NET WORK LISER

Discussion Paper Series

IZA DP No. 18517

April 2026

Human–AI Evaluation and Gender Transparency: Application Decisions in Competitive Hiring

Bernd Irlenbusch
University of Cologne, LSE and
IZA@LISER

Holger A. Rau
University of Duisburg-Essen and
University of Göttingen

Rainer Michael Rilke
WHU – Otto Beisheim School of Management

IZA DP No. 18517 April 2026

Human–AI Evaluation and Gender Transparency: Application Decisions in Competitive Hiring*

Abstract

LLMs are rapidly entering the hiring process, but their most pronounced effects may occur before any screening by changing who chooses to apply. We study how human versus LLMbased evaluation and gender transparency shape entry into competitive jobs. In a preregistered online experiment, participants first complete a Niederle and Vesterlund (2007) tournament task to measure competitive preferences, then prepare text-based job applications and decide whether to apply under each of four evaluation regimes — human only, LLM only, and two hybrid human-in-theloop configurations — while gender disclosure is randomized between subjects. LLM involvement reduces application rates, with stronger effects for women than men, including under hybrid designs. Effects are driven by non-competitive candidates; non-competitive women, the group most exposed to AI-induced deterrence, receive the strongest objective evaluations under pure AI assessment across all subgroups, yet are systematically underconfident and apply least often. Competitive men persistently apply and exhibit overconfidence-driven adverse selection, whereas competitive women show resilience to AI-induced deterrence while remaining well-calibrated under AI evaluation and exhibiting positive self-selection across regimes. We find no effects of gender transparency.

JEL classification

C92, J71, J24, O33

Keywords

AI hiring, LLMs, algorithm aversion, gender differences

Corresponding author

Bernd Irlenbusch
bernd.irlenbusch@uni-koeln.de

* This experiment was pre-registered on AsPredicted.org under #200,822 (November 22, 2024) and received ethical approval from the University of Cologne (reference 240009BI). Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC2126/2 – 390838866 and the Center for Social and Economic Behavior at the University of Cologne is gratefully acknowledged.

We thank Maria Cubel, Gerald Eisenkopf, Olaf Korn, Dorothea Kübler, Christiane Schwieren, Dirk Sliwka, and Stefan Traub for valuable discussions, and seminar and workshop participants at the University of Cologne, TU Hamburg, the Clausthal Behavioral Economics Workshop, and the Gender and STEM Workshops in Heidelberg and Kyoto for helpful comments.

1 Introduction

The use of artificial intelligence (AI) in hiring has expanded as firms seek to automate applicant screening and manage growing candidate pools more efficiently. Recent surveys suggest that nearly half of companies worldwide have integrated AI into their human resource operations (Tidio Editorial Team, 2024). Advances in natural language processing now enable AI systems to evaluate qualifications and rank candidates at scale (Gan et al., 2024; Tambe et al., 2019). Proponents argue that algorithmic hiring promises more standardized evaluations and reduced human bias (Alexander et al., 2025; Raghavan et al., 2020). Others identify ethical risks, including the introduction of algorithmic bias, privacy losses, a lack of transparency and explainability, obfuscation of accountability, and the potential loss of human oversight (Hunkenschroer and Luetge, 2022). As a result, emerging regulations—most prominently the European Union's AI Act—classify AI-based hiring systems as "high-risk" applications that require human oversight (European Parliament and Council, 2024).

Organizations therefore face design choices along a spectrum of AI involvement, ranging from purely human decisions to hybrid human-in-the-loop configurations to fully automated systems. Understanding how candidates respond to different degrees of AI integration—particularly hybrid arrangements increasingly mandated by regulation—is critical to designing effective and equitable hiring systems. Yet whether such configurations affect candidates' willingness to apply and whether these effects differ by gender remain open questions. This issue is especially salient given persistent gender disparities in labor market outcomes, including the gender pay gap (OECD, 2023). If AI evaluation deters qualified candidates, particularly women, from applying, it may widen rather than narrow these gaps. Our first research question, therefore, asks: *How do men's and women's application decisions change as AI plays a larger role in hiring evaluation?*

