantly differ across the two worker groups (see Section 4.1) and second, the time spent on the screening pages does not differ across worker groups (see Section 5.3).

[12] Note that the hiring decision did not impact workers, and this was salient to managers. This feature is common in similar experiments (e.g. Bohren et al., 2025; Barron et al., 2025) and rules out social preferences. Therefore, the decision (not) to engage in favoritism can be interpreted as a result of the trade-off between incentives and self-image concerns.

with the profiles of all eight workers, one by one, in a random order. For each of the profiles, managers must insert simple information about the workers' name and quiz score in order to proceed. Thereby, I prevent managers from mindlessly clicking through the profiles. The screening stage is announced as a way to get a "first impression" of the workers. Participants know that there will be a hiring decision, but they do not know when it will occur or which other decisions may follow. This ensures that participants in RECALL and REMINDER experience the screening stage in exactly the same way. It also means that managers are not primed to take notes.

Distraction stage In the distraction stage, managers engage in a real-effort task for eight minutes. The real-effort task consists of counting zeros in matrices containing zeros and ones. Managers receive a piece-rate incentive for this task. They can only proceed to the next matrix when they have correctly solved the previous one. After eight minutes, managers are automatically advanced to the hiring stage.

Hiring stage The hiring stage elicits the main outcome variables: the recall of the quiz scores and the hiring decisions. After being reminded of the incentives, Managers assess each worker one by one, in a random order. For each worker, managers decide whether or not to hire this worker. Afterwards, they provide a recommendation that serves as a complementary continuous outcome variable. The recommendation is the answer to the question *"How strongly do you recommend this worker?"* on a scale from 0 to 100. The recommendation decision is not directly incentivized but still consequential. Managers are informed that participants of an auxiliary experiment, run shortly after the main experiment, would be shown the recommendations to base their hiring decision on.[13]

Managers in both treatment groups make the same decisions, the only difference is whether they are reminded of the quiz score (RECALL) or must rely on their memories (REMINDER). In RECALL, managers see each profile, however, they cannot see the quiz score. Instead, they see question marks (see also Figure 1) and have to recall the quiz score. Managers receive incentives for precise recall. For the randomly selected worker that is relevant for the payoff, they receive 0.50 GBP if their recall is correct and 0.25 GBP if their recall is within 1 of the correct number. Additionally, I elicit people's subjective memory errors to be able to measure managers' (over)confidence related to their memory quality.[14] Managers in REMINDER also make the hiring and recommendation decision but based on the complete profiles that include the quiz scores. To equalize the amount of engagement with the profiles between the treatments, managers in REMINDER engage in an incentivized control task, the repetition task, that involves re-entering the worker's quiz score. As stating the recalled quiz score may have an influence on

[13] Appendix C.4 describes this auxiliary experiment in more detail.

[14] In particular, I use the subjective error method proposed by Bosch-Rosa et al. (2024) and ask: *"How far away do you think your recall is from the true quiz score?"*.

the hiring decision per se and vice versa, I randomize whether managers first make the hiring and recommendation decision or whether they first face the recall or repetition task, respectively.

In both treatments, after all the hiring decisions have been made, managers have to rank the workers from 1 to 8. This decision comes as a surprise and is incentivized via a lottery to ensure managers reveal their true preferences in the ranking. The lottery picks one worker. The higher a worker is ranked, the higher the probability that this worker is picked in the lottery. For the picked worker, the manager receives a bonus of 2 GBP with probability 10% times the quiz score that this worker achieved. These incentives are only paid out for one manager per day.

Final questionnaire The experiment ends with a final questionnaire. At the beginning of the questionnaire, managers provide their perceptions and strategies in the experiment. I use this information in exploratory analyses to derive insights into the mechanisms that are at play. The first set of questions asks managers to provide insights into their hiring strategy by i) stating how important the (memories of the) quiz score as well as the worker's gender were for the hiring decision, and ii) describing their hiring strategy in a free text form. A second set of questions aims at gaining meta-cognitive insights and asks iii) how many male and female workers they recalled having hired, and iv) (only in RECALL) how many quiz scores they recalled correctly. A third set of questions only applies to RECALL and concerns memorization strategies. In particular, managers had to state v) whether they have used memory-enhancing tools (and if so, which), vi) how difficult they found the recall task, as well as vii) which strategies they used if they were not sure they remembered the quiz score correctly.

Afterwards, managers answer survey questions that I exploit in heterogeneity analyses. I chose questionnaires that capture personal characteristics and attitudes that may be associated with engaging in favoritism and the psychological costs associated to 'doing something bad'.[15] The questionnaire also elicits some basic demographics (age, gender, and whether the participant visited university) as well as risk and time preferences.[16] These are used as control variables. Finally, managers have to provide their guess regarding the purpose of the study.

[15]First, I use the experimentally validated survey module by Schudy et al. (2024) that measures preferences for being and being seen as honest. Then, I elicit agreement on a five-point Likert scale to two items related to being and being seen as a discriminator (*"I am not a discriminator"* and *"It is important to me not to be seen as discriminatory by others"*) as well as one item related to equality (*"Every person should be treated equally"*) and one to meritocracy (*"Workers should be rewarded based on their performance"*). The questionnaire further contains a survey instrument by Aquino and Reed II (2002) that aims to capture the importance of moral identity and the GP5 scale by Cohen et al. (2014) that captures the proneness to feelings of guilt. Additionally, as a proxy for home-grown gender attitudes, I measure the degree of sexist attitudes using the ambivalent sexism inventory designed for men (Glick and Fiske, 1996) and the corresponding version for females (Glick and Fiske, 1999). These questionnaires aim at discovering two distinct aspects of attitudes towards the other gender: benevolent and hostile sexism.

[16]I elicit risk and time preferences according to short survey questions from the Global Preference Survey (Falk et al., 2023). Risk preferences are measured as an answer between 0 and 10 to the question *"How willing are you to take risks, in general?"* and patience is measured as an answer between 0 and 10 to the question *"How willing are you to give up something that is beneficial for you today in order to benefit from that in the future?"*.

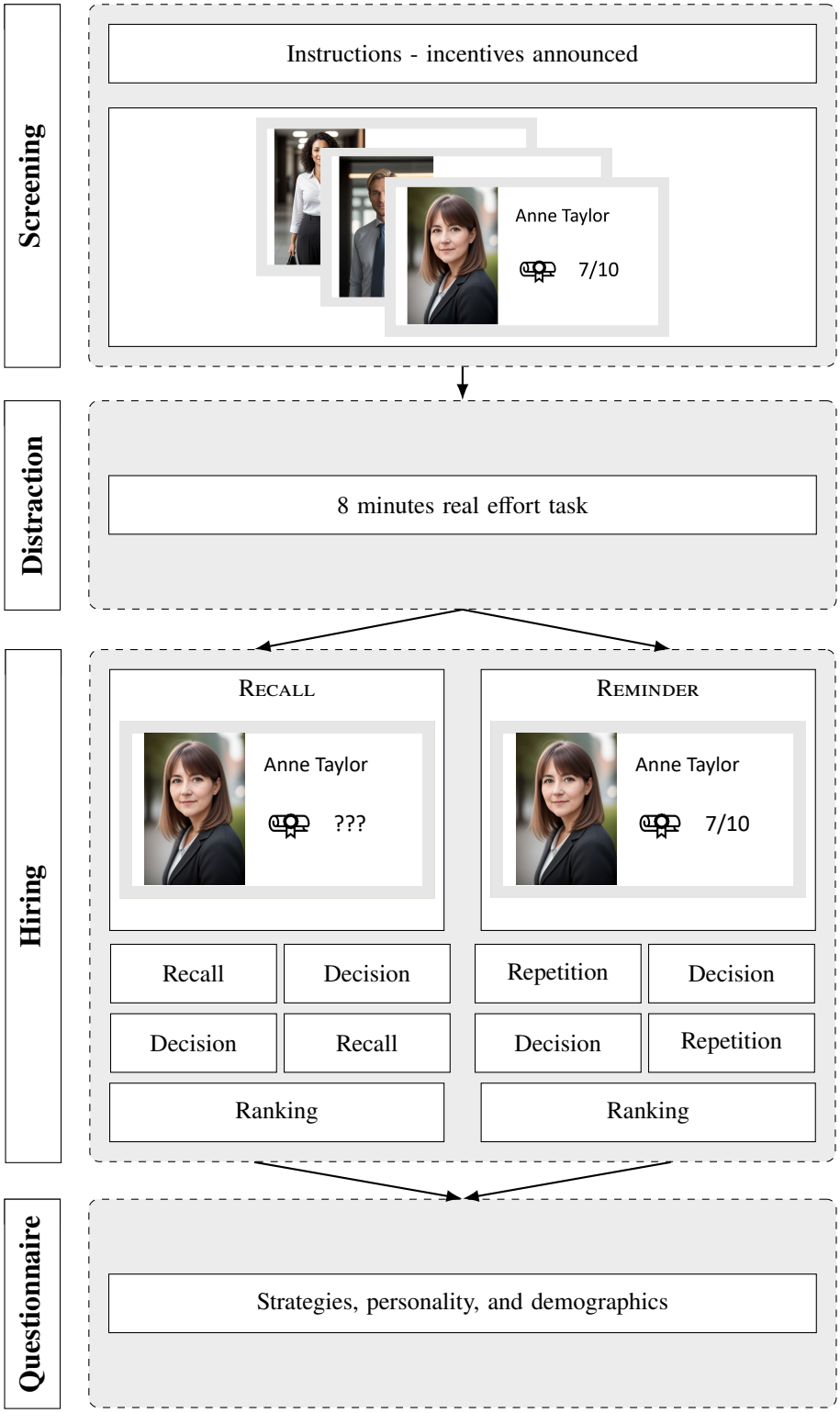


Figure 1: Overview of the experimental design

2.3 Procedures

The survey was administered with Qualtrics. I obtained IRB approval from the ethics committee at LMU Munich (Project 2024-12) and the project has been preregistered at: `https://aspredicted.org/yrhb-s9kt.pdf`. Any deviations from the preregistrations are documented in Appendix B.

To add external validity, I ran the experiment with a gender-balanced Prolific sample consisting of white-collar workers. In particular, I restricted the sample of Prolific participants to residents of the UK and Ireland who i) are employees who report managerial experience[17], ii) have at least a high school diploma or a vocational training diploma, iii) are between 25 and 65 years old, and iv) have an approval rate (share of studies by the participants that were accepted by the respective researchers) of at least 95%.

I collected data in November 2024. Median participation time was 31:42 minutes. The aim was to recruit 800 participants. After excluding observations based on preregistered criteria (participants were excluded if they do not pass one of two attention checks or if they state to have used memory aids), the sample size consists of 756 participants.[18] Randomization into treatments was determined by random variables within the system. I provide a balance table in the Appendix (Table A.21).

3 Conceptual framework and hypotheses

Setting I develop a simple theoretical framework to structure the manager's decision process and motivate the hypotheses. The decision maker (i.e., the manager) faces the choice whether to hire a worker ($d = 1$) or not ($d = 0$). The manager has two pieces of evidence about the worker: The (objective) quiz score q and the group identity g. The worker can belong to the manager's preferred group ($g = 1$) or not ($g = 0$). Quiz scores in the population of workers are distributed according to a distribution F with support on $Q \subseteq [0, 1]$. This distribution is known to the manager. The attributes q and g are independent. In the experiment, the attribute q corresponds to the quiz score divided by 10. PREF-workers correspond to $g = 1$, while NOPREF-workers correspond to $g = 0$. The distribution of quiz scores is a uniform distribution on $\{3, 4, 6, 7\}$, so $Q = \{0.3, 0.4, 0.6, 0.7\}$. By design, the quiz score q and the group identity g are independent.

When a manager decides to hire a worker, this worker works on a project. The probability that this project is successful depends on the worker's quality q. The manager's utility from hiring a worker depends on both the project's success and the group identity g. Following the incentive structure from the experiment, the manager's expected payoff from hiring the worker

[17] The reported managerial experience is part of the platform's pre-screeners and is determined according to the answer to the question: *"Do you have any experience being in a management position?"*

[18] The main results are robust to including participants who stated to have used memory aids or applying stricter exclusion criteria (see Section A.5).

is $b * q + \bar{\alpha} * g$, and the expected payoff of not hiring is $b * o$. Here, b is the level of the bonus ($= 2$ in the experiment), and o the success probability of the outside option ($= 0.5$ in the experiment). The parameter $\bar{\alpha} > 0$ reflects the benefits of engaging in favoritism. It is the weight a payoff-maximizing manager places on the group identity ($= 0.5$ in the experiment). Thus, for NOPREF-workers, hiring a worker is payoff-maximizing if and only if the quiz score is above 5, while for PREF-workers, hiring a worker with a quiz-score above 2.5 is payoff-maximizing. Generally, since $\bar{\alpha} > 0$, a payoff-maximizing agent conditions the decision on the group identity and thereby engages in favoritism.

The manager takes the decision in an environment $e \in \{0, 1\}$ that corresponds to either of the two experimental treatments, RECALL ($e = 1$) and REMINDER ($e = 0$). I assume that in a world where the manager has to rely on memory ($e = 1$), she is always aware of a worker's group identity g, but has to base the decision on memories of the worker's quality q, that is, $\tilde{q}$, a noisy memory of q. The memory $\tilde{q}$ depends on the memory quality $\lambda \in [0, 1]$ but not the group identity q. For illustrative purposes, I assume the following parametrization:

$$\tilde{q} = \begin{cases} q & \text{with probability } \lambda \\ \sim F & \text{with probability } (1 - \lambda). \end{cases} \tag{1}$$

Due to the noise, the managers' overall utility will be lower when relying on memories. She may hire some low-quality workers whom she misremembers as being of high quality, and vice versa. However, the fraction of hired workers in each group will be the same in environment $e = 1$ as under $e = 0$.

A behavioral manager I now consider an extension of the basic setting where I allow for a psychological component in managers' utility functions. Managers may not maximize expected payoff. Instead, they may want to "do the right thing" and not engage in favoritism. Specifically, the manager trades off costs and benefits from engaging in favoritism. In my experiment, the utility gain from favoritism is a monetary payoff, while the costs are non-tangible.[19] I model this idea by replacing $\bar{\alpha}$ with a function $\alpha(\theta, e) : [0, 1] \times \{0, 1\} \to [0, \bar{\alpha}]$ that determines to what extent the utility is impacted by the group identity.

The function $\alpha(\theta, e)$ reflects how the manager's self-image type $\theta \in [0, 1]$ and the environment $e \in \{0, 1\}$ determines the degree of favoritism. The higher the managers type, the less weight the manager places on the group identity signal. That is, the function is decreasing in θ, $\frac{\delta\alpha(\theta,e)}{\delta\theta} < 0$, and it holds that $\alpha(0, 1) = \alpha(0, 0) = \bar{\alpha} > 0$. I also assume that $\alpha(\theta, 1) > \alpha(\theta, 0) \quad \forall\theta \in (0, 1]$. Intuitively, environment $e = 1$ involves an 'excuse' that makes it 'easier' to engage in favoritism. This can be micro-founded with a dual-self model that includes

[19] In different applications, the utility gains from favoritism could also reflect non-tangible benefits such as anticipatory utility from spending time with a worker from the preferred group. Conversely, costs of engaging in favoritism may be tangible, for example, when a company rule forbids favoritism and a manager faces consequences for not obeying.

self-image concerns (Bénabou and Tirole, 2006; Grossman and Van Der Weele, 2017). Not having access to perfect quality signals alleviates image concerns from using g as a future self can attribute the hire of a low quality worker from the preferred group to incorrect memories.

Motivated memory I now derive the first hypothesis which concerns the memories of the behavioral agent. Biasing the quality signal to align with the respective group identity allows the manager to favor workers from the preferred group without consciously using group identity as a decision criterion. Such motivated memories may be the result of strategic memory distortions in a dual-self-process (Bénabou and Tirole, 2002; Chew et al., 2020; Hagenbach and Koessler, 2022) or arise through a rather unconscious process of endogenous memory rehearsal (Conlon, 2024). Motivated memories imply that if $g = 1$, $\tilde{q}$ will be biased upwards and if $g = 0$, $\tilde{q}$ will be biased downwards. The extent of the distortions will depend both on the memory quality, and the degree of memory distortion. For illustrative purposes, I assume the following parametrization.

$$\tilde{q} = \begin{cases} q & \text{with probability } \lambda \\ \sim F & \text{with probability } (1-\lambda)(1-\mu) \\ g * \bar{q} + (1-g) * \underline{q} & \text{with probability } (1-\lambda)\mu. \end{cases} \tag{2}$$

Here, $\lambda \in [0,1]$ is the manager i's memory quality and $\mu \in [0,1]$ measures the manager's cognitive flexibility needed to distort beliefs, i.e., the proneness to form motivated beliefs. $\lambda = 1$ implies perfect recall. If $\lambda < 1$, memory is imperfect. With probability $(1-\lambda)(1-\mu)$, the blank memory will be filled by a random variable, distributed according to the prior F. With probability $(1-\lambda)\mu$, the quality signal of the worker from the preferred group will be misremembered as the highest value on the support of F, $\bar{q}$, and the quality signal of the worker from the non-preferred group will be misremembered as the lowest value of the distribution F, $\underline{q} < \bar{q}$. For example, in the experiment, with probability $(1-\lambda)\mu$ a manager remembers Pref-workers of having a quiz score of 7 and Nopref-workers of having a quiz score of 3. As a consequence, the expected value of the quality signal is:

$$\mathbb{E}_F[\tilde{q}] = \lambda * \underbrace{q}_{\text{correct recall}} + (1-\lambda)(1-\mu) * \underbrace{\mathbb{E}_F[q]}_{\text{noise}} + (1-\lambda)\mu * \underbrace{(g\bar{q} - (1-g)\underline{q})}_{\text{bias}}. \tag{3}$$

This equation implies that a manager with parameters $\lambda < 1$ and $\mu = 0$ will have noisy but unbiased memory – as a payoff-maximizing agent would have. Conversely, a manager with $\lambda < 1$ and $\mu > 0$ will on average recall the quality signal of a worker from the preferred group ($g = 1$) by $(1-\lambda)\mu(\bar{q} - \underline{q})$ higher compared to a worker from the non-preferred group ($g = 0$). Thus, there will be a difference between average recalled quality attributes of workers of both groups if and only if $\lambda < 1$ and $\mu > 0$. This difference, $(1-\lambda)\mu(\bar{q} - \underline{q})$, is increasing in the flexibility for memory distortions and decreasing in the memory quality.

People's memory is generally imperfect ($\lambda < 1$) and previous literature on motivated memory (see Amelio and Zimmermann, 2023, for a review) and motivated reasoning to justify discrimination (see Eyting, 2022) suggests the prevalence of types with $\mu > 0$. In line with this recent evidence, I assume the existence of types with $\lambda < 1$ and $\mu > 0$, giving rise to my core hypothesis regarding recalls:

Hypothesis 1. *On average, managers will recall quality attributes of* PREF*-workers to be higher than those of* NOPREF*-workers.*

Hiring decisions based on memory Finally, I derive the second hypothesis regarding the effect of memory reliance on favoritism in hiring decisions. Two core aspects differ when decisions are based on memories of quality attributes ($e = 1$) compared to decisions based on the deterministic attributes itself ($e = 0$). First, the manager replaces a worker's attribute q by the noisy cognitive representation $\tilde{q}$ (which by Hypothesis 1 tends to be higher for workers with $g = 1$ compared to workers with $g = 0$).[20] Second, the environment determines the weight the manager places on the group identity, $\alpha(\theta, 1) \geq \alpha(\theta, 0)$.

The expected utility of hiring a worker for a manager in situation $e = 0$ (REMINDER) is now:

$$EU(d = 1, e = 0) = b * q + \alpha(\theta, 0) * g. \tag{4}$$

Conversely, the expected utility of hiring a worker for a manager in situation $e = 1$ (RECALL) is now:

$$EU(d = 1, e = 1) = b * \tilde{q} + \alpha(\theta, 1) * g. \tag{5}$$

Combining these equations with Equation (3), gives:

$$\begin{aligned} &EU(d = 1, e = 1) - EU(d = 1, e = 0) \\ &\quad = b * \big[(\lambda - 1)q + (1 - \lambda)(1 - \mu)\mathbb{E}_F[q] + (1 - \lambda)\mu\big(g\bar{q} + (1 - g)\underline{q}\big)\big] \\ &\qquad + [\alpha(\theta, 1) - \alpha(\theta, 0)] * g. \end{aligned} \tag{6}$$

Here, the two channels that lead to an increase in favoritism in situation $e = 1$ compared to $e = 0$ become visible. First, through the presence of an 'excuse' in the RECALL treatment, the expected utility of hiring depends more strongly on the group identity attribute ($\alpha(\theta, 1) - \alpha(\theta, 0) > 0$). The second effect operates through biased memory, as, in expectation, $\tilde{q} > q$ for workers from the preferred group and $\tilde{q} < q$ for workers from the non-preferred group.[21] In particular, the

[20] The implicit assumption behind this is that managers are ex-post naive and do not correctly trace back the true attribute based on introspection about the distortion in the recall. This assumption is based on recent evidence on ex-post naivete of memory distortions Conlon (2024). However, assuming only partial naivete would not change the overall conclusions.

[21] Note that these comparisons do not necessarily hold for every worker, as memory is also noisy. The direct effect, however, distorts the utility related to hiring upwards for every worker of the preferred group.

manager profits by $b(1-\lambda)\mu\ (\bar{q}-\underline{q})$ more from using the group identity attribute for hiring.[22] Together, these considerations lead to the following hypothesis:

Hypothesis 2. *Favoritism (i.e., the gap in the hiring probability between* Pref- *and* Nopref-*workers) is higher in the* Recall *treatment than in the* Reminder *treatment.*

The two channels additively impact the expected utility from hiring and thereby the hiring decision. Exploratory analyses will disentangle the relative importance of the presence of each channel.

The direct effect is closely related to the concept of 'implicit discrimination' (Barron et al., 2025; Cunningham and De Quidt, 2022). Barron et al. (2025) show that there are individuals who appear unbiased in transparent environments but begin to discriminate once ambiguity provides a form of 'moral wiggle room.' In such settings, decision makers can rationalize their actions as merit-based, even when these actions involve favoritism. Reliance on memory generates a similar opportunity for rationalization. However, whereas the ambiguity in Barron et al. (2025) and Cunningham and De Quidt (2022) arises from contextual uncertainty (in particular, when multiple performance dimensions make it difficult to verify whether a choice reflects favoritism), the ambiguity in my experiment stems from cognitive uncertainty (i.e., the noise in memories).

There are alternative underlying causes that could in theory contribute to the direct effect. One such cause is ambiguity aversion. An ambiguity-averse manager might decide to hire a preferred worker in Recall, because the ambiguity about the success probability makes the sure utility gain from hiring workers of group $g = 1$ relatively more attractive. However, in a choice regarding a worker of group $g = 0$, ambiguity in the success probability of the project makes the outside option relatively more attractive. Therefore, ambiguity aversion would also predict that managers in Recall would hire workers of group $g = 0$ less often compared to Reminder. I can directly test for this pattern in the data to rule out ambiguity aversion as an alternative to the 'excuse' in driving the direct effect.

Another alternative cause that may contribute to the direct effect could be the belief in a positive correlation between quality and group identity: If the group $g = 1$ had on average a higher quality than the group $g = 0$, filling the blanks with memories following Equation (2) would be rational. This reasoning extends to the situation where a manager mistakenly believes one group to be of higher quality. In my experiment, I rule out such beliefs by having identical performance distributions across worker groups and showing all workers to the managers. However, in settings with group differences where (inaccurate) statistical discrimination or stereotypes are present, this mechanism could amplify the effect of memory reliance on favoritism. Thus, my estimates of memory-induced favoritism could be seen as a lower bound.

[22] Note that due to the abstraction from group differences, pure noise in memory does not affect favoritism through this channel. Instead, the noise-term $(1-\lambda)(1-\mu)\mathbb{E}_F[q]$ implies a form of 'behavioral attenuation' (Enke et al., 2024), that is, choice elasticity with respect to attribute q is reduced in Recall compared to Reminder.

4 Results

In this section, I will test Hypothesis 1 (motivated memory), using the data from the RECALL treatment only. Then, I will compare the hiring decisions in the RECALL and REMINDER treatment to test Hypothesis 2 ('memory-induced favoritism'). Finally, I will provide evidence on the mechanisms.

4.1 Memory

The data exhibits substantial memory imperfections in recalled quiz scores, which is a necessary condition for motivated memory distortions. Overall, only 29.3% of recalled quiz scores are correct. Among incorrect recalls, the median deviation from the true quiz score is 2. In line with these results, 90% of managers perceived the memory task difficult or very difficult. However, on average, managers' memory appears well calibrated. The average remembered quiz score is exactly 5. Also, 94.7% of recalls are within 3 and 7.[23]

To test whether managers recall workers from their preferred group more favorably (Hypothesis 1), I compare the recalled quiz score for workers from PREF to workers from NOPREF within the RECALL treatment. Figure 2 shows the distribution of the recalled quiz scores separately for PREF (orange) and NOPREF (blue). The results reveal that the distribution of recalled quiz scores is shifted to the right for the preferred group.

For a formal analysis, I estimate the following model using OLS.

$$\mathtt{y}_{ij} = \beta_0 + \beta_1 \mathtt{pref}_{ij} + \gamma' X_i + \mu_j + \varepsilon_{ij}, \tag{7}$$

where $\mathtt{y}_{ij}$ is the recalled quiz score of manager i for worker j, and $\mathtt{pref}_{ij}$ is an indicator that is one when worker j belongs to manager i's preferred group. X_i is a vector of manager characteristics (gender, age, whether the participant visited university, risk preferences, and patience) and μ_j are worker fixed effects. Standard errors are clustered on the manager level. Throughout the main analyses, I will follow my pre-registration and control for worker fixed effects and manager characteristics in some specifications.

Column (1) of Table 1 shows the OLS regression results using Specification (7). The estimated coefficient of 0.199 represents an increase of the recalled quiz score by 12.3% of a standard deviation. This treatment effect is statistically significant ($p = 0.002$) and robust to including worker fixed effects and manager characteristics in Column (2). Summarized, these results confirm Hypothesis 1.

Result 1. *On average, managers recall quiz scores of PREF-workers to be higher than quiz scores of NOPREF-workers.*

[23] In Appendix A.5, I provide additional analysis showing that recall outperforms pure noise.

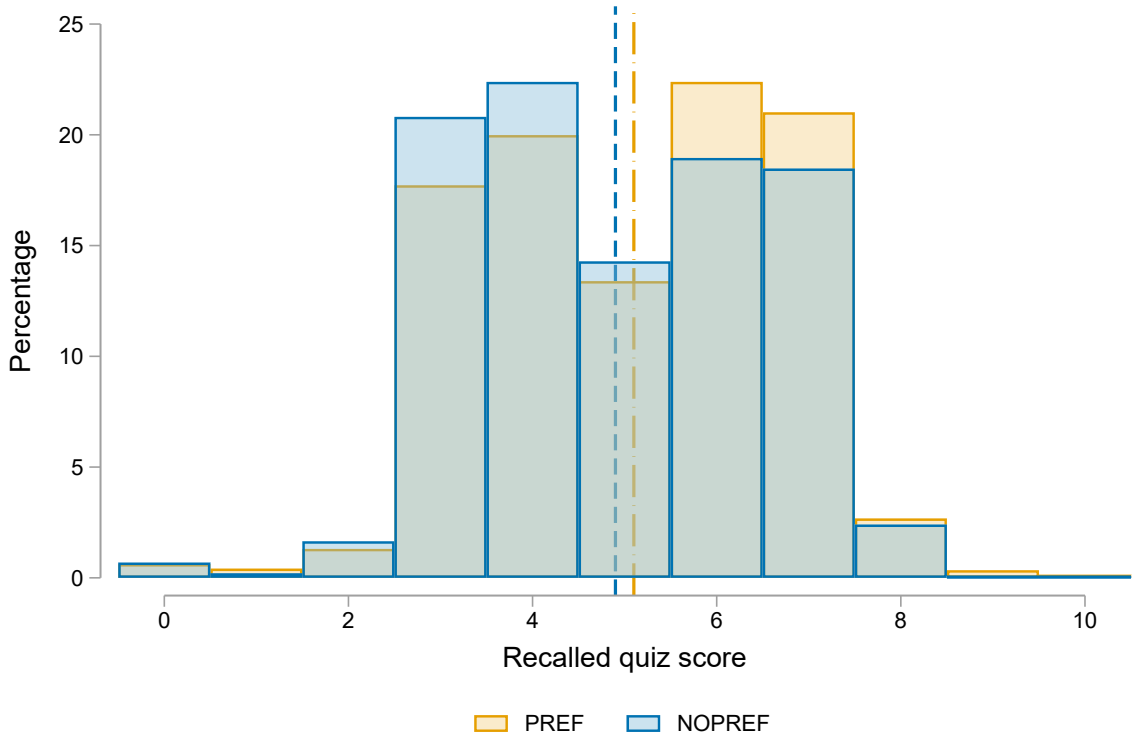


Figure 2: Distribution of recalled quiz scores

Notes: The figure shows the distribution of recalled quiz scores separately for PREF (orange) and NOPREF (blue) workers. The vertical lines indicate means of the respective group (dash/dot/orange for PREF and dashed/blue for NOPREF.

The remainder of this section aims to dig deeper into the memory mechanisms. Support for Hypothesis 1 may arise through either the extensive or intensive margin. On the extensive margin, differences in recollection may reflect variation in selective recall. That is, the probability that a recall is correct may depend on the worker's group identity. On the intensive margin, differences in recalled quiz scores may stem from both the direction and magnitude of memory errors. That is, conditional on a recall being incorrect, the error could depend on the group identity.

To explore the extensive margin, memory precision, Columns (3) and (4) of Table 1 report results from OLS regressions with the probability of correct recall as outcome variable.[24] The estimates show that PREF does not significantly affect this fraction. However, this effect masks potential heterogeneity. While managers may selectively forget low quiz scores of PREF-workers, they may selectively forget high quiz scores of NOPREF-workers. Indeed, the high quiz scores 6 and 7 are recalled correctly with almost 10 pp higher probability for PREF compared to NOPREF workers (see Table A.7). This is a sizable effect taking into account that the average probability of correct recall is 29.3%.

[24] I report the results of linear probability models for all binary outcome variables. The results are robust to using a logit specification, see Section A.5.

Table 1: Regression results: Effects of preferences on memory outcomes

	recall		correct recall		overestimation		abs. memory error	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Preferred group	0.199***	0.198***	0.014	0.015	0.047**	0.024	-0.105**	-0.104**
	(0.06)	(0.06)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
Constant	4.905***	4.277***	0.286***	0.143*	0.475***	0.635***	2.127***	1.766***
	(0.05)	(0.25)	(0.01)	(0.08)	(0.01)	(0.06)	(0.03)	(0.19)
Worker fixed effects		✓		✓		✓		✓
Manager characteristics		✓		✓		✓		✓
N	2912	2912	2912	2912	2059	2059	2059	2059

Notes: The table shows results from OLS regressions. The dependent variables are the recalled quiz score (Columns 1 and 2), the probability of correct recall in Columns (3) and (4), "overestimation": an indicator that is 1 if the recall of a incorrectly recalled quiz score is higher than the true quiz score in Columns (5) and (6) and the absolute memory error in Columns (7) and (8). Independent variables are the indicator for PREF and control variables in the even columns. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Columns (5) to (8) use the restricted sample of observations for which the recalled quiz score does not coincide with the actual quiz score. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

To explore the intensive margin, I concentrate on the subsample of incorrect recalls and first assess whether an incorrect recall is an over- or underestimation of the actual quiz score. Columns (5) and (6) of Table 1 indicate that a wrong recollection of a PREF-worker's quiz score is 4.7 pp more likely to be an overestimation (rather than an underestimation) compared to the recollection of a NOPREF-worker's quiz score. However, this effect is not robust to adding control variables. Finally, Columns (7) and (8) show regressions with the absolute memory error as outcome variable, again conditional on incorrect recalls. The average absolute memory error for PREF-workers is significantly smaller than the average absolute memory error for NOPREF-workers. Taken together, mostly memory precision and the magnitude of memory errors contribute to the gap in recalled quiz scores.

4.2 Hiring decisions

To test Hypothesis 2, namely whether reliance on memory exacerbates favoritism, I compare favoritism, which I measure as the gap in the hiring probability between PREF- and NOPREF-workers, in RECALL and REMINDER.

Figure 3 shows the hiring probability for PREF and NOPREF workers by treatment. On average, in REMINDER, a NOPREF-worker is hired with 40.6 percent probability. A NOPREF-worker is hired with 50.4 percent probability in REMINDER. Thus, the baseline level of favoritism is 9.8 pp. Supporting the idea that managers face image costs when explicitly engaging in favoritism, this difference is sizable but considerably lower than what would be predicted for expected payoff-maximizing managers. Further evidence for psychological factors stems from the fact that in

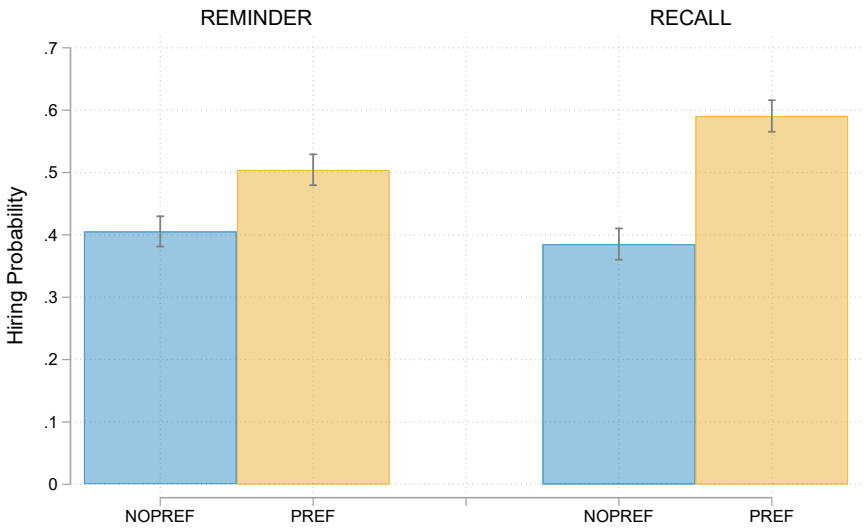


Figure 3: Hiring decision

Notes: The figures shows the hiring probability of PREF- and NOPREF-workers, seperately for the RECALL and REMINDER treatments. The gray bars show 95% Confidence intervals.

REMINDER, the level of favoritism is lower for managers who have stronger preferences against lying, a higher propensity to experience guilt, and higher moral image costs. An overview of potential moderators in Figure A.1 reveals that the characteristics related to gender do not impact the hiring gap, suggesting that aversion to favoritism is not specific to gender. Turning to the RECALL-treatment, we see that – while the hiring probability of NOPREF workers stays approximately at the level of the REMINDER treatment (38.5 percent), the hiring probability of PREF workers increases to 59.1 percent. Thus favoritism, the gap in the hiring probability between PREF- and NOPREF-workers more than doubles to 20.6 pp, showing that reliance on memory increases favoritism.

To formally test Hypothesis 2, I estimate the following model:

$$\mathtt{y}_{ij} = \beta_0 + \beta_1 \mathtt{pref}_{ij} + \beta_2 \mathtt{recall}_i + \beta_3 \mathtt{pref}_{ij} * \mathtt{recall}_i + \gamma' X_i + \mu_j + \varepsilon_{ij}. \tag{8}$$

The outcome variable $\mathtt{y}_{ij}$ is the respective decision of manager i for worker j, the indicator $\mathtt{pref}_{ij}$ is one when worker j belongs to manager i's preferred group, and $\mathtt{recall}_i$ is an indicator that is one when manager i is in the RECALL condition. X_i is a vector of manager characteristics (gender, age, whether the participant visited university, risk preferences, and patience) and μ_j are worker fixed effects. Standard errors are clustered on the manager level.

Column (1) of Table 2 shows the results of estimating Specification (8) using OLS with the binary hiring decision as the outcome variable. The gap in the hiring probability between workers in PREF and NOPREF is 10.7 pp larger in RECALL compared to REMINDER. These results

Table 2: Regression results: Effects of PREF and RECALL on personnel decisions

	hire		recommendation		ranking	
	(1)	(2)	(3)	(4)	(5)	(6)
PREF	0.099***	0.099***	4.112***	4.126***	-0.864***	-0.865***
	(0.01)	(0.01)	(0.44)	(0.44)	(0.05)	(0.05)
RECALL	-0.020	-0.019	-1.949**	-1.868**	0.013	0.009
	(0.02)	(0.02)	(0.78)	(0.79)	(0.06)	(0.06)
PREF x RECALL	0.107***	0.105***	2.895***	2.821***	-0.027	-0.019
	(0.02)	(0.02)	(0.92)	(0.92)	(0.11)	(0.11)
Constant	0.406***	0.122**	48.217***	31.084***	4.932***	5.780***
	(0.01)	(0.05)	(0.37)	(2.34)	(0.03)	(0.07)
Worker fixed effects		✓		✓		✓
Manager characteristics		✓		✓		✓
N	6048	6048	6048	6048	6048	6048

Notes: The table shows results from OLS regressions. The dependent variable is the hiring probability in Columns (1) and (2), the recommendation in Columns (3) and (4) an the rank in Columns (5) and (6). Independent variables are the indicator for PREF, the indicator that is 1 for managers in RECALL, as well as the interaction of both. Further control variables in the even columns include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

are not affected by the inclusion of worker fixed effects and manager characteristics in Column (2) and corroborate the impressions from Figure 3 in providing support for Hypothesis 2.

Result 2. *Favoritism (i.e., the gap in the hiring probability between PREF- and NOPREF-workers) is higher in the RECALL treatment than in the REMINDER treatment.*

The experiment features two additional personnel decisions: a recommendation and a ranking decision. Regarding the recommendation decision, the general pattern is identical to the one of the hiring decision: The gap in the recommendations of PREF and NOPREF workers increases by $2.8 - 2.9$ points from REMINDER to RECALL (see Columns (3) and (4) in Table 2). The ranking decision by contrast shows a different pattern. In REMINDER, the average rank of workers in PREF and NOPREF differs significantly: While a NOPREF-worker is on average ranked 4.9, a PREF-worker is ranked on average 4.1, almost a full rank better. However, in comparison to the hiring and recommendation decision, the gap in the ranking is the same in RECALL. The regression results in Columns (5) and (6) of Table 2 confirm that the interaction term between PREF and RECALL is not statistically different from zero. Possibly, the necessity to use tie-breaking rules reduce the psychological disutility related to favorable treatment of PREF-workers, leaving no scope for memory to increase the hiring gap. Indeed, the ranking choices in the REMINDER treatment seem to be overall in line with the following strategy: Rank first the best PREF-worker, then the best NOPREF-worker, then the second best PREF-worker and so on, see Figure A.2. Out



Figure 4: Individual level analysis of hiring decisions

Notes: The figures shows jittered scatterplots. Each dot represents one manager and indicates her number of hired NOPREF and PREF workers. Panel A. shows the REMINDER treatment and Panel B. shows the RECALL treatment. The colors indicate a mangers' type. A manager who hires the same number of workers from PREF as of NOPREF is categorized as meritocrat (blue squares). A manager who hires the more PREF-workers than NOPREF-workers is categorized as favoritist (orange circles).

of the managers in REMINDER, 46% follow this strategy exactly. Thus, many managers may have reached a strategy that does not provide scope to increase favoritism.

Next, I turn to an individual-level analysis. Figure 4 shows jittered scatterplots in which each point corresponds to a manager, with the horizontal axis reporting the number of hired PREF workers and the vertical axis reporting the number of hired NOPREF workers. Panel A. shows the data from the REMINDER treatment and Panel B. shows the data from the RECALL treatment. The colors represent a manager's type. I classify managers according to whether they behave as favoritist or meritocrat. Meritocrats are those who hire the same number of workers from PREF as from NOPREF, and favoritists hire more PREF-workers than NOPREF-workers. The majority of managers in REMINDER is meritocrat (58.7%) while the fraction of meritocrats in RECALL is only 20.3% ($p < 0.001$, Wilcoxon rank-sum (Mann-Whitney) test). Among meritocrats in REMINDER, most (88.2%) follow the strategy to hire workers if and only if the quiz score is above five. In contrast, only 10.7% of managers in RECALL follow the strategy to hire workers if and only if the recalled quiz score is above five. Instead, the majority of managers in RECALL are favoritists: 62.9% of them hire more PREF workers than NOPREF workers. This fraction of

favoritists is significantly higher than in the REMINDER treatment, where favoritists constitute 35.7% ($p < 0.001$, Wilcoxon rank-sum (Mann-Whitney) test). In conclusion, relying on memory affects hiring decisions by changing the distribution of the type classification.