Whether AI involvement encourages or deters applications depends on competing mechanisms. On one hand, substantial evidence documents algorithm aversion: individuals are often reluctant to trust algorithmic decisions, even when algorithms outperform human judgment (Dietvorst et al., 2015, 2018; Jussupow et al., 2024). If candidates perceive AI systems as opaque or inflexible, algorithmic screening may discourage participation. Should women exhibit stronger algorithm aversion than men, AI adoption could inadvertently widen gender gaps by disproportionately deterring female applicants. On the other hand, algorithmic systems may be perceived as more objective and less susceptible to interpersonal bias than human evaluators. For women who anticipate discrimination in human screening contexts (Charness et al., 2020), AI involvement may reduce expected bias and thereby encourage applications. Supporting this view, recent evidence indicates that acceptance of algorithmic hiring increases when candidates are informed that algorithms do not use gender in their predictions (Dargnies et al., 2026). These considerations motivate our second question: *Does the effect of AI involvement differ when candidates' gender is disclosed versus concealed in the application materials—and does this vary for men and women?*

A further, potentially critical moderator of application behavior is competitive preference. In the foundational tournament paradigm of Niederle and Vesterlund (2007), men exhibit stronger competitive preferences than women, displaying greater willingness to enter performance-based competitions. Job

markets closely resemble such tournaments: applicants compete for limited positions, with only a subset succeeding. Competitive preferences predict educational specialization, career track selection, and earnings (Buser et al., 2014, 2024; Henning et al., 2026). Because competitiveness shapes entry decisions in tournament settings and is closely linked to labor-market success, it may also influence how candidates respond to AI-mediated evaluation. Competitive individuals, who are inherently more willing to enter uncertain environments, may be less sensitive to changes in evaluation architecture and thus more resilient to AI-induced deterrence. We therefore extend the Niederle and Vesterlund (2007) framework to the context of algorithmic hiring—embedding their tournament-entry paradigm directly in our experimental design to elicit competitive preferences—and ask: *Do competitive preferences moderate how men and women respond to algorithmic versus human evaluation?*

Prior work has begun to address these questions, though important gaps remain. Dargnies et al. (2026) document algorithm aversion and the role of gender transparency in AI acceptance, but in a design where all participants submit applications, capturing preferences over evaluators rather than the participation margin. Field evidence from Avery et al. (2024) demonstrates that AI adoption can increase female application rates in tech hiring. However, their design cannot fully disentangle whether candidates respond to the evaluation architecture itself, anticipated transparency levels, or broader organizational signals. Neither study examines how competitive preferences moderate responses to algorithmic systems, nor do they consider participation decisions under hybrid human–AI evaluation regimes or LLM-based assessment of standardized application materials. We address these gaps by embedding application decisions in an incentivized competitive hiring environment, employing the Niederle and Vesterlund (2007) tournament-entry paradigm to elicit competitive preferences, using ChatGPT to evaluate actual written application materials—rather than simple algorithmic predictions—across the full spectrum from pure human to pure AI evaluation including hybrid configurations, and randomizing gender transparency between subjects. We discuss the related literature and derive our hypotheses in Section 2.

We test our hypotheses in an online experiment with a two-stage structure. The first stage elicits competitive preferences using a standard real-effort tournament-entry paradigm (Niederle and Vesterlund, 2007). The second stage embeds application decisions in a hiring environment featuring human, AI, and hybrid evaluation procedures, in which participants first prepare a written application file that is evaluated and then decide whether to apply. Evaluation differs across four regimes: purely human evaluation (Human only), purely AI evaluation (GPT only), and two hybrid regimes in which human and AI evaluations interact (Human w/ GPT Advice and GPT w/ Human Advice); each participant faces all four regimes while gender transparency is manipulated between subjects. Application files are scored with higher scores yielding higher monetary payoffs and are used to rank candidates within fixed groups, creating a tournament-style entry decision with a high-payoff success option and a safe outside alternative. Incentives are structured to ensure incentive compatibility for both candidates and evaluators.

We find clear aversion to algorithmic hiring that intensifies with AI involvement. Compared to Human only evaluation, application rates decline by 4.6 percentage points for women and 3.2 percentage points for men under pure AI evaluation. Crucially, deterrence effects emerge even in hybrid human-AI systems, suggesting that the mere presence of algorithmic input—not just who makes final decisions—shapes

participation choices. This finding has important implications for human-in-the-loop requirements in the EU AI Act: mandating human involvement does not eliminate candidates' aversion to AI-involved processes.