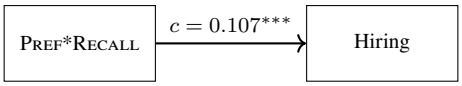
4.3 Mechanisms: Direct vs. indirect effects

The preceding section documents that reliance on memory more than doubles the hiring gap between favored and unfavored workers. I also documented that managers remember favored workers more positively. Biased memories are, however, unable to explain the full effect. A back-of-the-envelope calculation shows that in the REMINDER treatment, an additional quiz score on average leads to a 21.5 pp higher hiring probability (see Column (3) in Table A.1). Therefore, a 0.2 difference in quiz scores (Section 4.1) cannot explain the full treatment effect of more than 10 pp (Section 4.2). This motivates a mediation analysis to quantify the importance of the effect that is mediated through motivated memories.

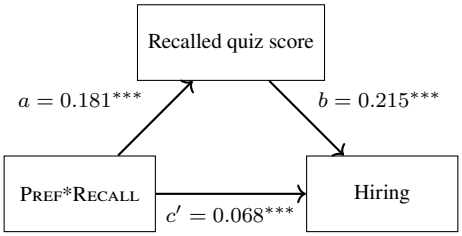
As derived in the theoretical framework (Section 3), the total treatment effect is composed of two mechanisms, one operating through distortions in recalled performance, and one operating through changes in how group identity is weighted when decisions rely on memory. To decompose these two effects, I conduct a causal mediation analysis (Baron and Kenny, 1986). In the mediation analysis, I assess the extent to which the effect of RECALL on favoritism (i.e., the effect of PREF∗RECALL on the hiring decision) is due to changes in the mediator variable (i.e., the recalled quiz score). This constitutes the indirect (mediation) effect. Because the distribution of true quiz scores is identical across PREF and NOPREF, any systematic mediation operates through differential recall rather than underlying quality. The remaining effect – the direct effect – captures the additional weight managers in RECALL place on group identity beyond what is explained by distortions in recalled quiz scores.

Figure 5 illustrates the results of the mediation analysis. Panel A presents the total effect of RECALL on favoritism (i.e., the difference in the hiring probability between PREF- and NOPREF-workers). This 10.7 pp increase was already documented in Table 2, Column (1). Panel B of Figure 5 shows the decomposition of this effect into the indirect effect, which operates through positive distortions in the recalled quiz score of PREF-workers compared to NOPREF-workers ($a \times b$), and the direct effect of RECALL in increasing favoritism (c'). I report the results of the corresponding regressions in Table A.1. The indirect, mediated effect is $a \times b = 0.039$.[25] Thus, around 36% of the total effect is mediated through distortions in the recalled quiz score. This also means that the direct effect of PREF explains around two thirds of the total effect and is

[25]Note that the numbers have to be taken with a grain of salt. First, noise in the memory elicitation may lead to an underestimation of the indirect effect (Gillen et al., 2019). Second, the causal interpretation of the mediation analysis relies on the strong assumption of sequential ignorability (Imai et al., 2010), which posits that not only the treatment assignment has to be random, but also the mediator has to be independent of the outcome given the treatment. The second part of this assumption cannot be tested. The sensitivity analysis proposed by Imai et al. (2010) however suggests that the indirect effect is positive even for mild violations of the assumption, see Figure A.3.



A. Total effect model (c)



B. Mediation model: Direct effect (c') and indirect effect ($a \times b$)

Figure 5: Mediation Through Recalled Quiz Score

Notes: This figure quantifies the impact of RECALL on favoritism (i.e., the difference in hiring between PREF and NOPREF), and decomposes the total effect into a direct effect and an indirect effect, mediated through positively distorted recalled quiz scores of PREF-workers. Panel A.: c is the total effect of RECALL on favoritism. Panel B.: a is the effect of RECALL on the recalled quiz scores of PREF-workers compared to NOPREF-workers. b and c' are respectively the impact of the recalled quiz score on the hiring decision and RECALL on favoritism in a model that incorporates both variables. The numbers reported in the Figure are based on the regression results in Table A.1.

thus essential for the difference in the hiring gap between RECALL and REMINDER. Akin to the concept of 'implicit discrimination' (Barron et al., 2025), the ambiguity that arises due to the noise in memory provides an excuse to hire the preferred group.[26]

Result 3. *Both motivated memories and reliance on memory per se drive the increase in favoritism in RECALL compared to REMINDER.*

5 Additional results

This section discusses additional results that shed light on the mechanisms at play, as well as an overview of heterogeneous treatment effects and conducted robustness checks.

5.1 Awareness of memory-induced favoritism

To shed light on managers' awareness that reliance on memory curbs favoritism, I leverage data from my questionnaire. Managers' reported number of female and male hires, their free-form descriptions of their hiring strategies, and their stated importance of gender in hiring decisions collectively suggest that managers seem to be unaware of the widening hiring gap between PREF and NOPREF from the REMINDER to the RECALL treatment.

[26]In principle, ambiguity aversion could also explain an increase in the hiring probability of the PREF workers as hiring them yields a sure payment. Ambiguity aversion would however also lead to a lower hiring probability of NOPREF workers when decisions are based on memories. My data does not reflect this pattern (see Section 4.2). It is unlikely that ambiguity aversion explains the increase in hiring of PREF-workers.

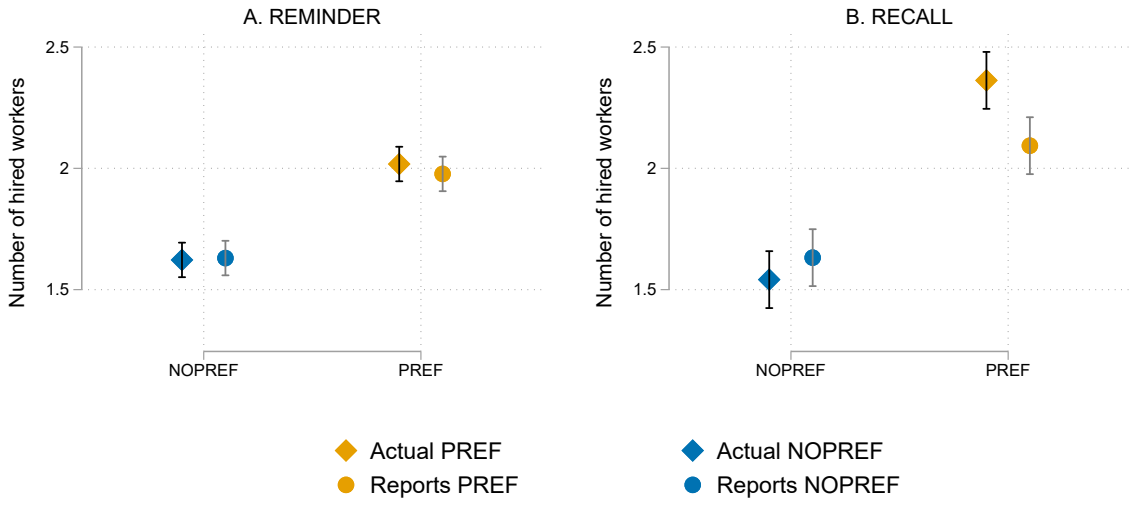


Figure 6: Number of workers hired: actual vs. stated

Notes: The graph shows the average number of hires (diamonds) and average reported number of hires (circles) separately for NOPREF- and PREF-workers. Panel A shows the results for the REMINDER treatment, while Panel B shows the results for the RECALL treatment. The bars show 95% Confidence intervals.

To examine whether the 'excess' use of gender in RECALL is unconscious, I added two simple questions in the questionnaire to provide insights into metacognition: *"Please recall: How many male workers did you hire?"* and *"Please recall: How many female workers did you hire?"*. Figure 6 shows the actual average number of workers hired in comparison to the reports. Overall, recollections are well calibrated. Managers are on average aware of their baseline bias, that is, that they hire less NOPREF-workers than PREF-workers in REMINDER. However, in RECALL, they understate the number of hired PREF-workers. In particular, they do not incorporate the additional 'memory-induced favoritism' into their reports. As the observed favoritism in REMINDER is reflected in self-reports, it is unlikely that the discrepancy between the actual number of hired workers and the reports thereof result from social desirability bias. Therefore, I interpret this result as suggestive evidence for naivete about 'memory-induced favoritism'.

This pattern is consistent with existing evidence on subconscious discrimination (Carlana, 2019; Bertrand et al., 2005). It also aligns with findings by Saccardo and Serra-Garcia (2023), who find that individuals who choose environments conducive to cognitive flexibility still engage in self-serving belief distortion. Moreover, evidence on managers' self-stated strategies corroborates this interpretation. First, more free-form strategies in REMINDER make use of the gender than in RECALL (74.8% vs. 64.7%, Two-sample Wilcoxon rank-sum (Mann–Whitney) test, $p = 0.003$).[27] Second, managers in REMINDER rate the gender as more important than

[27] I use GPT-4o to classify the free-form answers to the question: *"When you were asked to hire workers. What were your considerations that determined whether you hired a candidate or not? Please answer in 2-3 sentences."* according to whether managers stated strategies made use of gender or not. Appendix A.3 documents the procedure in more detail.

managers in RECALL in a closed-form question that yields a non-binary measure of the importance of the worker's gender for the hiring decision.[28]

5.2 Welfare effects

What consequences has 'memory-induced favoritism' for the welfare of the different players?

The firm is interested in hiring workers with high quiz scores. Not surprisingly, the average quiz score of a hired worker is lower in RECALL (5.29) than in REMINDER (6.33) due to the noise in memory. However, this gap is especially pronounced for NOPREF-workers. The quality loss of the average PREF-worker due to memory is 0.25 quiz scores less than the average quality loss of a NOPREF-worker ($p < 0.001$, p-value from OLS regression). In the long run, when employers learn from their experiences in hiring, such differences could foster inaccurate statistical discrimination (Lepage, 2024).

While the focus of this study is the managers and the workers do not experience actual payoff consequences (to avoid social preferences to drive the treatment effect), a suitable welfare measure would be whether or not they are hired. Workers in PREF benefit from the memory-reliance of managers. They are 8.6 pp more likely to be hired in RECALL compared to REMINDER ($p < 0.001$, p-value from OLS regression). In comparison, there is no significant effect of RECALL vs. REMINDER for NOPREF workers ($p = 0.263$, p-value from OLS regression). This suggests that while advantaged groups can benefit, the gap between advantaged and disadvantaged groups increases due to reliance on memory. In the long-run, under-representation of the disadvantaged groups could foster inaccurate statistical discrimination (Struth and Thon, 2022; Fonseca and McCrea, 2023). Across worker groups, it is the low-quality workers who benefit from RECALL. The probability a worker with quiz score below 5 is hired is 33 pp more likely in RECALL compared to REMINDER. Thus, memory-based decisions would lead to a decrease in merit-based hiring in general.

Finally, from the perspective of the managers, the expected utility constitutes the ideal welfare measure. As I cannot measure managers' utility, I can only use material utility (expected payoff) as a welfare approximation. There is no difference in average payoff related to the hiring decisions between managers in RECALL and REMINDER (two-sample Wilcoxon rank-sum (Mann-Whitney) test, $p = 0.978$). Thus, with my calibration, the loss due to lower quality of workers in RECALL is compensated for by the benefit from engaging in favoritism. Naturally, in other settings, one of the effects could dominate.

[28] The precise question was: *"How important was the worker's gender for your hiring decision?"*. Managers had to provide their answers on a scale from 0 to 100. Importance of the gender was assessed 60 on average in REMINDER and 56 on average in RECALL (two-sample Wilcoxon rank-sum (Mann-Whitney) test, $p = 0.023$).

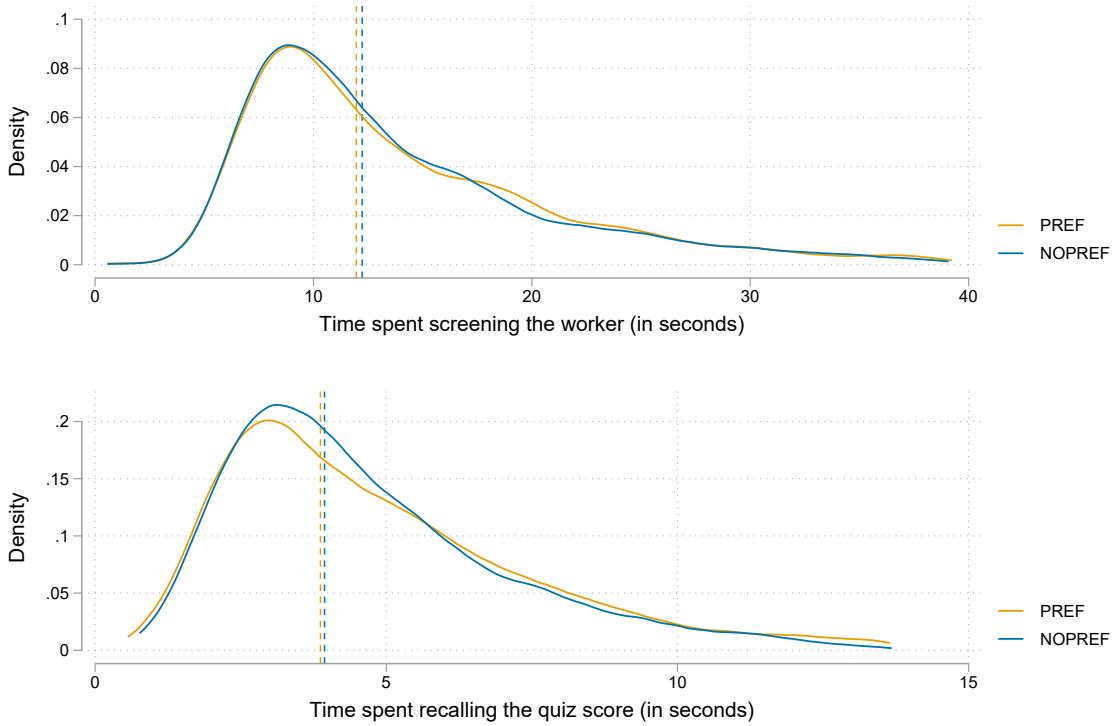


Figure 7: Distribution of page times

Notes: The figure shows distributions of the screening time (above graph, data from both treatments) and the recall time (lower graph, only RECALL treatment), truncated at the 95th percentile. The dashed vertical lines indicated medians of the respective groups.

5.3 Attention and effort

Theoretical models of motivated memory (Bénabou and Tirole, 2002; Hagenbach and Koessler, 2022) assume that memory is distorted at the encoding stage. Alternatively, biased memories may also form because of endogenous memory rehearsal (Conlon, 2024). Experimental evidence has shown that providing very high incentives in the decoding stage can eliminate motivated memory biases (Zimmermann, 2020; Gödker et al., 2025), indicating that effort at the retrieval stage may be relevant. While I cannot disentangle the exact moment at which memories become biased, the early announcement of the incentives allows me to capture distortions arising throughout the whole memory process. Suggestive evidence on the importance of encoding and decoding may be derived from examining page times. In particular, I can use page times as proxies for attention (time spent screening a worker) and effort (time spent recalling a worker's quiz score). As visualized in Figure 7, the distributions of screening and recall times do not look systematically different for PREF and NOPREF. This impression is supported by regression analyses (Table A.2). Thus, attention and effort are not likely to explain the gap in recalls.

5.4 Cognitive flexibility and ex-post rationalization

Reflecting on a worker's quiz score may limit cognitive flexibility in the decisions (Saccardo and Serra-Garcia, 2023). With my design, I can explicitly test this possibility as I vary whether the recall of the quiz score is elicited before or after the hiring decision. I find that having to state the memory before making a decision does not significantly impact the main conclusions regarding the personnel choices, see Table A.3. However, directionally, the evidence supports the intuition that not explicitly thinking about the quiz score before making a decision increases cognitive flexibility and thus fosters engagement in 'motivated favoritism'.

People adapt their beliefs and preferences to ex-post rationalize their decisions to reduce cognitive dissonance (Festinger, 1957). The psychological literature documents 'choice-supportive misremembering', the phenomenon of misremembering choice-related information that boosts the chosen option and/or demotes the forgone options (see Lind et al., 2017, for a review). To investigate the possibility of ex-post rationalization, I compare the memory outcomes when recall is elicited before to those elicited after choices. The memory outcomes elicited before the choices are not significantly different from those elicited after the choices (see Table A.4). However, directionally, the patterns are in line with an additional distortion in the memories rising from an ex-post adjustment of beliefs to 'excuse' favoritism.

5.5 Memory confidence

To investigate memory confidence, I use two complementary measures: overestimation and overprecision. The data from the final questionnaire features a measure of overconfidence that follows the concept of overestimation. In particular, I ask: *'How many of your recalled quiz scores do you think were correct?"*. While the answer to the question is 3.65 on average, the actual number of correct recalls is 2.34 on average (Wilcoxon signed-rank test, $p < 0.001$). Thus, managers tend to overestimate how well they performed. Furthermore, the perceived recall quality correlates with the actual number of correct recalls (correlation 0.389, $p < 0.001$) and the stated difficulty to remember the quiz scores (correlation 0.475, $p < 0.001$).

To collect a complementary measure defined on the observation level (manager$*$worker), I use the definition of overprecision: the excessive confidence that one knows the truth. To calculate overprecision, I subtract the subjective absolute memory error (reported) from the objective absolute memory error ($|$quiz score $-$ recalled quiz score$|$). On average, managers tend to be underprecise. They believe that their absolute memory error is 75% larger than it really is. The correlation between objective and subjective absolute memory errors is 0.069 and therefore very weak. Thus, managers are poor in predicting their recall quality with the subjective error method. Taken together, the data suggests that the overprecision elicitation may not have worked in this context, especially considering the rare occurrence of underprecision

(see Moore, 2023). Still, I investigate how overprecision varies across worker groups.[29] It is theoretically unclear whether and how overprecision varies with PREF. For example, if people are aware that they state biased memories, they could state a high absolute subjective error for PREF-workers to provide themselves with an 'excuse' to be wrong. It is however also possible that participants are aware of their biased beliefs therefore pretend to remember correctly by stating a low absolute subjective error for PREF-workers. The regression results in Table A.5 suggest that PREF does not significantly affect overprecision. This result should be taken with a grain of salt given the concerns with the subjective error measure as described above.

5.6 Heterogeneity by quiz score

To explore whether the results depend on workers' qualities, I first ask: How does the increase in favoritism depend on the quiz scores? Figure 8 shows the hiring probability as a function of the quiz score separately for PREF- and NOPREF-workers. Panel A shows the hiring probabilities in the REMINDER-treatment, where managers have full information on workers' quiz scores. Reassuringly, the hiring probability is increasing in the quiz score. The hiring probability for NOPREF workers who score lower than 5 is very close to 0 (2.8%) and jumps to a higher level for workers with quiz scores above 5 (78.3%). Moreover, the gap in the hiring probability between PREF- and NOPREF-workers is significant for every quiz score. Turning to the RECALL-treatment in Panel B, we see that the slope is substantially less steep, which naturally reflects that the recalls of quiz scores are noisy (a form of 'behavioral attenuation', see Enke et al., 2024). Most importantly, favoritism is higher in RECALL than in REMINDER for all quiz scores (see Panel C of Figure 8). This 'memory-induced favoritism' is significant for most quiz scores: quizscore 3: $p < 0.001$, quizscore 4: $p = 0.140$, quizscore 6: $p = 0.003$, quizscore 7: $p < 0.001$ (Wilcoxon rank (Mann-Whitney) tests).

While favoritism in RECALL seems to be highest for workers with quiz score 7, the effect of RECALL on favoritism is not specifically driven by high quiz scores (see the triple interaction effect in Table A.6). By contrast, the highest quiz scores drive motivated memory distortions. Those workers with quiz score equal to 7 from the PREF-group are on average recalled to have a by 0.37 higher quiz score compared to workers from NOPREF with the same quiz score. (see Table A.7). To align these two observations, I conduct the mediation analysis separately for workers with quiz score 7 and the remaining workers. Indeed, for workers with the highest quiz scores, the indirect effect is more than twice as large as the indirect effect via motivated memories in the full sample and explains 43% of the total effect of RECALL on favoritism (see Figure A.4 and Table A.8). For quiz scores below 7, the indirect effect is small and insignificant (see Figure A.5 and Table A.9).

[29] As overestimation is elicited only once for each manager, I am not able to conduct similar analyses using this confidence measure.

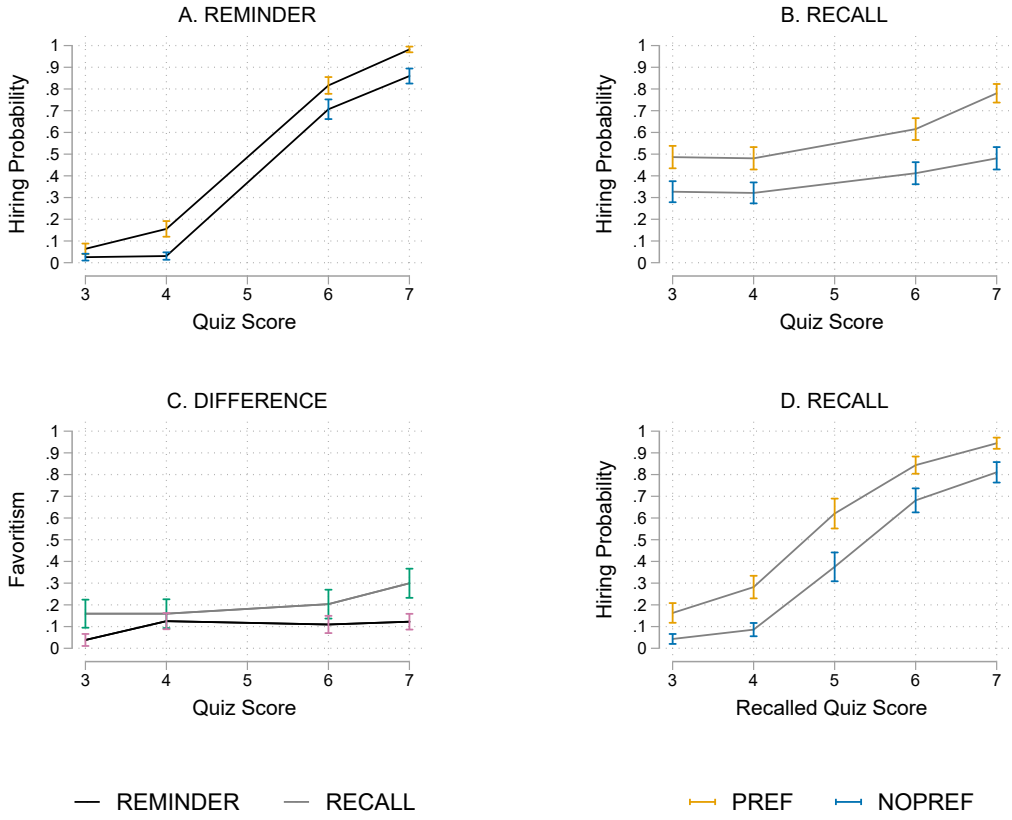


Figure 8: Hiring probabilities by quiz score

Notes: The top panels show the hiring probability as a function of the quiz score separately for PREF- and NOPREF-workers for the REMINDER treatment (Panel A) and the RECALL treatment (Panel B). Panel C shows favoritism (i.e. the difference between the hiring probability of PREF and NOPREF workers) as a function of the quiz score separately for the REMINDER and the RECALL treatment. Panel D shows the hiring probability as a function of the *recalled* quiz score separately for PREF- and NOPREF-workers for the RECALL treatment. For better visibility, the data on which Panel D is based excludes recalls below 3 and above 7. The colored bars show 95% Confidence intervals.

Panel D of Figure 8 shows again the hiring probability in RECALL, but as a function of the *recalled* quiz score.[30] The slope in this graph is similar compared to the slope in Panel A. That is, the way people take their recalled quiz score into account is similar to how they take the actual quiz score into account. However, when basing the decision on recalled quiz scores, favoritism increases, illustrating the direct effect.

5.7 Heterogeneity by gender

While any preferences for or against a particular gender are orthogonal to the treatment variation, the treatment effect could be moderated by gender identity or gender preferences. To examine this dimension of heterogeneity, I first consider the role of the manager's gender. Second, I consider whether it matters if females or males constitute the PREF group, and third, I consider whether the PREF group is of the same gender as the manager. I run the main regression

[30] To enhance clarity, I excluded the 5.3% of observations with recalls below 3 or above 7 from the analysis.

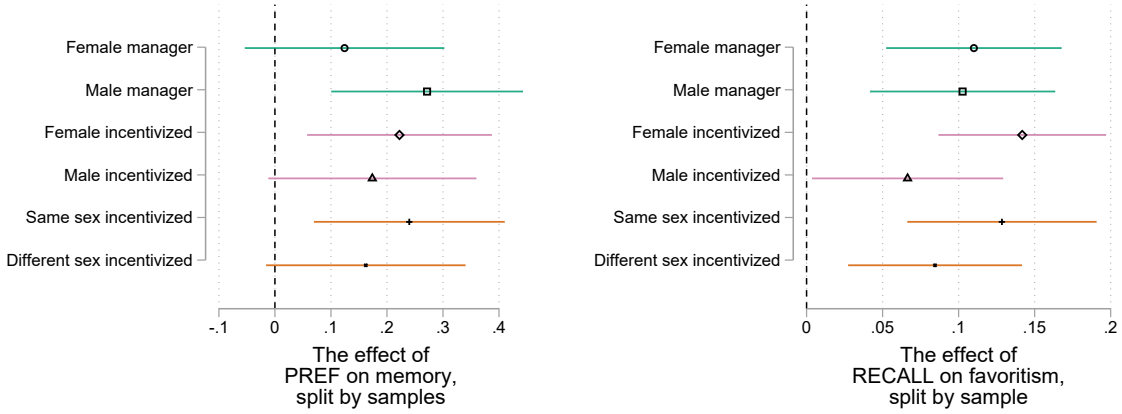


Figure 9: Heterogeneity by gender

Notes: The figure shows interaction coefficients from regression. The left panel uses data from the RECALL treatment and shows the effect of PREF on the recalled quiz scores separately for each subsample. The right panel shows the interaction effect of PREF and RECALL on the hiring decision across the different subsample. The graphs show 95% Confidence intervals.

separately for the respective samples (e.g. female manager and male manager). The estimates of the coefficients of interest (i.e., PREF for the recalled quiz scores and PREF*RECALL for the hiring decision) are visualized in Figure 9.

The results offer two takeaways. First, 'memory-induced favoritism' is stronger when females constitute the PREF group (Wald test, $p = 0.077$). Second, none of the main results is solely driven by a specific gender constellation.

5.8 Robustness checks

I provide a series of robustness checks. These show that the main results are robust to a) using stricter and more lenient exclusion criteria (showing that the results are highly robust to excluding possibly confused participants), b) using PDS Lasso to select control variables, c) estimating Logit regressions for the binary outcome variables, and d) adjusting for multiple hypothesis testing. I provide all details in Appendix A.5, where I also present a balance test and simulations to show that memory is better than random noise.

6 Conclusion

This study provides causal evidence on the role of memory for favoritism in hiring decisions. Exogenously manipulating the possibility to forget informative signals, I find that managers recall workers of their preferred group more favorably. Further, allowing managers to forget objective quality measures of workers significantly increases favoritism in personnel decisions.

Such 'memory-induced favoritism' has detrimental welfare implications not only for firms that miss out on the optimal allocation of talent but also for workers who are not evaluated objectively based on their performance.

The experiment offers a proof of concept for a mechanism that can sustain discrimination in the long run. While the objective quality signals in my experiment are simple and deterministic, uncertainty and more complex information may reinforce this mechanism in real-world settings. Especially in modern work environments with non-routine, analytical tasks, precise and objective performance measures are often not available. Instead, subjective assessments are used, which have been shown to be crucially influenced by preferences, for example, due to social connections (Breuer et al., 2013) or gender (Benson et al., 2024). Even in technology-aided hiring, managers may override algorithmic recommendations (Hoffman et al., 2018). Because subjective evaluations rely on a multitude of different recalled signals, my findings suggest that memory-based assessments exhibit greater favoritism. This mechanism can help to explain the evidence on persistent favoritism.

In situations where reliance on imperfect memory causes favoritism, this paper has shown that rigorous bookkeeping and reminders of quality signals can be valuable. Thereby, my findings highlight that efforts to promote fairness in hiring and evaluation decisions should not only target attitudinal changes – for example through the use of effective diversity training (Arslan et al., 2025) and the establishment of non-discriminatory corporate values (Drechsel-Grau and Holub, 2025), but also the design of the environment in which decisions are made, for instance, by increasing the frequency of evaluations. Future research needs to test the effectiveness of such interventions in organizations.

References

Amelio, A. and F. Zimmermann (2023). Motivated memory in economics—a review. *Games 14*(1), 15.

Aquino, K. and A. Reed II (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology 83*(6), 1423.

Arslan, C., E. H. Chang, S. Chilazi, I. Bohnet, and O. P. Hauser (2025). Behaviorally designed training leads to more diverse hiring. *Science 387*(6732), 364–366.

Badalyan, S., D. Korlyakova, and R. Rehák (2023). Disclosure discrimination: An experiment focusing on communication in the hiring process. *CERGE-EI Working Paper Series* (743).

Bandiera, O., I. Barankay, and I. Rasul (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica 77*(4), 1047–1094.

Baron, R. M. and D. A. Kenny (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology 51*(6), 1173.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2025). Explicit and implicit belief-based gender discrimination: A hiring experiment. *Management Science 71*(2), 1600–1622.

Bartoš, V., M. Bauer, J. Chytilová, and F. Matějka (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review 106*(6), 1437–1475.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Belot, M. and M. Schröder (2023). Remember me? The role of gender and racial attributes in memory. *Journal of Behavioral and Experimental Economics 104*, 102008.

Bénabou, R. and J. Tirole (2002). Self confidence and personal motivation. *The Quarterly Journal of Economics 117*(3), 817–915.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Benson, A., D. Li, and K. Shue (2024). Potential and the gender promotions gap. *Available at SSRN 4747175*.

Bertrand, M., D. Chugh, and S. Mullainathan (2005). Implicit discrimination. *American Economic Review 95*(2), 94–98.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2025). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics 107*(3), 605–620.

Bohren, J. A., P. Hull, and A. Imas (2025). Systemic discrimination: Theory and measurement. *The Quarterly Journal of Economics 140*(3), 1743–1799.

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

Bosch-Rosa, C., B. Kassner, and S. Ahrens (2024). Overconfidence and the political and financial behavior of a representative sample. *Berlin School of Economics Discussion Paper 51*.

Breuer, K., P. Nieken, and D. Sliwka (2013). Social ties and subjective performance evaluations: an empirical investigation. *Review of Managerial Science 7*(2), 141–157.

Burks, S. V., B. Cowgill, M. Hoffman, and M. Housman (2015). The value of hiring through employee referrals. *The Quarterly Journal of Economics 130*(2), 805–839.

Campos-Mercade, P. and F. Mengel (2024). Non-bayesian statistical discrimination. *Management Science 70*(4), 2549–2567.

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

Carlson, R. W., M. A. Maréchal, B. Oud, E. Fehr, and M. J. Crockett (2020). Motivated misremembering of selfish decisions. *Nature Communications 11*(1), 2100.

Cavounidis, C., K. Lang, and R. Weinstein (2024). The boss is watching: How monitoring decisions hurt black workers. *The Economic Journal 134*(658), 485–514.

Celebi, C. and S. Penczynski (2024). Using large language models for text classification in experimental economics. *Working paper*.

Chew, S. H., W. Huang, and X. Zhao (2020). Motivated false memory. *Journal of Political Economy 128*(10), 3913–3939.

Cilliers, J., N. Elashmawy, and D. McKenzie (2024). Using post-double selection lasso in field experiments. *Policy Research Working Paper 10931*.

Coffman, K., S. Kostyshak, and P. Saygin (2025). Choosing and using information in evaluation decisions. *arXiv:2507.13798*.

Cohen, T. R., Y. Kim, and A. Panter (2014). The five-item guilt proneness scale (gp-5). *Differences 92*, 109–112.

Conlon, J. J. (2024). Memory rehearsal and belief biases. *Working paper*.

Cordes, C., J. Friedrichsen, and S. Schudy (2025). Motivated procrastination. *CESifo Working Paper No. 11072*.

Cullen, Z. and R. Perez-Truglia (2023). The old boys' club: Schmoozing and the gender gap. *American Economic Review 113*(7), 1703–1740.

Cunningham, T. and J. De Quidt (2022). Implicit preferences. *CEPR Discussion Paper No. DP17343*.

Dana, J., R. A. Weber, and J. X. Kuang (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory 33*(1), 67–80.

Deserranno, E., P. Kastrau, and G. León-Ciliotta (2025). Promotions and productivity: the role of meritocracy and pay progression in the public sector. *American Economic Review: Insights 7*(1), 71–89.

Di Tella, R., R. Perez-Truglia, A. Babino, and M. Sigman (2015). Conveniently upset: avoiding altruism by distorting beliefs about others' altruism. *American Economic Review 105*(11), 3416–3442.

Dickinson, D. L., D. Masclet, and E. Peterle (2018). Discrimination as favoritism: The private benefits and social costs of in-group favoritism in an experimental labor market. *European Economic Review 104*, 220–236.

Drechsel-Grau, M. and F. Holub (2025). Are male bosses bad for women's careers? Evidence from a multinational corporation. *CESifo Working Paper No. 11622*.

Enke, B., T. Graeber, R. Oprea, and J. Yang (2024). Behavioral attenuation. *NBER working paper No. w32973*.

Esponda, I., R. Oprea, and S. Yuksel (2023). Seeing what is representative. *The Quarterly Journal of Economics 138*(4), 2607–2657.

Evsyukova, Y., ®. Rusche®, and W. Mill (2025). Linkedout? A field experiment on discrimination in job network formation. *The Quarterly Journal of Economics 140*(1), 283–334.

Eyting, M. (2022). Why do we discriminate? The role of motivated reasoning. *SAFE Working Paper No. 356*.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press, Stanford CA.

Fonseca, M. A. and A. McCrea (2023). The role of shortlisting in shifting gender beliefs on performance: Experimental evidence. *Working paper*.

Frederiksen, A., F. Lange, and B. Kriechel (2017). Subjective performance evaluations and employee careers. *Journal of Economic Behavior & Organization 134*, 408–429.

Friebel, G., M. Heinz, M. Hoffman, and N. Zubanov (2023). What do employee referral programs do? Measuring the direct and overall effects of a management practice. *Journal of Political Economy 131*(3), 633–686.

Gagnon, N., K. Bosmans, and A. Riedl (2025). The effect of gender discrimination on labor supply. *Journal of Political Economy 133*(3), 1047–1081.

Galeotti, F., C. Saucet, and M. C. Villeval (2020). Unethical amnesia responds more to instrumental than to hedonic motives. *Proceedings of the National Academy of Sciences 117*(41), 25423–25428.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy 127*(4), 1826–1863.

Glick, P. and S. T. Fiske (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology 70*(3), 491 – 512.

Glick, P. and S. T. Fiske (1999). The ambivalence toward men inventory: Differentiating hostile and benevolent beliefs about men. *Psychology of Women Quarterly 23*(3), 519–536.

Gödker, K., P. Jiao, and P. Smeets (2025). Investor memory. *The Review of Financial Studies 38*(6), 1595–1640.

Grossman, Z. and J. J. Van Der Weele (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association 15*(1), 173–217.

Grunewald, A., V. Klockmann, A. von Schenk, and F. von Siemens (2023). Are biases contagious? the influence of communication on motivated beliefs. *Available at SSRN 4508834*.

Hagenbach, J., N. Jacquemet, and P. Sternal (2025). The motivated memory of noise. *Games and Economic Behavior 152*, 257–275.

Hagenbach, J. and F. Koessler (2022). Selective memory of a psychological agent. *European Economic Review 142*, 104012.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature 42*(4), 1009–1055.

Heursen, L., S. Friess, and M. Chugunova (2025). Reputational concerns and advice-seeking at work. *Available at SSRN 4538216*.

Hicks, R. and D. Tingley (2011). Causal mediation analysis. *The Stata Journal 11*(4), 605–619.

Hirmas, A. and J. Hausfeld (2024). Hiring for others: The role of intermediaries in discrimination. *Available at SSRN 5013215*.

Hoffman, M., L. B. Kahn, and D. Li (2018). Discretion in hiring. *The Quarterly Journal of Economics 133*(2), 765–800.

Huffman, D., C. Raymond, and J. Shvets (2022). Persistent overconfidence and biased memory: Evidence from managers. *American Economic Review 112*(10), 3141–3175.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 23*(1), 51–71.

Kerwin, J., N. Rostom, and O. Sterck (2024). Striking the right balance: Why standard balance tests over-reject the null, and how to fix it. *IZA Discussion Papers No. 17217*.

Lepage, L.-P. (2024). Experience-based discrimination. *American Economic Journal: Applied Economics 16*(4), 288–321.

Li, K. K. (2013). Asymmetric memory recall of positive and negative events in social interactions. *Experimental Economics 16*, 248–262.

Lind, M., M. Visentini, T. Mäntylä, and F. Del Missier (2017). Choice-supportive misremembering: A new taxonomy and review. *Frontiers in Psychology 8*, 2062.

Miserocchi, F. (2023). Discrimination through biased memory. *Working paper*.

Moore, D. A. (2023). Overprecision is a property of thinking systems. *Psychological Review 130*(5), 1339.

Müller, M. W. (2022). Selective memory around big life decisions. *Working paper*.

Pikulina, E. and D. Ferreira (2025). Subtle discrimination. *Journal of Finance, forthcoming*.

Rackstraw, E. (2025). Bias-motivated updating in the labor market. *Available at SSRN 4278076*.

Radbruch, J. and A. Schiprowski (2025). Interview sequences and the formation of subjective assessments. *Review of Economic Studies 29*(2), 1226–1256.

Riaño, J. F. (2023). Bureaucratic nepotism. *Available at SSRN 3995589*.

Roy-Chowdhury, V. (2024). Self-confidence and motivated memory loss: Evidence from schools. *Available at SSRN 4986405*.

Ruebeck, H. (2025). Causes and consequences of perceived workplace discrimination. *Available at SSRN 4799864*.

Saccardo, S. and M. Serra-Garcia (2023). Enabling or limiting cognitive flexibility? Evidence of demand for moral commitment. *American Economic Review 113*(2), 396–429.

Sarsons, H. (2022). Interpreting signals in the labor market: Evidence from medical referrals. *Working paper*.

Saucet, C. and M. C. Villeval (2019). Motivated memory in dictator games. *Games and Economic Behavior 117*, 250–275.

Schudy, S., S. Grundmann, and L. Spantig (2024). Individual preferences for truth-telling. *CESifo Working Paper No. 11521*.

Schwardmann, P., E. Tripodi, and J. J. Van der Weele (2022). Self-persuasion: Evidence from field experiments at international debating competitions. *American Economic Review 112*(4), 1118–1146.

Schwardmann, P. and J. Van der Weele (2019). Deception and self-deception. *Nature Human Behaviour 3*(10), 1055–1061.

Sial, A. Y., J. R. Sydnor, and D. Taubinsky (2023). Biased memory and perceptions of self-control. *NBER Working Paper No. w30825*.

Stötzer, L. and F. Zimmermann (2024). A note on motivated cognition and discriminatory beliefs. *Games and Economic Behavior 147*, 554–562.

Struth, L. and M. Thon (2022). Discrimination, quotas, and stereotypes. *Available at SSRN 4184611*.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review 110*(2), 337–61.