However, this overall pattern masks substantial heterogeneity by competitive preferences. Algorithm aversion is driven primarily by non-competitive candidates, who withdraw significantly from AI-involved processes. In stark contrast, competitive individuals—particularly competitive men—maintain stable application rates regardless of AI involvement. If competitive individuals are disproportionately male and non-competitive individuals withdraw from AI-involved hiring, AI systems may inadvertently increase gender gaps in applicant pools, even when algorithms themselves are unbiased. Our design enables us to evaluate all application files regardless of candidates' entry decisions, allowing us to assess selection quality for both applicants and non-applicants—a counterfactual that is rarely available in field settings. This reveals that competitive men's persistent application behavior reflects overconfidence rather than rational self-selection: they apply despite holding significantly lower objective hiring probabilities than those who abstain, particularly under human evaluation. This adverse selection disappears when AI makes hiring decisions. Interestingly, non-competitive women — the group most deterred by AI involvement — receive the strongest objective evaluations under pure AI assessment across all subgroups, yet are systematically underconfident and apply least often, compounding existing labor market disadvantages and mirroring the pattern in Niederle and Vesterlund (2007) where high-performing women fail to enter competitions despite strong chances of winning. Competitive women show smaller overconfidence gaps than men and are well-calibrated under AI evaluation, exhibiting no adverse selection across regimes. This gender asymmetry in how competitive preferences shape selection quality has important implications for understanding who enters AI-mediated hiring processes.

Our findings contribute to understanding how AI hiring affects labor market participation and gender equity. We demonstrate that algorithm aversion emerges with written application materials evaluated by modern AI systems, and that competitive preferences critically shape application behavior in these contexts. AI adoption can widen gender gaps through multiple mechanisms: non-competitive women face the strongest deterrence, while competitive men's overconfidence leads to adverse selection. Human-in-the-loop configurations can limit these effects, though their effectiveness depends on who retains decision authority and on candidates' competitive preferences.

2 Related Literature and Hypotheses

Prior to data collection, we pre-registered our study, hypotheses, and analysis on AsPredicted.[1] Our experimental design addresses two central questions: (1) How do men's and women's application decisions change as AI plays a larger role in hiring evaluation? (2) Does the effect of AI involvement differ when candidates' gender is disclosed versus concealed in the application materials—and does this vary for men and women? Given competing predictions in the literature, we formulate non-directional hypotheses

[1] AsPredicted.org (#200,822, dated November 22, 2024); the pre-registration is available at `https://aspredicted.org/pq72-4vw6.pdf`.

reflecting genuine uncertainty regarding the direction of effects and test them against the null hypotheses of no effects.

2.1 AI in Hiring and the Effect of Evaluation Regimes

The integration of artificial intelligence into recruitment is altering how firms screen applicants and how candidates perceive their chances of success. A central question in recent research is whether AI involvement encourages or deters diverse applicant pools, a tension defined by the competing mechanisms of "algorithm aversion" and "algorithm appreciation" (Jussupow et al., 2024).

Several studies provide evidence that AI can be a powerful tool for reducing gender disparities by attracting female candidates. Avery et al. (2024) find that the prospect of AI assessment can more than double the fraction of top applicants who are women in the tech sector. This effect is largely driven by female jobseekers' beliefs that AI evaluation is more objective and less prone to interpersonal prejudice than human alternatives. Similarly, Pisanelli (2022) demonstrates that automated resume screening can reduce gender gaps in shortlisting by 43 percentage points by mitigating the unconscious stereotypes of human recruiters. Furthermore, Ip (2025) and Awad et al. (2023) show that while awareness of potential bias deters women, the implementation of debiased algorithms eliminates the gender gap in application decisions for competitive "quant" roles. Schulte Steinberg and Hohenberger (2023) and Koch-Bayram et al. (2023) find that women's preference for AI increases among those who believe in its potential to reduce bias or have experienced workplace discrimination.

However, these potential gains are often offset by systemic "algorithm aversion" —a general reluctance to trust algorithmic decisions even when they outperform humans (Dietvorst et al., 2015). Newman et al. (2020) argue that algorithms are often perceived as reductionistic, assuming they quantify personal attributes into mere metrics while failing to holistically consider qualitative context. This perception leads to lower fairness ratings for AI compared to humans, particularly in subjective tasks like hiring. Fumagalli et al. (2022) further show that while high-performers may prefer algorithms for their perceived meritocracy, lower-performing workers prefer human recruiters who are seen as more prone to weighing personal characteristics over raw task performance.