Online appendix

Contents

A Additional analyses **ii**

A.1 Additional figures ii

A.2 Additional tables v

A.3 Classification of free-form answers xii

A.4 Heterogeneity by manager characteristics xiv

A.5 Robustness checks xv

B Deviations from the pre-registration **xxv**

C Auxiliary experiments **xxvi**

C.1 Auxiliary experiment 1: Workers xxvi

C.2 Prompts to generate the workers' pictures xxvi

C.3 Auxiliary experiment 2: Selection of workers xxvii

C.4 Auxiliary experiment 3: The HR experiment xxviii

D Experimental material **xxix**

D.1 Instructions xxix

A Additional analyses

A.1 Additional figures

This Appendix section contains figures that are discussed in the main part of the text.

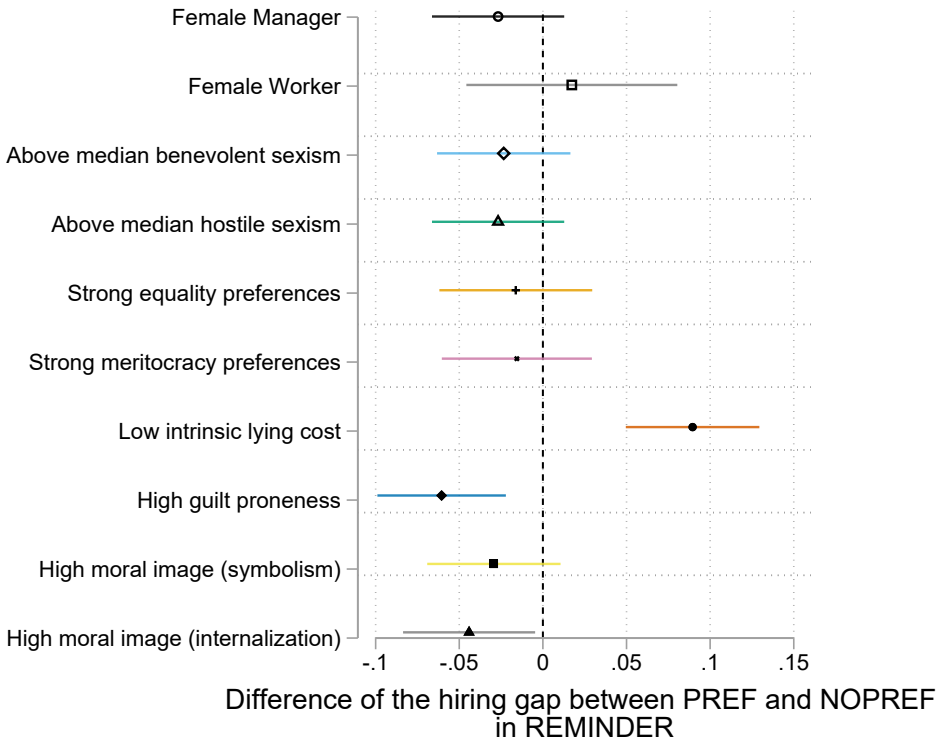


Figure A.1: Heterogeneity of favoritism in REMINDER

Notes: The graph shows the coefficients estimates of β_3 from the regression $hire_{ij} = \beta_0 + \beta_1 * pref_{ij} + \beta_2 * characteristic_i + \beta_3 * pref_{ij} * characteristic_i$, where $characteristic_i$ is the respective dummy variable in case of a binary characteristic and a dummy variable that indicates an above median value otherwise. The graphs show 95% Confidence intervals.

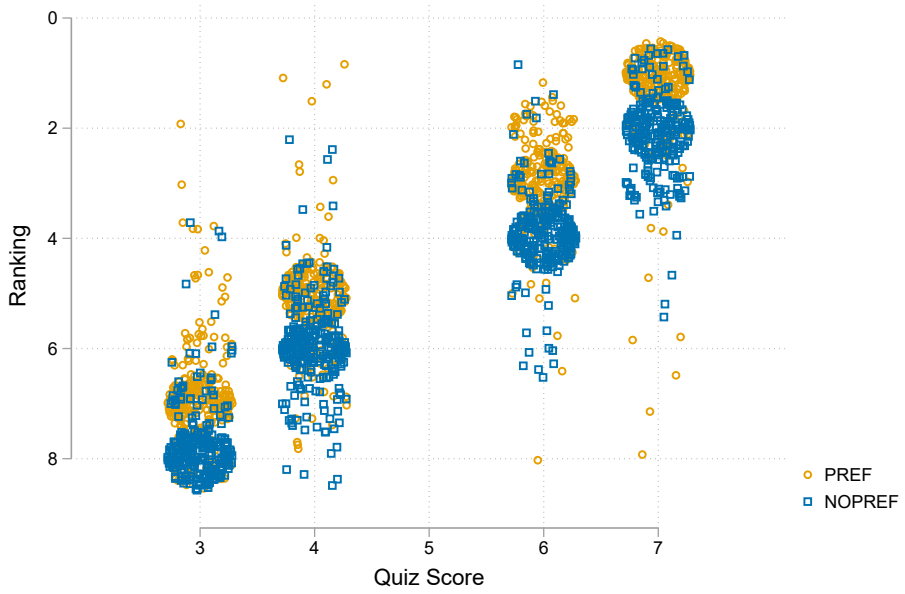


Figure A.2: Scatterplot of ranking decisions in REMINDER

Notes: The graph shows a jittered scatterplot of the ranking decision of managers in REMINDER as a function of the quiz score.

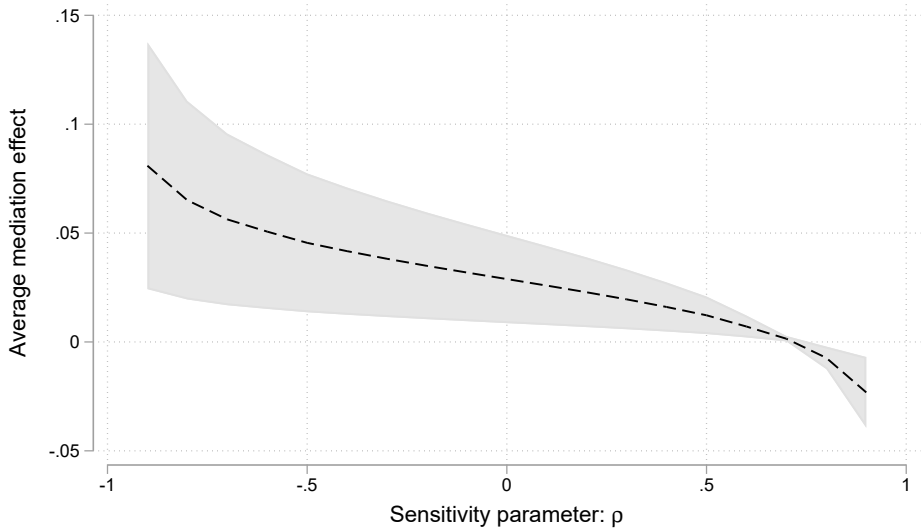
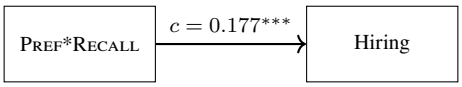
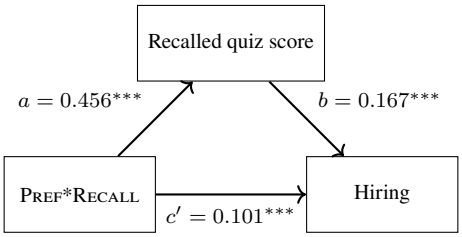


Figure A.3: Mediation through recalled quiz score: Sensitivity analysis

Notes: The graph shows how sensitive the mediation effect is relative to the sensitivity parameter ρ. The shaded area shows 95%-Confidence intervals. This sensitivity analysis follows the `medsens` procedure from the `mediation` package (Hicks and Tingley, 2011), which evaluates violations of the sequential ignorability assumption arising from unobserved confounding between the mediator and the outcome. The parameter $\rho = \text{Corr}(\varepsilon_{Mi}, \varepsilon_{Yi})$ measures sensitivity to unobserved confounding and is equal to the correlation between the error terms in the mediation model ($M_i = \alpha_0 + \alpha_1 T_i + \varepsilon_{Mi}$, see also Column (2) in Table A.1) and outcome model ($Y_i = \beta_0 + \beta_1 T_i + \beta_2 M_i + \varepsilon_{Yi}$, see also Column (3) in Table A.1)). Under sequential ignorability, one assumes $\rho = 0$. Deviations from zero represent increasing degrees of unobserved confounding that jointly affect the mediator and outcome.



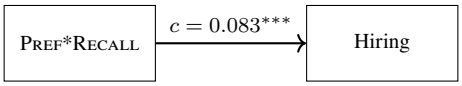
A. Total effect model (c)



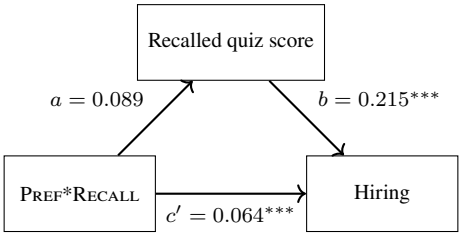
B. Mediation model: Direct effect (c') and indirect effect ($a \times b$)

Figure A.4: Mediation Through Recalled Quiz Score: Quizscore = 7

Notes: This figure quantifies the impact of Recall on favoritism (i.e., the difference in hiring between Pref and Nopref), and decomposes the total effect into a direct effect and an indirect effect, mediated through positively distorted recalled quiz scores of Pref-workers. Panel A.: c is the total effect of Recall on favoritism. Panel B.: a is the effect of Recall on the recalled quiz scores of Pref-workers compared to Nopref-workers. b and c' are respectively the impact of the recalled quiz score on the hiring decision and Recall on favoritism in a model that incorporates both variables. The numbers reported in the Figure are based on the regression results in Table A.8.



A. Total effect model (c)



B. Mediation model: Direct effect (c') and indirect effect ($a \times b$)

Figure A.5: Mediation Through Recalled Quiz Score: Quizscore < 7

Notes: This figure quantifies the impact of Recall on favoritism (i.e., the difference in hiring between Pref and Nopref), and decomposes the total effect into a direct effect and an indirect effect, mediated through positively distorted recalled quiz scores of Pref-workers. Panel A.: c is the total effect of Recall on favoritism. Panel B.: a is the effect of Recall on the recalled quiz scores of Pref-workers compared to Nopref-workers. b and c' are respectively the impact of the recalled quiz score on the hiring decision and Recall on favoritism in a model that incorporates both variables. The numbers reported in the Figure are based on the regression results in Table A.9.

A.2 Additional tables

This Appendix Section contains figures that are discussed in the main part of the text.

Table A.1: Regression results: Mediation analysis

	(1)	(2)	(3)
	hire	recalled quiz score	hire
PREF	0.099***	0.018	0.095***
	(0.01)	(0.01)	(0.01)
RECALL	-0.020	-0.088*	-0.001
	(0.02)	(0.05)	(0.02)
PREF*RECALL	0.107***	0.181***	0.068***
	(0.02)	(0.06)	(0.02)
Recalled quiz score			0.215***
			(0.00)
Constant	0.406***	4.994***	-0.667***
	(0.01)	(0.01)	(0.02)
N	6048	6048	6048

Notes: This table shows the OLS regression coefficients underlying Figure 5. The coefficient on the interaction term in Column (1) corresponds to c, the coefficient on the interaction term in Column (2) corresponds to a, the interaction term in Column (3) corresponds to c' and the coefficient on the recalled quiz score corresponds to b. Note that I defined the recalled quiz score to be the repeated quiz score for participants in REMINDER. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.2: Regression results: Page times

	screeningtime		recall/repetition-time	
	(1)	(2)	(3)	(4)
PREF	-0.444	-0.434	0.061	0.060
	(0.38)	(0.38)	(0.15)	(0.15)
RECALL	-0.903	-0.707	0.427*	0.475**
	(0.68)	(0.66)	(0.24)	(0.23)
PREF*RECALL	0.382	0.329	0.154	0.158
	(0.51)	(0.51)	(0.20)	(0.20)
Constant	16.637***	6.471***	5.043***	1.700**
	(0.48)	(1.98)	(0.17)	(0.73)
Worker fixed effects		✓		✓
Manager characteristics		✓		✓
N	6048	6048	6048	6048

Notes: The table shows results from OLS regressions. The dependent variables are the time spent screening a worker and the time spent recalling the worker's quiz score. Both variables are winzorized at the 99th percentile. Independent variables are the indicator for PREF, the indicator that is 1 for managers in RECALL as well as the interaction. Even columns also include control variables. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.3: Regression results: Heterogeneity of memory results by memory elicitation timing

	hire		recommendation		ranking	
	(1)	(2)	(3)	(4)	(5)	(6)
PREF	0.111***	0.108***	4.107***	4.005***	-0.836***	-0.825***
	(0.01)	(0.01)	(0.66)	(0.65)	(0.07)	(0.07)
Hiring first	0.038*	0.037*	-0.524	-0.498	0.030	0.043
	(0.02)	(0.02)	(0.73)	(0.74)	(0.05)	(0.05)
PREF*Hiring first	-0.026	-0.020	0.011	0.261	-0.060	-0.086
	(0.02)	(0.02)	(0.87)	(0.86)	(0.10)	(0.10)
RECALL	-0.037	-0.036	-3.020***	-2.873***	-0.003	0.002
	(0.03)	(0.03)	(1.10)	(1.09)	(0.08)	(0.08)
RECALL*Hiring first	0.035	0.036	2.265	2.114	0.033	0.014
	(0.04)	(0.04)	(1.56)	(1.55)	(0.12)	(0.11)
PREF*RECALL	0.082***	0.084***	2.105*	2.200*	0.006	-0.004
	(0.03)	(0.03)	(1.23)	(1.23)	(0.16)	(0.16)
PREF*RECALL*Hiring first	0.053	0.045	1.661	1.301	-0.067	-0.029
	(0.04)	(0.04)	(1.85)	(1.84)	(0.23)	(0.23)
Constant	0.388***	0.110**	48.461***	31.574***	4.918***	5.760***
	(0.01)	(0.05)	(0.52)	(2.37)	(0.03)	(0.07)
Worker fixed effects		✓		✓		✓
Manager characteristics		✓		✓		✓
N	6048	6048	6048	6048	6048	6048

Notes: The table shows results from OLS regressions. The dependent variable is the hiring probability in Columns (1) and (2), the recommendation in Columns (3) and (4) and the rank in Columns (5) and (6). Independent variables are the indicator for PREF, the indicator that is 1 for managers in RECALL as well as the interaction. Further, independent variables include an indicator that is one if the hiring and recommendation decision comes before the memory elicitation as well as it's interaction with the other variables. Even columns also include control variables. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.4: Regression results: Heterogeneity of memory-induced favoritism by memory elicitation timing

	recalled quiz score		correct recall		overestimation		abs. memory error	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PREF	0.150***	0.150***	0.027	0.028	0.077***	0.040	-0.189***	-0.200***
	(0.05)	(0.05)	(0.02)	(0.02)	(0.03)	(0.03)	(0.07)	(0.06)
Recall elicitation first	0.064	0.062	0.035	0.038	0.053*	0.041*	-0.133*	-0.157**
	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)	(0.02)	(0.07)	(0.07)
PREF*recall first	-0.084	-0.084	-0.025	-0.025	-0.058	-0.031	0.163*	0.186**
	(0.06)	(0.06)	(0.03)	(0.03)	(0.04)	(0.04)	(0.08)	(0.08)
Constant	4.917***	4.114***	0.267***	0.119	0.448***	0.607***	2.195***	1.866***
	(0.04)	(0.12)	(0.02)	(0.08)	(0.02)	(0.07)	(0.06)	(0.20)
Worker fixed effects		✓		✓		✓		✓
Manager characteristics		✓		✓		✓		✓
N	6048	6048	2912	2912	2059	2059	2059	2059

Notes: The table shows results from OLS regressions. The dependent variables are the recalled quiz score in Columns (1) and (2), the probability of correct recall in Columns (3) and (4), "overestimation": an indicator that is 1 if the recall of a incorrectly recalled quiz score is higher than the true quiz score in Columns (5) and (6) and the absolute memory error in Columns (7) and (8). Independent variables are the indicator for PREF, an indicator that is one if memory is elicited before the hiring decision, as well as the interaction of both. Even columns also include control variables. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Columns (5) and (8) use the restricted sample of observations for which the recalled quiz score does not coincide with the actual quiz score. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.5: Regression results: Overprecision

	overprecision	
	(1)	(2)
PREF	-0.103	-0.103
	(0.07)	(0.07)
Constant	-1.052***	0.025
	(0.11)	(0.80)
N	2912	2912

Notes: The table shows results from OLS regressions. The dependent variable is overprecision, defined as: (reported subjective absolute memory error − true absolute memory error). Independent variables are the indicator for PREF, as well as further control variables in Column (2). These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.6: Regression result: Heterogeneity of memory-induced favoritism by quiz score

	hire		recommendation		ranking	
	(1)	(2)	(3)	(4)	(5)	(6)
PREF	0.038***	0.038***	2.860***	2.865***	-0.816***	-0.818***
	(0.01)	(0.01)	(0.58)	(0.58)	(0.06)	(0.06)
qs4	0.005	-0.019	10.043***	9.064***	-1.788***	-1.757***
	(0.01)	(0.02)	(0.43)	(0.65)	(0.05)	(0.08)
qs6	0.681***	0.695***	36.291***	37.052***	-3.717***	-3.847***
	(0.02)	(0.03)	(0.70)	(0.90)	(0.05)	(0.09)
qs7	0.834***	0.823***	49.393***	48.971***	-5.492***	-5.527***
	(0.02)	(0.02)	(0.80)	(0.92)	(0.05)	(0.08)
PREF*qs4	0.087***	0.087***	0.811	0.831	0.071	0.071
	(0.02)	(0.02)	(0.62)	(0.63)	(0.07)	(0.08)
PREF*qs6	0.071***	0.071***	2.046***	2.050***	-0.145*	-0.145*
	(0.02)	(0.02)	(0.70)	(0.70)	(0.07)	(0.08)
PREF*qs7	0.084***	0.084***	2.153***	2.162***	-0.115	-0.114
	(0.02)	(0.02)	(0.67)	(0.67)	(0.08)	(0.08)
RECALL	0.301***	0.303***	17.786***	17.845***	-2.338***	-2.342***
	(0.03)	(0.03)	(1.22)	(1.23)	(0.12)	(0.12)
RECALL*qs4	- -0.011	-0.009	-10.222***	-10.169***	1.621***	1.619***
	(0.03)	(0.03)	(1.42)	(1.42)	(0.17)	(0.17)
RECALL*qs6	-0.596***	-0.596***	-31.513***	-31.502***	3.294***	3.295***
	(0.04)	(0.04)	(1.54)	(1.53)	(0.17)	(0.17)
RECALL*qs7	-0.680***	-0.680***	-37.203***	-37.181***	4.490***	4.491***
	(0.04)	(0.04)	(1.75)	(1.75)	(0.19)	(0.19)
PREF*RECALL	0.121***	0.120***	3.910***	3.879***	-0.008	0.001
	(0.04)	(0.04)	(1.49)	(1.49)	(0.18)	(0.18)
PREF*RECALL*qs4	-0.087*	-0.089*	-2.872	-2.977	0.288	0.292
	(0.05)	(0.05)	(2.03)	(2.03)	(0.24)	(0.24)
PREF*RECALL*qs6	-0.027	-0.027	-1.873	-1.895	0.003	-0.000
	(0.05)	(0.05)	(2.02)	(2.02)	(0.23)	(0.23)
PREF*RECALL*qs7	0.056	0.055	0.685	0.640	-0.366	-0.370
	(0.05)	(0.05)	(2.01)	(2.01)	(0.23)	(0.23)
Constant	0.026***	-0.005	24.286***	18.078***	7.681***	7.760***
	(0.01)	(0.05)	(0.56)	(2.28)	(0.04)	(0.05)
Worker fixed effects		✓		✓		✓
Manager characteristics		✓		✓		✓
N	6048	6048	6048	6048	6048	6048

Notes: The table shows results from OLS regressions. The dependent variable is the hiring probability in Columns (1) and (2), the recommendation in Columns (3) and (4) and the rank in Columns (5) and (6). Independent variables are the indicator for PREF, the indicator for the RECALL treatment, dummies for the worker's quiz scores, where a quiz score of 3 is the omitted category, and all interactions. Further, I include control variables in the even columns. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Columns (5) to (8) use the restricted sample of observations for which the recalled quiz score does not coincide with the actual quiz score. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.7: Regression results: Heterogeneity of memory results by quiz score

	recall		correct recall		overestimation		abs. memory error	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PREF	0.082	0.083	-0.036	-0.035	0.002	-0.001	0.026	0.038
	(0.11)	(0.11)	(0.03)	(0.03)	(0.02)	(0.02)	(0.10)	(0.10)
qs4	0.091	-0.040	-0.003	0.015	-0.300***	-0.333***	-0.614***	-0.742***
	(0.11)	(0.13)	(0.03)	(0.04)	(0.03)	(0.04)	(0.09)	(0.11)
qs6	0.357***	0.270*	-0.033	-0.026	-0.720***	-0.737***	-0.454***	-0.432***
	(0.11)	(0.14)	(0.03)	(0.04)	(0.03)	(0.04)	(0.10)	(0.11)
qs7	0.799***	0.784***	0.058*	0.043	-0.905***	-0.905***	0.152	0.071
	(0.12)	(0.14)	(0.03)	(0.04)	(0.02)	(0.02)	(0.12)	(0.13)
PREF*qs4	-0.088	-0.100	0.008	0.010	-0.012	-0.011	-0.061	-0.076
	(0.15)	(0.15)	(0.05)	(0.05)	(0.04)	(0.04)	(0.13)	(0.13)
PREF*qs6	0.181	0.187	0.096**	0.094**	0.065	0.070	-0.097	-0.117
	(0.15)	(0.15)	(0.05)	(0.05)	(0.04)	(0.04)	(0.13)	(0.13)
PREF*qs7	0.374**	0.372**	0.096**	0.095**	0.043	0.046*	-0.403**	-0.413***
	(0.15)	(0.15)	(0.05)	(0.05)	(0.03)	(0.03)	(0.16)	(0.16)
Constant	4.593***	4.420***	0.280***	0.148*	0.950***	0.988***	2.366***	2.478***
	(0.08)	(0.25)	(0.02)	(0.08)	(0.01)	(0.06)	(0.08)	(0.20)
Worker fixed effects		✓		✓		✓		✓
Manager characteristics		✓		✓		✓		✓
N	2912	2912	2912	2912	2059	2059	2059	2059

Notes: The table shows results from OLS regressions. The dependent variables are the recalled quiz score (Columns (1) and (2)), the probability of correct recall in Columns (3) and (4), "overestimation": an indicator that is 1 if the recall of a incorrectly recalled quiz score is higher than the true quiz score in Columns (5) and (6) and the absolute memory error in Columns (7) and (8). Independent variables are the indicator for PREF, and dummies for the worker's quiz scores, where a quiz score of 3 is the omitted category. Further, I include control variables in the even columns. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Columns (5) to (8) use the restricted sample of observations for which the recalled quiz score does not coincide with the actual quiz score. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.8: Regression results: Mediation analysis for quiz scores of 7

	(1) hire	(2) recalled quiz score	(3) hire
PREF	0.122***	-0.000	0.122***
	(0.02)	(0.02)	(0.02)
RECALL	-0.379***	-1.587***	-0.114***
	(0.03)	(0.09)	(0.03)
PREF*RECALL	0.177***	0.456***	0.101***
	(0.04)	(0.11)	(0.03)
Recalled quiz score			0.167***
			(0.01)
Constant	0.860***	6.980***	-0.305***
	(0.02)	(0.02)	(0.07)
N	1512	1512	1512

Notes: This table shows the OLS regression coefficients underlying Figure A.4. The coefficient on the interaction term in Column (1) corresponds to c, the coefficient on the interaction term in Column (2) corresponds to a, the interaction term in Column (3) corresponds to c' and the coefficient on the recalled quiz score corresponds to b. Note that I defined the recalled quiz score to be the repeated quiz score for participants in REMINDER. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.9: Regression results: Mediation analysis for quiz scores smaller than 7

	(1) hire	(2) recalled quiz score	(3) hire
PREF	0.091***	0.025*	0.086***
	(0.01)	(0.01)	(0.01)
RECALL	0.099***	0.411***	0.011
	(0.02)	(0.05)	(0.02)
PREF*RECALL	0.083***	0.089	0.064***
	(0.02)	(0.07)	(0.02)
Recalled quiz score			0.215***
			(0.01)
Constant	0.254***	4.332***	-0.677***
	(0.01)	(0.01)	(0.02)
N	4536	4536	4536

Notes: This table shows the OLS regression coefficients underlying Figure A.5. The coefficient on the interaction term in Column (1) corresponds to c, the coefficient on the interaction term in Column (2) corresponds to a, the interaction term in Column (3) corresponds to c' and the coefficient on the recalled quiz score corresponds to b. Note that I defined the recalled quiz score to be the repeated quiz score for participants in REMINDER. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.3 Classification of free-form answers

Following the procedure and recommendations by Celebi and Penczynski (2024), I use GPTs API to classify the answers to the free-form text that results from the self-stated hiring strategies according to whether they used gender or not. I used GPT-4o and implemented n-shot prompting (i.e., providing examples), a technique that has been shown to improve performance, and set the temperature to 0, as recommended. Figure A.6 shows the prompt in `Markdown` language. This procedure classified 225 strategies as *0* (i.e., gender did not influence the decision), 524 strategies as *1* (i.e., gender influenced the decision), and failed to classify 7 strategies ($< 1\%$). Additionally, a research assistant classified a random sample of 10% of observations. The human and GPT coded classification coincide in most of the cases (96.2%).

```
# General Task
- Evaluate a manager's explanation of their hiring decisions to determine whether the **gender of the worker** influenced the decision.

# Context
- This data comes from an online laboratory experiment where managers hire workers based on short profiles that contain:
    + a name
    + a picture
    + a quiz score (a signal of the worker's quality)
- The experiment has two main stages:
    + **Stage 1**: Managers view all worker profiles, including the quiz score.
    + **Stage 2**: Managers view the same profiles again and make hiring decisions.
        - **Important**: Depending on the treatment group, in Stage 2, quiz scores are **shown** again or **not shown**
        - If the quiz score is not shown, managers must **recall** them from memory.
- Payoffs for hiring decisions:
    + If a manager **does not hire** a worker: they receive **2 GBP with 50% probability**.
    + If a manager **does hire** a worker:
        - They receive **2 GBP with probability q × 10%**, where **q** is the quiz score.
        - Additionally, managers receive a **fixed bonus** based on the worker's gender.
        - Which gender corresponds to the bonus is randomized:
            + Either: If the worker is **female**, manager receives **0.50 GBP**
            + Or: If the worker is **male**, manager receives **0.50 GBP**
- After hiring, managers also rank the workers.
- At the end, each manager provides a **free-text explanation** of their hiring strategy.

# Classification Task
- Determine whether the manager’s explanation indicates **use of gender identity** in their decision-making.
- Label the explanation as:
    + `1` if **gender influenced** the decision.
    + `0` if **gender did not influence** the decision.

# Examples
1a. "quiz score" -> `0`
1b. "high quiz score and females" → `1`
1c. "high quiz score and males" → `1`
2. "My decision depended mainly on the quiz score but also on the gender." → `1`
3a. "I could not remember anything, so I ranked females higher" → `1`
3b. "I could not remember anything, so I ranked males higher" → `1`
3c. "I calculated the total expected payoff, using the quiz score and information on gender" → `1`
4a. "If two had the same quiz score, I favored female" -> `1`
4b. "If two had the same quiz score, I favored male" -> `1`
5. "How they looked like" → `0`
6. "Quiz score was important; I did not use the gender" → `0`
7. "The higher the score the more likely I hired" → `0`

# Constraint
- Refrain from providing an explanation for your classification.
- Provide a final and single classification for each manager.
- Follow the output format below:

# Output Format
Please respond with either 1 or 0, and nothing else
```

Figure A.6: Prompt for the classification of hiring strategies

A.4 Heterogeneity by manager characteristics

In this Appendix section, I assess whether treatment effects differ along any of the preregistered dimensions. A first set of heterogeneity dimensions concerns gender aspects: The gender of the manager, the gender of the worker and the degree of sexist attitudes (measured using short versions of the gender-specific ambivalent sexism inventories, see Glick and Fiske, 1996, 1999). A second set of heterogeneity dimensions are related to intrinsic costs and benefits of belief distortions and engagement in favoritism. These include attitudes toward equality and meritocracy (measured using the level of agreement on a five-point Likert scale towards the statements *"every person should be treated equally"* and *"workers should be rewarded based on their performance"*), preferences for truth-telling (implicit lying costs, measured with the survey module by Schudy et al., 2024), proneness to feelings of guilt (Cohen et al., 2014), and two dimensions of moral self-image (symbolism and internalization, see Aquino and Reed II, 2002).

For the recalled quiz score as outcome variable, I estimate the following model with OLS.

$$\texttt{y}_{ij} = \beta_0 + \beta_1 \texttt{pref}_{ij} + \beta_2 \texttt{abovemedian}_i + \beta_3 \texttt{abovemedian}_i * \texttt{pref}_{ij} + \varepsilon_{ij}, \tag{9}$$

where $\texttt{y}_{ij}$ is the recalled quiz score of manager i for worker j, $\texttt{pref}_{ij}$ an indicator that is one when worker j belongs to manager i's preferred group, and $\texttt{abovemedian}_i$ a dummy variable that indicates whether the manager scores above median in the respective heterogeneity dimension. In the case where the dimension of heterogeneity is measured as a binary variable, `abovemedian` is replaced with this binary variable. Standard errors are clustered on the manager level. The left panel of Figure A.7 shows the coefficients of interest (β_3) with its confidence intervals for each of the dimensions of heterogeneity. The results show little systematic heterogeneity. In particular, the difference in recalled quiz scores between PREF and NOPREF workers is not significantly effected by either of the gender-related dimensions. The only dimension of heterogeneity that impacts the treatment effect significantly with a p-value below 0.05 is the preference for equality: Managers who strongly agree to the statement *"Every person should be treated equally"* (71% of participants strongly agree to the statement) distort their memory in favor of PREF-workers less compared to less equality concerned managers.

To assess heterogeneous treatment effects on the hiring decision, I estimate the following model with OLS.

$$\begin{aligned} \texttt{y}_{ij} =& \beta_0 + \beta_1 \texttt{pref}_{ij} + \beta_2 \texttt{recall}_i + \beta_3 \texttt{pref}_{ij} * \texttt{recall}_i + \beta_4 \texttt{abovemedian}_i \\ &+ \beta_5 \texttt{abovemedian}_i * \texttt{pref}_{ij} + \beta_6 \texttt{abovemedian}_i * \texttt{recall}_i \\ &+ \beta_7 \texttt{abovemedian}_i * \texttt{pref}_{ij} * \texttt{recall}_i + \varepsilon_{ij}, \end{aligned} \tag{10}$$

where $\texttt{y}_{ij}$ is the binary hiring indicator, $\texttt{pref}_{ij}$ an indicator that is one when worker j belongs to manager i's preferred group, and $\texttt{recall}_i$ an indicator that is one when manager i is in

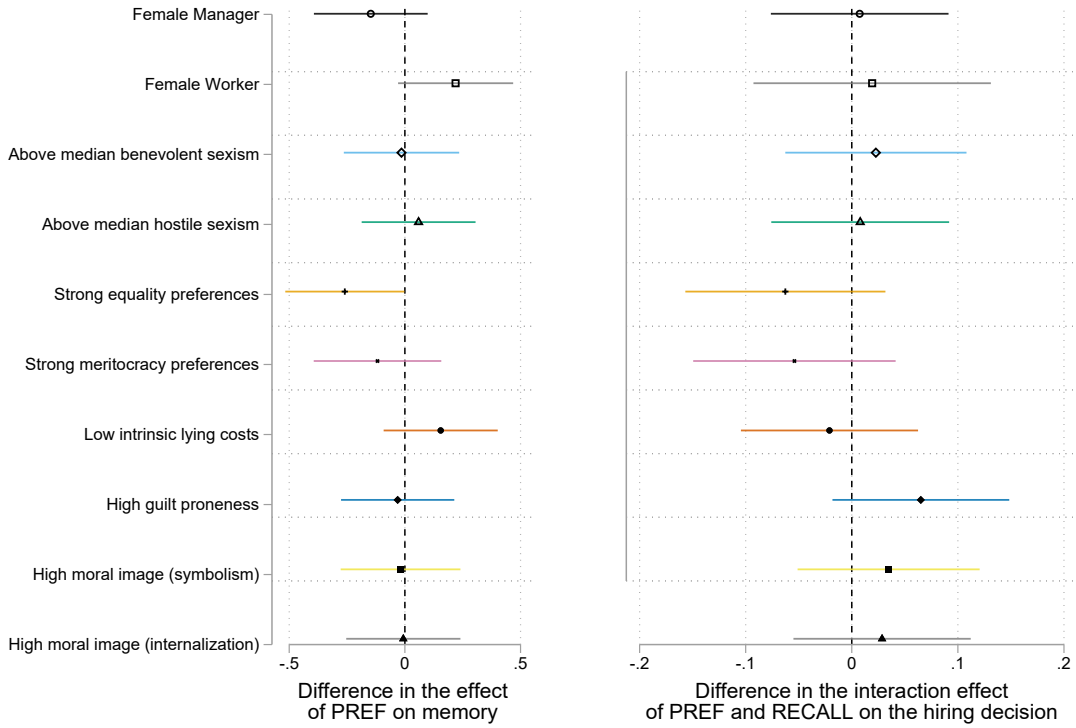


Figure A.7: Heterogeneity

Notes: The left panels shows coefficients β_3 of Regression (9) using the data from the RECALL treatment, and the right panel shows coefficients β_7 from Regression (10). The colored bars show 95% Confidence intervals

the RECALL condition. Again, $\texttt{abovemedian}_i$ is a dummy variable that indicates whether the manager scores above median in the respective heterogeneity dimension. In the case where the dimension of heterogeneity is measured as a binary variable, `abovemedian` is replaced by this binary variable. Standard errors are clustered on the manager level. The right panel of Figure A.7 shows the coefficients of interests (β_7) and it's confidence intervals for each of the dimensions of heterogeneity. I cannot detect heterogeneous treatment effects across any of the dimensions.[1]

A.5 Robustness checks

Stricter and more lenient exclusion criteria The main results are robust when including participants that used memory aids, see Tables A.10 and A.11. They are also robust to the exclusion of participants who guessed the research question sufficiently close, see Tables A.12 and A.13. A research assistant coded whether participants correctly guessed the research question. An answer was classified as "sufficiently close" if it mentioned both memory and favoritism (e.g. "discrimination" or "gender differences").

[1]Note that the absence of statistically significant heterogeneity should not be interpreted as evidence of no heterogeneity, as the study is not sufficiently powered to detect subtle differences across subgroups.

Table A.10: Robustness of memory results to including participants who took notes

	recalled quiz score		correct recall		overestimation		abs. memory error	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PPREF	0.202***	0.202***	0.016	0.017	0.047**	0.026	-0.096**	-0.093**
	(0.06)	(0.06)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)
Constant	4.915***	4.279***	0.305***	0.181*	0.477***	0.636***	2.125***	1.778***
	(0.05)	(0.24)	(0.01)	(0.10)	(0.01)	(0.06)	(0.03)	(0.19)
Manager characteristics		✓		✓		✓		✓
Worker fixed effects		✓		✓		✓		✓
N	3048	3048	3048	3048	2093	2093	2093	2093

Notes: The table shows results from OLS regressions. The dependent variables are the recalled quiz score in Columns (1) and (2), the probability of correct recall in Columns (3) and (4), "overestimation": an indicator that is 1 if the recall of a wrongly recalled quiz score is higher than the true quiz score in Columns (5) and (6) and the absolute memory error for incorrect recalls in Columns (7) and (8). Independent variables are the indicator for PREF, as well as further control variables in the even columns. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). The sample includes participants who stated to have taken notes in the screening stage. These participants are not included in the main analyses. Columns (5) to (8) use the restricted sample of observations for which the recalled quiz score does not coincide with the actual quiz score. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.11: Robustness of memory-induced favoritism to including participants who took notes

	hire		recommendation		ranking	
	(1)	(2)	(3)	(4)	(5)	(6)
PREF	0.099***	0.099***	4.112***	4.125***	-0.864***	-0.865***
	(0.01)	(0.01)	(0.44)	(0.44)	(0.05)	(0.05)
RECALL	-0.025	-0.023	-2.011***	-1.928**	0.021	0.017
	(0.02)	(0.02)	(0.76)	(0.77)	(0.06)	(0.06)
PREF x RECALL	0.105***	0.103***	3.014***	2.932***	-0.042	-0.034
	(0.02)	(0.02)	(0.90)	(0.90)	(0.11)	(0.11)
Constant	0.406***	0.134***	48.217***	30.884***	4.932***	5.782***
	(0.01)	(0.05)	(0.37)	(2.29)	(0.03)	(0.06)
Worker fixed effects		✓		✓		✓
Manager characteristics		✓		✓		✓
N	6184	6184	6184	6184	6184	6184

Notes: The table shows results from OLS regressions. The dependent variable is the hiring probability in Columns (1) and (2), the recommendation in Columns (3) and (4) an the rank in Columns (5) and (6). Independent variables are the indicator for PREF, the indicator that is 1 for managers in RECALL as well as the interaction. Further control variables in the even columns include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). The sample includes participants who stated to have taken notes in the screening stage. These participants are not included in the main analyses. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.12: Robustness of memory results to including participants who guessed the research question

	recalled quiz score		correct recall		overestimation		abs. memory error	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PREF	0.195***	0.194***	0.017	0.017	0.049**	0.031*	-0.095**	-0.093**
	(0.06)	(0.06)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
Constant	4.900***	4.371***	0.285***	0.143*	0.473***	0.654***	2.122***	1.776***
	(0.05)	(0.26)	(0.01)	(0.08)	(0.01)	(0.06)	(0.04)	(0.19)
Manager characteristics		✓		✓		✓		✓
Worker fixed effects		✓		✓		✓		✓
N	2808	2808	2808	2808	1984	1984	1984	1984

Notes: The table shows results from OLS regressions. The dependent variables are the recalled quiz score in Columns (1) and (2), the probability of correct recall in Columns (3) and (4), "overestimation": an indicator that is 1 if the recall of a wrongly recalled quiz score is higher than the true quiz score in Columns (5) and (6) and the absolute memory error for incorrect recalls in Columns (7) and (8). Independent variables are the indicator for PREF, as well as further control variables in the even columns. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). The sample excludes participants who guessed the research question sufficiently close. These participants are included in the main analyses. Columns (5) to (8) use the restricted sample of observations for which the recalled quiz score does not coincide with the actual quiz score. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.13: Robustness of memory-induced favoritism to including participants who guessed the research question

	hire		recommendation		ranking	
	(1)	(2)	(3)	(4)	(5)	(6)
PREF	0.099***	0.099***	4.133***	4.154***	-0.863***	-0.865***
	(0.01)	(0.01)	(0.44)	(0.44)	(0.05)	(0.05)
RECALL	-0.021	-0.020	-2.112***	-2.039**	0.012	0.008
	(0.02)	(0.02)	(0.80)	(0.80)	(0.06)	(0.06)
PREF x RECALL	0.107***	0.105***	3.011***	2.933***	-0.025	-0.017
	(0.02)	(0.02)	(0.94)	(0.94)	(0.12)	(0.12)
Constant	0.406***	0.132***	48.199***	31.359***	4.931***	5.771***
	(0.01)	(0.05)	(0.37)	(2.36)	(0.03)	(0.07)
Worker fixed effects		✓		✓		✓
Manager characteristics		✓		✓		✓
N	5928	5928	5928	5928	5928	5928

Notes: The table shows results from OLS regressions. The dependent variable is the hiring probability in Columns (1) and (2), the recommendation in Columns (3) and (4) an the rank in Columns (5) and (6). Independent variables are the indicator for PREF, the indicator that is 1 for managers in RECALL as well as the interaction. Further control variables in the even columns include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). The sample excludes participants who guessed the research question sufficiently close. These participants are included in the main analyses. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Finally, the main results are robust to excluding participants who show irregularities in their hiring decisions. One way to do so is to exclude managers who show inconsistencies in their hiring decisions. I classify a manager as inconsistent if their hiring decisions of each group of workers as a function of their recalled or repeated quiz score is not weekly increasing, see Tables A.14 and A.15. A second way would be to exclude participants who do not hire workers with (recalled) quiz score above 5, which implies choosing a first order stochastically dominated lottery, see Tables A.16 and A.17. The point estimates of the main treatment effects are even somewhat larger in these reduced samples, indicating that the treatment differences are explicitly not driven by these inconsistencies that may be the cause of confusion nor by participants who may have misunderstood incentives and made stochastically dominated choices.