Hybrid systems that combine human and AI judgment provide a potential middle ground. In a vignette study, Gonzales et al. (2022) find that "augmented" approaches—where humans are supplemented by AI—can mitigate adverse reactions compared to purely automated systems. Newman et al. (2020) suggest that human-AI partnerships only alleviate perceptions of unfairness if the human remains the default decision-maker. Hoffman et al. (2018) demonstrate that hiring managers who overrule automated, non-AI recommendations tend to recruit lower-productivity workers. Lacroux and Martin-Lacroux (2022) find that recruiters are behaviorally more influenced by algorithmic recommendations than they admit.

While this literature establishes that both algorithm aversion and appreciation shape responses to AI hiring, it also shows that willingness to apply can depend on the specific evaluation regime (e.g., varying levels of AI involvement) or on personal characteristics such as gender. Given the mixed evidence, we formulate non-directional hypotheses and test them against the null of no regime differences.

Hypothesis. ***1a*** *[Female Response to Evaluation Regime] The fraction of female participants applying for the position differs across the four evaluation regimes (Human only, Human w/ GPT Advice, GPT w/ Human Advice, GPT only).*

Hypothesis. ***1b*** *[Male Response to Evaluation Regime] The fraction of male participants applying for the position differs across the four evaluation regimes (Human only, Human w/ GPT Advice, GPT w/ Human Advice, GPT only).*

Our within-subject design allows us to test these hypotheses by comparing each participant's willingness to apply across all four evaluation regimes, providing statistical power to detect regime-specific effects while controlling for individual heterogeneity.

2.2 Gender Transparency

A second critical factor in recruitment design is gender transparency—whether the candidate's gender is revealed or concealed during the evaluation process. This transparency may fundamentally alter a candidate's willingness to apply, particularly if they anticipate that evaluators will use gender as a basis for discriminatory treatment (Bohnet, 2016; Charness et al., 2020). Prior experimental evidence by Dargnies et al. (2026) suggests that workers prefer algorithmic hiring significantly more often when they know the algorithm is gender-blind and does not use "gender profiling" in its predictions. Similarly, Ip (2025) demonstrates that awareness of potential gender bias in an algorithm significantly deters qualified women from applying, but that gender-blind algorithms are perceived as the fairest by both men and women and successfully restore female participation.

The effect of transparency may also depend on the perceived objectivity of the evaluator. Avery et al. (2024) show that informing candidates they will be assessed by AI rather than a human can more than double the fraction of top female applicants because women believe AI is less prone to interpersonal prejudice. On the demand side, they find that while human evaluators score women lower than men when names are visible, this gap disappears when names—and thus gender—are hidden. Interestingly, the introduction of AI scores closed the gender gap even when gender information remained visible to the human evaluators, suggesting that algorithmic input can counteract the biases triggered by transparency.

However, reactions to transparency are often nuanced by the social context of the evaluation. Pethig and Kroenung (2023) find that women's preference for algorithms increases specifically when the human alternative is a man (an outgroup evaluator). They suggest that women seek "relative algorithmic objectivity" as a shield in situations where they believe their gender identity might lead to disadvantage. Conversely, Schulte Steinberg and Hohenberger (2023) find that individuals generally prefer female human evaluators over AI, indicating that gender disclosure may have positive or negative effects depending on the gender of the recipient.

Finally, the medium of evaluation may also influence how transparency is perceived. Zhang and Yencha (2022) observe that fairness perceptions and the acceptability of hiring algorithms vary significantly between resume screening and video screening, with women generally finding automated evaluation less acceptable than men.

Taken together, these competing mechanisms imply that the effect of gender transparency is theoretically ambiguous. By testing these effects across humans, AI, and hybrid regimes, this study identifies whether concealing gender can mitigate the algorithm aversion that often suppresses participation at the labor market margin. Thus, we formulate non-directional hypotheses separately for female and male participants:

Hypothesis. ***2a*** *[Female Response to Gender Transparency] Gender transparency (revealing versus concealing candidates' gender to evaluators) influences female participants' application decisions.*

Hypothesis. ***2b*** *[Male Response to Gender Transparency] Gender transparency (revealing versus concealing candidates' gender to evaluators) influences male participants' application decisions.*

Our between-subject design—where one cohort experiences all evaluation regimes with gender disclosed and another experiences the same regimes with gender concealed—enables causal identification of transparency effects across screening architectures.

2.3 The Role of Competitive Preferences

Competitive preferences are a vital moderator of how candidates respond to different screening regimes. The foundational work of Niederle and Vesterlund (2007) established that men enter competitive tournaments twice as often as women, even when performance is equal—a gap driven by differences in overconfidence and taste for competition. Research on gender gaps in competitive preferences suggests that women are less likely to self-select into competitive environments (Niederle and Vesterlund, 2007), so changes in evaluation regimes that reduce perceived competitiveness or bias may affect women's willingness to apply differently than men's.