Table A.14: Robustness of memory results to excluding inconsistent participants

	recalled quiz score		correct recall		overestimation		abs. memory error	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PREF	0.260***	0.258***	0.009	0.009	0.057***	0.031*	-0.115**	-0.115***
	(0.07)	(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
Constant	4.861***	4.303***	0.291***	0.143	0.464***	0.597***	2.121***	1.836***
	(0.05)	(0.27)	(0.01)	(0.09)	(0.01)	(0.07)	(0.04)	(0.21)
Manager characteristics		✓		✓		✓		✓
Worker fixed effects		✓		✓		✓		✓
N	2512	2512	2512	2512	1769	1769	1769	1769

Notes: The table shows results from OLS regressions. The dependent variables are the recalled quiz score in Columns (1) and (2), the probability of correct recall in Columns (3) and (4), "overestimation": an indicator that is 1 if the recall of a wrongly recalled quiz score is higher than the true quiz score in Columns (5) and (6) and the absolute memory error for incorrect recalls in Columns (7) and (8). Independent variables are the indicator for PREF, as well as further control variables in the even columns. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). The sample excludes participants who show inconsistencies in the sense that their hiring decisions within each group of workers as a function of their recalled or repeated quiz score is not weekly increasing. These participants are included in the main analyses. Columns (5) to (8) use the restricted sample of observations for which the recalled quiz score does not coincide with the actual quiz score. Standard errors clustered at the individual level are reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.15: Robustness of memory-induced favoritism to excluding inconsistent participants

	hire		recommendation		ranking	
	(1)	(2)	(3)	(4)	(5)	(6)
PREF	0.099***	0.099***	4.125***	4.138***	-0.888***	-0.889***
	(0.01)	(0.01)	(0.45)	(0.45)	(0.05)	(0.05)
RECALL	-0.029	-0.027	-2.455***	-2.489***	0.017	0.014
	(0.02)	(0.02)	(0.83)	(0.84)	(0.06)	(0.06)
PREF x RECALL	0.116***	0.115***	2.790***	2.736***	-0.034	-0.027
	(0.02)	(0.02)	(0.98)	(0.98)	(0.12)	(0.12)
Constant	0.404***	0.127**	48.174***	33.570***	4.944***	5.805***
	(0.01)	(0.05)	(0.37)	(2.36)	(0.03)	(0.07)
Worker fixed effects		✓		✓		✓
Manager characteristics		✓		✓		✓
N	5504	5504	5504	5504	5504	5504

Notes: The table shows results from OLS regressions. The dependent variable is the hiring probability in Columns (1) and (2), the recommendation in Columns (3) and (4) an the rank in Columns (5) and (6). Independent variables are the indicator for PREF, the indicator that is 1 for managers in RECALL as well as the interaction. Further control variables in the even columns include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). The sample excludes participants who show inconsistencies in the sense that their hiring decisions within each group of workers as a function of their recalled or repeated quiz score is not weekly increasing. These participants are included in the main analyses. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.16: Robustness of memory results to excluding participants who do not hire well performing workers

	recalled quiz score		correct recall		overestimation		abs. memory error	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PREF	0.342***	0.340***	0.014	0.014	0.085***	0.061***	-0.121**	-0.118**
	(0.08)	(0.08)	(0.02)	(0.02)	(0.02)	(0.02)	(0.05)	(0.05)
Constant	4.754***	4.055***	0.273***	0.096	0.433***	0.523***	2.147***	1.786***
	(0.06)	(0.32)	(0.02)	(0.10)	(0.02)	(0.08)	(0.04)	(0.24)
Manager characteristics		✓		✓		✓		✓
Worker fixed effects		✓		✓		✓		✓
N	1896	1896	1896	1896	1365	1365	1365	1365

Notes: The table shows results from OLS regressions. The dependent variables are the recalled quiz score in Columns (1) and (2), the probability of correct recall in Columns (3) and (4), "overestimation": an indicator that is 1 if the recall of a wrongly recalled quiz score is higher than the true quiz score in Columns (5) and (6) and the absolute memory error for incorrect recalls in Columns (7) and (8). Independent variables are the indicator for PREF, as well as further control variables in the even columns. These include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). The sample excludes participants who do not hire workers with (recalled) quiz score above 5, which implies choosing a first order stochastic dominated lottery. These participants are included in the main analyses. Columns (5) to (8) use the restricted sample of observations for which the recalled quiz score does not coincide with the actual quiz score. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.17: Robustness of memory-induced favoritism to excluding participants who do not hire well performing workers

	hire		recommendation		ranking	
	(1)	(2)	(3)	(4)	(5)	(6)
PREF	0.049***	0.048***	3.636***	3.604***	-0.837***	-0.834***
	(0.01)	(0.01)	(0.44)	(0.43)	(0.06)	(0.06)
RECALL	-0.050***	-0.048**	-2.793***	-2.733***	0.071	0.069
	(0.02)	(0.02)	(0.96)	(0.97)	(0.07)	(0.07)
PREF x RECALL	0.158***	0.157***	3.822***	3.790***	-0.142	-0.138
	(0.03)	(0.03)	(1.12)	(1.11)	(0.14)	(0.14)
Constant	0.515***	0.195***	49.165***	32.109***	4.919***	5.747***
	(0.00)	(0.05)	(0.40)	(3.00)	(0.03)	(0.08)
Worker fixed effects		✓		✓		✓
Manager characteristics		✓		✓		✓
N	3984	3984	3984	3984	3984	3984

Notes: The table shows results from OLS regressions. The dependent variable is the hiring probability in Columns (1) and (2), the recommendation in Columns (3) and (4) an the rank in Columns (5) and (6). Independent variables are the indicator for PREF, the indicator that is 1 for managers in RECALL as well as the interaction. Further control variables in the even columns include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). The sample excludes participants who do not hire workers with (recalled) quiz score above 5, which implies choosing a first order stochastic dominated lottery These participants are included in the main analyses. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Lasso to select control variables In the main analysis, I control for worker fixed effects, and basic manager characteristics, following my pre-registration. I however elicit further variables related to manager's characteristics, that may correlate with the outcomes and thereby constitute adequate control variables. Therefore, I use post-double selection Lasso (Belloni et al., 2014) to select control variables for the between-subjects comparisons.[2] Table A.18 reports the results. As is common for randomized experiments (Cilliers et al., 2024), the Lasso procedure selects only few variables. In particular, the quiz score is selected for most outcome variables, but none of the manager characteristics are selected by the lasso.

[2]I include the following potential control variables: (i) worker's characteristics: gender, quiz score, an indicator for each worker, (ii) dummies for each of the 8 possible position a manager is confronted with a specific worker in the hiring stage, (iii) basic manager's characteristics: gender, age, university, risk preferences, and patience, (iv) further manager characteristics: Lying preferences, benevolent and hostile sexism, guilt proneness, moral image symbolization and internalization dimensions, preferences for equality and meritocracy, self-view as a discriminator, importance of not being seen as a discriminator, and number of tasks solved in the distraction stage, as well as (v) manager's data as per download from prolific (total approvals, university degree, ethnicity, nationality, employment type, whether English is the first language, and student status). I use the Stata command `pdslasso` with its default parameters and methods for estimation.

Table A.18: Robustness: Lasso to select control variables

	(1) recalled quiz score	(2) correct recall	(3) overestimate	(4) abs. memory error	(5) hire	(6) recommendation	(7) ranking
PREF	0.195***	0.014	0.026	-0.106**	0.099***	4.116***	-0.866***
	(0.06)	(0.02)	(0.02)	(0.04)	(0.01)	(0.44)	(0.05)
RECALL					-0.021	-1.938**	0.007
					(0.02)	(0.78)	(0.06)
PREF*RECALL					0.106***	2.840***	-0.015
					(0.02)	(0.92)	(0.11)
quiz score	0.215***	0.023***	-0.193***		0.146***	7.379***	-0.687***
	(0.02)	(0.01)	(0.01)		(0.01)	(0.29)	(0.03)
male worker, q3			0.104***	0.317***			0.236***
			(0.03)	(0.08)			(0.08)
female worker, q3			0.119***				
			(0.03)				
female worker, q4				-0.340***	-0.054***		
				(0.06)	(0.01)		
male worker, q6				-0.333***			
				(0.06)			
male worker, q6						3.475***	-0.520***
						(0.79)	(0.08)
female worker, q7	0.230**				0.035**	4.921***	-0.619***
	(0.10)				(0.02)	(0.81)	(0.08)
Position 7/8				0.302***			0.236***
				(0.07)			(0.08)
Constant	3.803***	0.173***	1.404***	2.213***	-0.321***	10.305***	8.480***
	(0.12)	(0.03)	(0.05)	(0.04)	(0.02)	(1.42)	(0.14)
N	2904	2904	2053	2053	6016	6016	6016

Notes: The table shows results from lasso estimation with the Stata command `pdslasso` and it's default parameters and settings. The control variables considered are: (i) worker's characteristics: gender, quiz score, an indicator for each worker, (ii) dummies for each of the 8 possible position a manager is confronted with a specific worker in the hiring stage, (iii) basic manager's characteristics: gender, age, university, risk preferences, and patience, (iv) further manager characteristics: Lying preferences, benevolent and hostile sexism, guilt proneness, moral image symbolization and internalization dimensions, preferences for equality and meritocracy, self-view as a discriminator, importance of not being seen as a discriminator, and number of tasks solved in the distraction stage, as well as (v) manager's data as per download from prolific (total approvals, university degree, ethnicity, nationality, employment type, whether English is the first language, and student status). The dependent variables are the recalled quiz score in Column (1), the probability of correct recall in Column (2), "overestimation": an indicator that is 1 if the recall of a wrongly recalled quiz score is higher than the true quiz score in Column (3), the absolute memory error for incorrect recalls in Column (4), the hiring probability in Column (5), the recommendation in Column (6), and the ranking in Column (7). Columns (1) to (4) use the sample of the RECALL treatment only. Columns (3) and (4) further restrict the sample to observations for which the recalled quiz score does not coincide with the actual quiz score. Note that the slightly reduced sample size compared to Tables 1 and 2 is due to three subjects that identified as non-binary/third gender or did not want to reveal their gender. These participants did not see the gender-specific sexism questionnaire, thus not all manager attitudes are defined. Standard errors clustered at the individual level are reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Logit regression for binary outcome variables. For the binary outcome variables (hiring, correct recall, and overestimation), the results are qualitatively similar when using logit regression instead of a linear probability model, see Table A.19.

Table A.19: Robustness: Logit for binary outcome variables

	correctrecall		overestimate		hire	
	(1)	(2)	(3)	(4)	(5)	(6)
PREF	0.0696	0.0720	0.188*	0.185	0.400***	0.542***
	(0.95)	(0.97)	(2.40)	(1.42)	(9.70)	(9.93)
RECALL					-0.0850	-0.104
					(-1.12)	(-1.02)
PREF*RECALL					0.434***	0.572***
					(4.89)	(4.76)
Constant	-0.916***	-1.620***	-0.100	0.655	-0.382***	-1.862***
	(-14.09)	(-4.11)	(-1.83)	(1.53)	(-9.32)	(-6.84)
Manager characteristics		✓		✓		✓
Worker fixed effects		✓		✓		✓
N	2912	2912	2059	2059	6048	6048

Notes: The table shows results from logit regressions. The dependent variables are the correct recall in Columns (1) and (2), "overestimation": an indicator that is 1 if the recall of a wrongly recalled quiz score is higher than the true quiz score in Columns (3) and (4) an the hiring decision in Columns (5) and (6). Independent variables are the indicator for PREF, the indicator that is 1 for managers in RECALL as well as the interaction. Further control variables in the even columns include worker fixed effects and manager's characteristics (gender, age, a dummy that is one if the individual visited university, risk preferences and patience). Columns (1) and (2) use the data of the RECALL treatment only. Columns (3) and (4) further excludes observations for which the recalled quiz score coincides with the actual quiz score. Columns (5) and (6) use the full sample of the RECALL and REMINDER treatment. Standard errors clustered at the individual level are reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Multiple hypothesis testing. Seven main and secondary outcome variables are part of the main analysis. The memory related outcomes include the recalled quiz score, whether the recall is incorrect, whether an incorrect recall is an overestimation, and the absolute memory error. Additionally, I test treatment differences in three personnel decisions: hiring, recommendation and ranking. To account for multiple-hypothesis testing, I calculate Westfall-Young step-down adjusted p-values, which control the family-wise error rate, treating the seven tests as one family. The effect of PREF on the recalled quiz score is still significant on the 5%-level ($pwyoung = 0.02$), the effect of PREF on the probability to overestimate and the absolute memory error are now not significant on the 5%-level anymore, but remain marginally significant ($pwyoung = 0.07$, each). Regarding the treatment effect of RECALL on favoritism, the coefficients on the interaction term remain significant on the 5%-level (hiring: $pwyoung = 0.01$ and recommendation: $pwyoung = 0.03$).

Active memory and choice vs. noise Remembering the quiz scores of eight workers is quite a difficult task. 38% of participants found it "very difficult" to remember the worker's quiz scores. Therefore, one may fear that participants did not try to actively remember the quiz scores but instead report noise. However, in an exercise inspired by Saucet and Villeval (2019), I can show that memory is better than random noise would be. To do so, I compare i) the recall quality and ii) the magnitude of memory errors in the experimental data with several simulated distributions.[3] As reported in Table A.20, the percentage of correct recalls is significantly higher in the experimental data compared to any of the simulated distributions. Furthermore, the absolute memory error is significantly lower in the experimental data compared to any of the simulated distributions.

Table A.20: Comparison with data from simulated distributions

	Probability of correct recall			Absolute memory error		
	Simulated data	Actual data	P-value (t-test)	Simulated data	Actual data	P-value (t-test)
	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)
Distribution a)	0.205	0.293	<0.001	1.70	1.47	<0.001
	(0.007)	(0.008)		(0.024)	(0.025)	
Distribution b)	0.251	0.293	<0.001	1.72	1.47	<0.001
	(0.008)	(0.008)		(0.025)	(0.025)	
Distribution c)	0.153	0.293	<0.001	1.57	1.47	=0.001
	(0.007)	(0.008)		(0.025)	(0.025)	
Distribution d)	0.153	0.293	<0.001	2.03	1.47	<0.001
	(0.007)	(0.008)		(0.028)	(0.025)	
Distribution e)	0.240	0.293	<0.001	1.60	1.47	<0.001
	(0.008)	(0.008)		(0.023)	(0.025)	
Distribution f)	0.198	0.293	<0.001	2.04	1.47	<0.001
	(0.007)	(0.008)		(0.029)	(0.025)	

Notes: In this table, I compare memory quality from simulated data to actual data. To assess memory quality, I use the probability of correct recall and the absolute memory error. The p-values are the result of two-sided t-tests. The simulated distributions are: a) a uniform distribution on {3,4,5,6,7}, a uniform distribution on {3,4,6,7}, c) a normal distribution with mean 5 and standard deviation 1, d) a normal distribution with mean 5 and standard deviation 2, e) in a first step a normal distribution with mean 5 and standard deviation 1, replacing values of 5 with randomly either 4, or 6, f) as e), but with standard deviation 2. All normal distributions are winzorized to [0,10].

Balance I test for joint orthogonality by estimating the following model:

$$\texttt{Treat}_i = \alpha + \beta X_i' + \varepsilon_i, \tag{11}$$

[3] In particular, these distributions are: a) a uniform distribution on {3,4,5,6,7}, a uniform distribution on {3,4,6,7}, c) a normal distribution with mean 5 and standard deviation 1, d) a normal distribution with mean 5 and standard deviation 2, e) a normal distribution with mean 5 and standard deviation 1, replacing values of 5 with randomly either 4, or 6, f) as e), but with standard deviation 2. All normal distributions are winzorized to [0,10].

where $\texttt{Treat}_i$ is i) a dummy that is 1 for participants in RECALL, and ii) a dummy that is 1 for participants for whom the PREF group corresponds to female workers and iii) a dummy that is 1 for participants for whom the memory elicitation comes prior to the hiring decision. X' is a vector of control variables. This includes data available on Prolific: Total approvals, university degree, ethnicity, nationality, employment type, whether English is the first language, and student status. Furthermore, I include the following pre-treatment variables from my survey: Gender and age. For outcomes i) and iii), I also include the number of tasks solved in the real-effort task and the time spent screening the workers. I do not use further manager characteristics, as these are elicited post treatment assignment. Following Kerwin et al. (2024), I estimate a logit model and perform an F-test of joint orthogonality using randomization inference p-values. Regarding the gender corresponding to the PREF-group and the order of memory and hiring decisions, randomization was successful ($p = 0.580$ and $p = 0.273$, respectively). Conversely, the memory randomization of RECALL is not sufficiently randomized ($p = 0.044$). Nonetheless, a balance table (Table A.21) indicates that none of the variables is 'unbalanced' across treatments. Therefore, following my pre-registration, I do not control for more characteristics in my analysis.

Table A.21: Balance table

	(1) REMINDER		(2) RECALL		(1)-(2) Pairwise t-test	
Variable	N	Mean/(SE)	N	Mean/(SE)	N	P-value
Total approvals	392	905.015 (54.200)	364	888.948 (56.482)	756	0.837
Bachelor degree	392	0.439 (0.025)	364	0.451 (0.026)	756	0.745
Master degree	392	0.342 (0.024)	364	0.291 (0.024)	756	0.135
White	392	0.768 (0.021)	364	0.753 (0.023)	756	0.627
British	392	0.781 (0.021)	364	0.780 (0.022)	756	0.990
Irish	392	0.038 (0.010)	364	0.025 (0.008)	756	0.289
Full time employed	392	0.776 (0.021)	364	0.731 (0.023)	756	0.154
English first language	392	0.906 (0.015)	364	0.901 (0.016)	756	0.834
Student	392	0.105 (0.015)	364	0.102 (0.016)	756	0.894
Female	392	0.500 (0.025)	364	0.492 (0.026)	756	0.821
Age	392	40.709 (0.550)	364	40.824 (0.551)	756	0.883
# tasks solved	392	6.143 (0.125)	364	6.118 (0.114)	756	0.884
Screening time	392	141.905 (5.505)	364	144.226 (9.089)	756	0.824

B Deviations from the pre-registration

The order in which the results are presented in the paper does not necessarily match the order of analyses in the pre-registration. I deviated from the pre-registration in one instance for the following reason:

- The original Hypothesis 2 was formulated as *"Motivated memories of preferred workers can foster discriminatory hiring decisions"*. The preregistered main analysis (comparison of favoritism between RECALL and REMINDER) however does not only capture the effect of motivated memories but also the effect of reliance on memory per se. Therefore, I decided to reformulate the hypothesis to *"Favoritism (i.e., the hiring gap between PREF- and NOPREF-workers) is higher in the RECALL treatment than in the REMINDER treatment"* and investigate the original Hypothesis 2 in the exploratory mediation analysis (Section 4.3).

Furthermore, I report a number of additional analyses and robustness checks that were not included in the pre-registration. These include:

- The individual level analysis at the end of Section 4.2
- The mediation analysis (Section 4.3)
- Results on awareness of 'memory-induced favoritism' (Section 5.1)
- The analysis of order effects to explore cognitive flexibility and ex-post rationalization (Section 5.4)
- Heterogeneity by workers' quiz score (Section 5.6, the preregistered heterogeneity analysis can be found in Appendix A.4)
- A subset of the robustness checks (Section A.5)

C Auxiliary experiments

To select the workers for my main experiment, I followed a pre-registered procedure (`https://aspredicted.org/2547-g7sh.pdf`) that includes two auxiliary experiments. I collected data in July 2024. The next three subsections describe these pre-experiments as well as intermediate steps in more detail. Section C.4 describes a third auxiliary experiment, run shortly after the main experiment in November 2025.

C.1 Auxiliary experiment 1: Workers

50 Prolific participants act as workers who solve quizzes. Each quiz is on a different topic: Science/Technology, Language/Psychology, and History/Politics. Heursen et al. (2025) invented and thoroughly tested these questions and kindly allowed me to borrow them. In each of the quizzes, there are 10 questions which are shown for 15 seconds each.[4] These quizzes have the beneficial feature of combining each question with a picture. This impedes a quick internet search. For each question there are 5 possible answers, of which one is correct. Workers earn a piece rate for correct answers. After completing all quizzes, workers respond to a short questionnaire.

The requirements for workers were similar to those of the managers, with the difference that they were not required to have managerial experience but needed to be between 25 and 35 years old. On average, participants got 17 out of 30 answers correct and there is no significant gender difference ($p = 0.531$, two-sided t-test).

C.2 Prompts to generate the workers' pictures

For each of the 50 workers who completed the pre-experiment, I created a unique picture using the free tool `https://perchance.org/ai-photo-generator` to generate portraits. These portraits match the real worker's features on several physical characteristics. I used the following prompts to create the pictures.

- For females: Portrait of a woman, aged [25-30 / 30-35], with [brown / blue / green / gray] eyes, [black / dark brown / light brown / dark red / light red / ash blond / medium bald / blond / light blond / white / gray / colorful dyed] [curly / wavy / straight hair] ([pixie cut / bob / shoulder length / long]), skin type [1 / 2 / 3 / 4 / 5 / 6]. The woman is dressed in business casual attire and has a friendly expression. The portrait should be suitable for a CV, showing a frontal view with the body visible up to the shoulders, and a blurred background.

[4] In the original version, questions were shown for 30 seconds. Due to the advancements in generative AI, I reduced the time to 15 seconds after testing with research assistants.

- For males: Portrait of a man, aged [25-30 / 30-35], with [brown / blue / green / gray] eyes, [black / dark brown / light brown / dark red / light red / ash blond / medium bald / blond / light blond / white / gray / colorful dyed] [curly / wavy / straight hair] ([buzz cut / shorter hair / slightly longer hair / shoulder length or longer]), skin type [1 / 2 / 3 / 4 / 5 / 6]. The man is dressed in business casual attire and has a friendly expression. The portrait should be suitable for a CV, showing a frontal view with the body visible up to the shoulders, and a blurred background.

Further, I selected the style "cinematic" and shape "portrait".

C.3 Auxiliary experiment 2: Selection of workers

50 Prolific participants residing in the UK or Ireland took part in a second pre-experiment. They acted as evaluators who tested the perception of names and the AI-generated portraits. Evaluators engaged in two tasks that are presented in a randomized order. In one of the tasks, evaluators report their perception of a list of common names in the UK. In the other task, evaluators report their perception of the AI-generated portraits. The aim of pre-study 2 is to exclude names and pictures that provoke atypical perceptions.

I shortlisted 16 unique names out of the most common surnames in the UK as well as the most common female and male baby names from 1994, i.e., for the cohort that was 30 at the time of data collection. For these names, evaluators had to assess whether they believe a person with this name is UK or Irish citizen. They also had to indicate to which income group they believe a person with this name belongs and what is the highest educational degree a person with this name has reached. I chose the 4 female and 4 male names that most evaluators perceived to be UK or Irish citizens, earning middle income and having A-levels, bachelor's or a master's degree.

In the picture-perception task, each evaluators rated a subset of the 50 portraits based on memorability, competence, intelligence, attractiveness and likability of the person in the picture. For each of the characteristics (memorability, competence, intelligence, attractiveness and likability), I count a point if an evaluator rates it within the second and the third quartile of the overall rating of this characteristic. I then excluded the pictures which points lie in the lowest quartile of points. Thus, I excluded the pictures that are rated most extreme along these characteristics.

In the final step, I selected a sample of 4 female and 4 male workers where each of the quiz scores 3,4,6, and 7 occurs once for each gender. In my sample of workers, this was only possible for the language and psychology quiz.[5] I then matched the 8 selected names with the 8 selected

[5]The quiz questions were chosen such that females and males perform similarly well (see Heursen et al., 2025). In my sample of 50 workers, the average performance in the language/psychology quiz is around 5 out of 10 and – in line with the stereotype – females performed better (on average 5.4) than men (on average 4.4).

workers. For each of these selected workers I created a mini-CV consisting of (i) the name (ii) the portrait and (iii) quiz score in the language and psychology quiz.

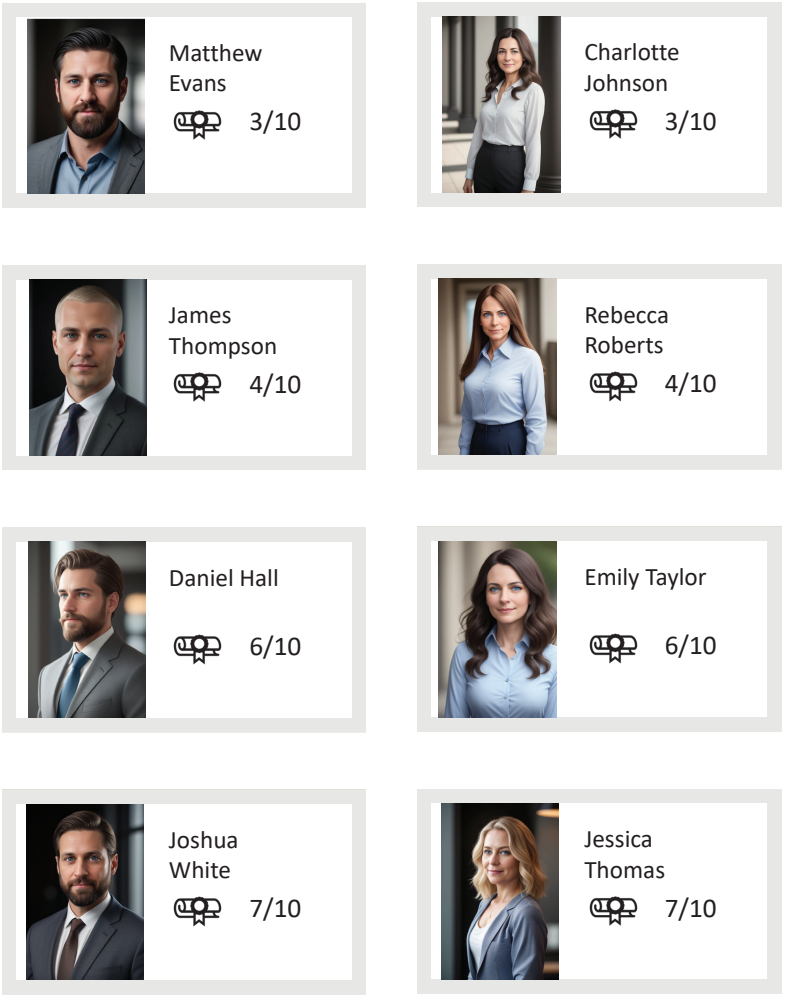


Figure C.1: CVs of the Selected Workers

C.4 Auxiliary experiment 3: The HR experiment

With the purpose to provide non-deceptive consequences for the recommendation decisions, I invited 6 Prolific participants that fulfill the same requirements as the managers acted as head of HR to an additional auxiliary data collection. In this auxiliary experiment, participants faced the same incentives as managers in the main experiment. However, instead of the brief CVs, they saw a histogram with the recommendations of all managers who faced exactly the same incentives (i.e., if females constitute the PREF-group for head of HR, they saw the recommendations by managers for whom females constitute the PREF-group). Compared to the main experiment, the structure was simplified such that the hiring decision follows directly after the instructions.

Also, participants only made the eight hiring decisions and faced no additional decisions. I do not analyze this data.

D Experimental material

D.1 Instructions

The instructions shown in this Appendix correspond to the RECALL treatment where PREF corresponds to male workers and the hiring decision comes prior to the recall task. For the screening (Part 1) and the hiring stage (Part 3), these instructions contain the pages related to one of the eight workers only. The remaining pages follow the same pattern. For the distraction stage (Part 2), one matrix is shown only. Depending on the speed, a participant would see several matrices (the median number of matrices solved is 6, the maximum 14). I only include selected questions of the final questionnaire (Part 4). Complete instructions are available upon request.

Part 1

In this experiment, you will take on the role of a **manager.** You will eventually hire workers to perform a task. In the following, you will learn about the workers, the task and your bonus payment. Please read the instructions carefully.

Please click "next" to see the instructions of part 1.

next

The Workers

The workers are **real** participants who took part in a previous experiment. In that experiment, the workers solved a quiz. They answered 10 questions and received a monetary bonus for each correct answer. Thus, they had an incentive to perform as well as possible.

The Questions

There were five available answer options for each question. One of these options was the correct answer. The workers had only 15 seconds to answer each question. Due to the format of the questions and the time limit, workers could not easily search for answers on the internet.

The Quiz Score

Depending on how well each worker did on the questions, they received a lower or higher quiz score. **The maximum attainable quiz score is 10, while the minimum score is 0.**

Worker CVs

In the following, you will see worker's CVs. These show:

- A name

- A picture
- The **quiz score**

The pictures and names enable you to distinguish between the workers during the experiment. Due to data protection policies, the names are modified and the pictures are AI-generated. The pictures match the real workers based on the following characteristics:

- Gender
- Eye color
- Skin color
- Hair color
- Hair style
- Approximate age

Here is an example CV:



Comprehension Questions

We want to make sure you carefully read and understood the above instructions. Please answer the following questions.

How high can the quiz score maximally be?

What is presented on a worker's CV? Select all that apply.

- Age
- Quiz score
- Picture
- Hobbies
- Name

next

Your Task

You will make hiring decisions for several workers. Your hiring decisions determine your bonus payment. As the workers already received a piece-rate for each correctly solved question, your hiring decision does not affect their payoffs.

Your goal is to hire workers who will successfully complete a project. Whether a project is successful depends on the worker's quiz score.

If you choose to hire a worker, your bonus payment has two components:

1. You receive a bonus payment of 2 GBP if the project the worker works on is successful. Whether a project is successful depends on a worker's quiz score. The project is successful with a probability of 10% times the quiz score. For

example, if the worker has a quiz score of 6, the success probability is 60%. **Note that the higher a worker's quiz score is, the higher the success probability is.**
2. **If the hired candidate is male, you will receive an extra 0.50 GBP. This extra payment is received irrespective of a worker's quiz score and whether the project is successful.**

If you choose not to hire a worker, the computer tosses a coin to determine your bonus payment: That is, you will receive the bonus payment of 2 GBP with 50% probability.

You will have to make decisions for 8 workers, of which one is randomly chosen to be payoff relevant. Since each decision has the potential to affect your final payment, you should carefully consider your choices.

Comprehension Questions

We want to make sure you carefully read and understood how your bonus payment will be determined. Please answer the following questions.
Imagine that, in the payoff relevant decision you chose to **hire** that worker. On which factors does your bonus payment depend? Select all that apply.

- A coin toss
- Worker's quiz score
- Worker's gender

Imagine that, in the payoff relevant decision you chose **not to hire** that worker. On which factors does your bonus payment depend? Select all that apply.

Worker's quiz score

A coin toss

Worker's gender

next

You will now see each of the 8 workers for a first impression. Your task is to fill in worker's information in a corresponding form. You can take as long as you like to screen each candidate. If you wish to proceed to the next profile, fill in the information and click "next".

Please click "next" to start the task of part 1.

next



Jessica
Thomas

7/10

What is the first letter in the worker's surname?

How many points did the worker score in the quiz?

next

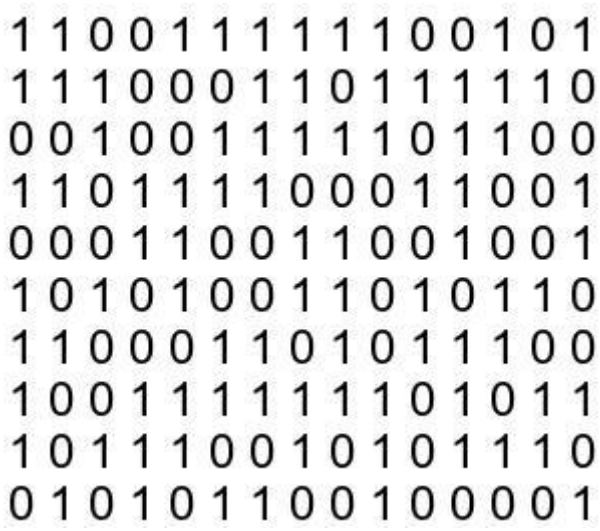
Part 2

For this part of the study, you can receive a bonus payment by working on a task.

Please click "next" to see the instructions of part 2.

next

The task is to count the number of zeros in a matrix consisting of ones and zeros. In the picture you can see the setup.



You enter the number of zeros in the field below the matrix. Then click on “next”. If you have entered the correct number, you will be taken to the next matrix. If your entry was incorrect, you will have to try again. You will have to wait a few seconds before you can submit a new entry. You can only proceed to the next matrix if you submitted the correct number.

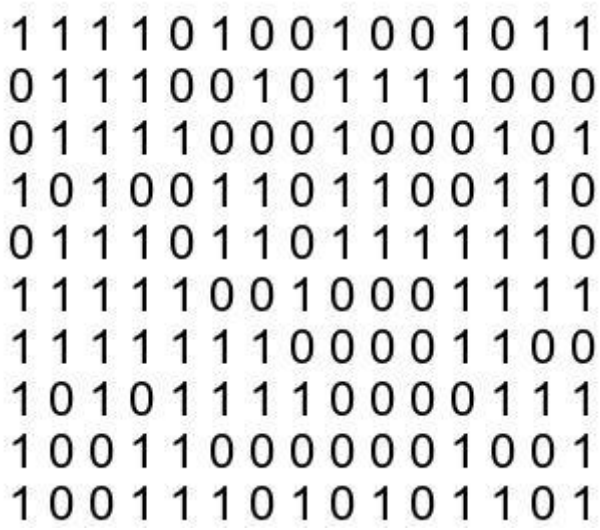
You will receive 0.10 GBP for every matrix you counted correctly.

Please click "next" to start the task of part 2.

next

Time Remaining: 06:13

How many zeros are in the matrix?



next

Part 3

Please click "next" to see the instructions of part 3.

next

Recall that in the beginning of this experiment you saw 8 workers.

Now, we will show you each worker again, asking you to make a hiring and a recommendation decision.

As explained earlier, note that your bonus payment for the hiring decision depends on the worker's **quiz score and gender.**

In particular:

If you choose to hire a worker, your bonus payment has two components:

1. You receive a bonus payment of 2 GBP if the project the worker works on is successful. Whether a project is successful depends on a worker's quiz score. The project is successful with a probability of 10% times the quiz score. For example, if a worker has a quiz score of 6 in the performance task, the success probability is 60%. **Note that the higher a worker's quiz score is, the higher the success probability is.**
2. **If the hired candidate is male, you will receive an extra 0.50 GBP. This extra payment is received irrespective of a worker's quiz score and whether the project is successful.**

If you choose not to hire a worker, the computer tosses a coin to determine your bonus payment: That is, you will receive the bonus payment of 2 GBP with 50% probability.

The recommendation decision has no impact on your payouts, but it may **influence decisions of future participants**.

In particular:

Participants of a future experiment will see your recommendation. They will take the same decisions and face the **same incentives as you do**. However, they cannot see the the workers' CVs and thus **base their decisions solely on the recommendations**.

After the hiring and recommendation decision, we will ask you what you recall about the quiz score.

For each worker, we will show you the corresponding name and picture to help you remember. We will also ask you how confident you are that your memory is correct.

You will receive a bonus for precise recall: For the worker that is relevant for your payoff, you will earn a bonus of 0.50 GBP if your recall is exactly correct, 0.25 GBP if you are within 1 of the correct number and 0 GBP if you are 2 or more outside the correct number.
Recall, that the decisions related to one of the workers is relevant for your payoff. For this payoff-relevant worker, you will receive the bonus related to the hiring decision and the bonus for the recall task.

Comprehension Questions

We want to make sure you carefully read and understood the instructions. Please answer the following question.

Who will see your recommendation decision?

- No one
- Participants of a future experiment
- The researchers
- Participants of this experiments

Imagine a worker has a quiz score of 5. What is the probability that you will receive the bonus of 2 GBP?

- less than 50%
- 50%
- more than 50%
- more information is needed to calculate the probability

This question is very simple. You have to choose the option "quiz score".

Which option are you asked to chose?

- gender
- worker CV
- quiz score
- bonus payment
- hiring decision
- recommendation decision

Please click "next" to start the task of part 3.



Charlotte Johnson

???

Do you want to hire this worker?

yes

no

How strongly would you recommend this worker?

0 10 20 30 40 50 60 70 80 90 100

next



Charlotte Johnson

???

Please recall: How well did the worker score in the quiz?

0	1	2	3	4	5	6	7	8	9	10

next



Charlotte Johnson

???

Please recall: How well did the worker score in the quiz?
Your answer:

3

How far away do you think your recall is from the true quiz score?

0 1 2 3 4 5 6 7 8 9 10

next

Now, we ask you to make one more decision regarding the workers: to **rank** them.

Please note that this task may have real consequences. You may be selected to participate in a lottery in which one worker is picked to the be hired. Your lottery payment will, as before, depend on the **quiz score** and **gender** of this worker. The higher you rank a worker, the more likely this worker is picked. **Therefore, you should rank the workers truthfully: Rank 1 should correspond to the worker you most want to hire, while Rank 8 should correspond to the worker you least want to hire.**

If you are selected for the lottery, we will notify you within the next 10 days.

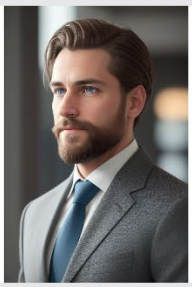
Here, you can read the exact explanation of the lottery if you are interested:

- One of the participants who, like you, do this experiment today, will be chosen to play a lottery.
- If you are selected to play the lottery, you can win an extra bonus payment.
- Each lottery ticket corresponds to a worker.
- You will then hire the winning ticket worker for an extra project.
- As before, the quiz score and gender of this worker will determine your bonus.
- The higher you rank a worker, the more lottery tickets you get for this worker.
- In particular, you will get 8 lottery tickets for the worker you ranked first, 7 lottery tickets for the worker you ranked second, etc.



Matthew Evans

???



Daniel Hall

???



Rebecca Roberts

???



Charlotte Johnson

???



Joshua White

???



James Thompson

???



Jessica Thomas

???



Emily Taylor

???

For your references, here a list of the workers and your recall of the quiz scores:

Matthew Evans	3	Daniel Hall	4
Rebecca Roberts	3	Charlotte Johnson	3
Joshua White	7	James Thompson	5
Jessica Thomas	6	Emily Taylor	4

To rank the listed workers drag & drop each name.

1. Emily Taylor
2. Charlotte Johnson
3. Joshua White
4. Matthew Evans
5. Rebecca Roberts
6. Jessica Thomas
7. James Thompson
8. Daniel Hall

next

Part 4

To complete the study, please answer a short survey about yourself and the study you just performed.

Please click "next" to start the survey.

next

How important were your memories of the quiz score for your hiring decision?

Note: You need to click on the slider even if you wish to select the default.

The qualification task mattered little for my hiring decision — The qualification task mattered a lot for my hiring decision

0 10 20 30 40 50 60 70 80 90 100

Move the slider to answer

How important was the worker's gender for your hiring decision?

Note: You need to click on the slider even if you wish to select the default.

The gender mattered little for my hiring decision — The gender mattered a lot for my hiring decision

0 10 20 30 40 50 60 70 80 90 100

Move the slider to answer

When you were asked to hire workers. What were your considerations that determined whether you hired a candidate or not? Please answer in 2-3 sentences.

next

Please recall: How many male workers did you hire?

0	1	2	3	4	5

Please recall: How many female workers did you hire?

0	1	2	3	4	5

next

How many of your recalled quiz scores do you think were correct?

0	1	2	3	4	5	6	7	8	9	10

When recalling the workers' quiz scores. Did you use any memory aid such as screenshots or notes to give your memory a helping hand? Please respond truthfully to this question. It has NO impact on your payment or your future invitations to participate in studies on Prolific.

Yes

No

next

What do you think this research was about?

next

Thank you for participating! Your payment will consist of:

Base payment: 3 GBP
Bonus payment related to your hiring decision: 2 GBP
Bonus payment from the recall task: 0.25 GBP
Bonus payment from the counting task: 0.1 GBP

The total payment will be transferred to your prolific account within 10 days.

next

Is there anything else you would like to tell us?

next