AI involvement can potentially interact with these preferences in important ways. Awad et al. (2023) and Ip (2025) find that debiased screening methods encourage qualified women to enter competitive positions, reducing the gender gap in application decisions. van Esch et al. (2019) identify that the likelihood of applying is influenced by technology-use motivation and novelty, though these benefits can be offset by AI-related anxiety. If competitive candidates are inherently more willing to enter uncertain environments, they may be less sensitive to changes in evaluation architecture. Non-competitive candidates, by contrast, may be more susceptible to algorithm aversion, particularly if they perceive AI evaluation as adding an additional layer of uncertainty to an already daunting competitive process.

These considerations suggest that the effects documented in our hypotheses are unlikely to be uniform across the competitiveness distribution. We do not formulate separate pre-registered hypotheses for competitive preferences but elicit them before the main experiment following Niederle and Vesterlund (2007) and treat them as a central moderating variable in our empirical analysis.

2.4 Contribution and Gaps Addressed

While the literature has established critical foundations, our study addresses several gaps that remain even in the most closely related experimental and field work. Dargnies et al. (2026) provide important

experimental evidence on algorithm aversion in hiring contexts. However, their design captures preferences over evaluators among participants who always submit applications, rather than the participation margin—the decision of whether to apply at all—which is an important channel through which AI adoption may affect labor market access and gender representation. Moreover, their study employs algorithmic predictions based on regression models rather than the large language models increasingly prevalent in hiring. Field evidence from Avery et al. (2024) persuasively demonstrates that AI adoption increases female application rates in tech hiring, but field settings cannot easily isolate whether candidates respond to evaluation architecture or the broader organizational signal that AI adoption may convey. They also lack the counterfactual data necessary to assess self-selection out of the hiring process. Neither study examines how competitive preferences moderate responses to algorithmic systems, nor do they consider participation decisions under hybrid human–AI evaluation regimes or LLM-based assessment of standardized application materials.

We address these gaps in several ways. First, we focus on the participation margin by embedding application decisions in an incentivized competitive hiring environment that mirrors the structure of the Niederle and Vesterlund (2007) tournament: candidates face a fundamental choice between applying for the job—with a high payoff if selected but zero otherwise—and taking a safe outside option, directly paralleling the tournament-versus-piece-rate tradeoff in their paradigm. Second, we employ ChatGPT to evaluate actual written application materials rather than simple algorithmic predictions, providing participants with an authentic evaluation experience that reflects the natural language processing capabilities of modern AI hiring tools. Third, our within-subject design allows us to observe how the same individuals respond across the full spectrum of AI involvement—from pure human to pure AI evaluation, including hybrid human-in-the-loop configurations increasingly mandated by regulation. Fourth, by eliciting competitive preferences, we can test whether algorithm aversion operates uniformly or disproportionately affects specific subgroups defined by gender and willingness to compete.

3 Experimental Design

To test our hypotheses, we conducted a preregistered incentivized online experiment that combines a standard elicitation of competitive preferences with application decisions in a hiring environment featuring human, AI, and hybrid evaluation procedures. The design allows us to identify (i) how candidates respond to different levels of AI involvement in evaluation, (ii) whether these responses vary with the disclosure of candidates' gender to evaluators, and (iii) whether competitive preferences moderate these effects (see Figure 1 for a schematic presentation of the experimental design).

Our setting has four main advantages. First, it approximates an important feature of real-world hiring by requiring candidates to produce short written application materials that are evaluated by human and AI screeners. Second, because all application files are evaluated regardless of later application choices, we observe both task performance and application quality for eventual applicants and non-applicants—a counterfactual that is typically unobservable in real labor markets. Third, candidates' rankings depend on evaluator scores rather than on other candidates' application decisions, simplifying strategic reasoning.

Fourth, employer incentives are aligned with identifying the most capable candidate, independent of realized hiring outcomes.

The experiment consists of two stages. Participants were informed at the outset that they would complete both stages and that one randomly selected stage would determine their final payment.[2] Instructions were provided sequentially: participants learned the details of Stage 2 only after completing Stage 1. Before Stage 1, participants completed a brief questionnaire on gender, age, and educational attainment. They were informed that some of this information might later be shared anonymously with another participant, depending on treatment assignment.

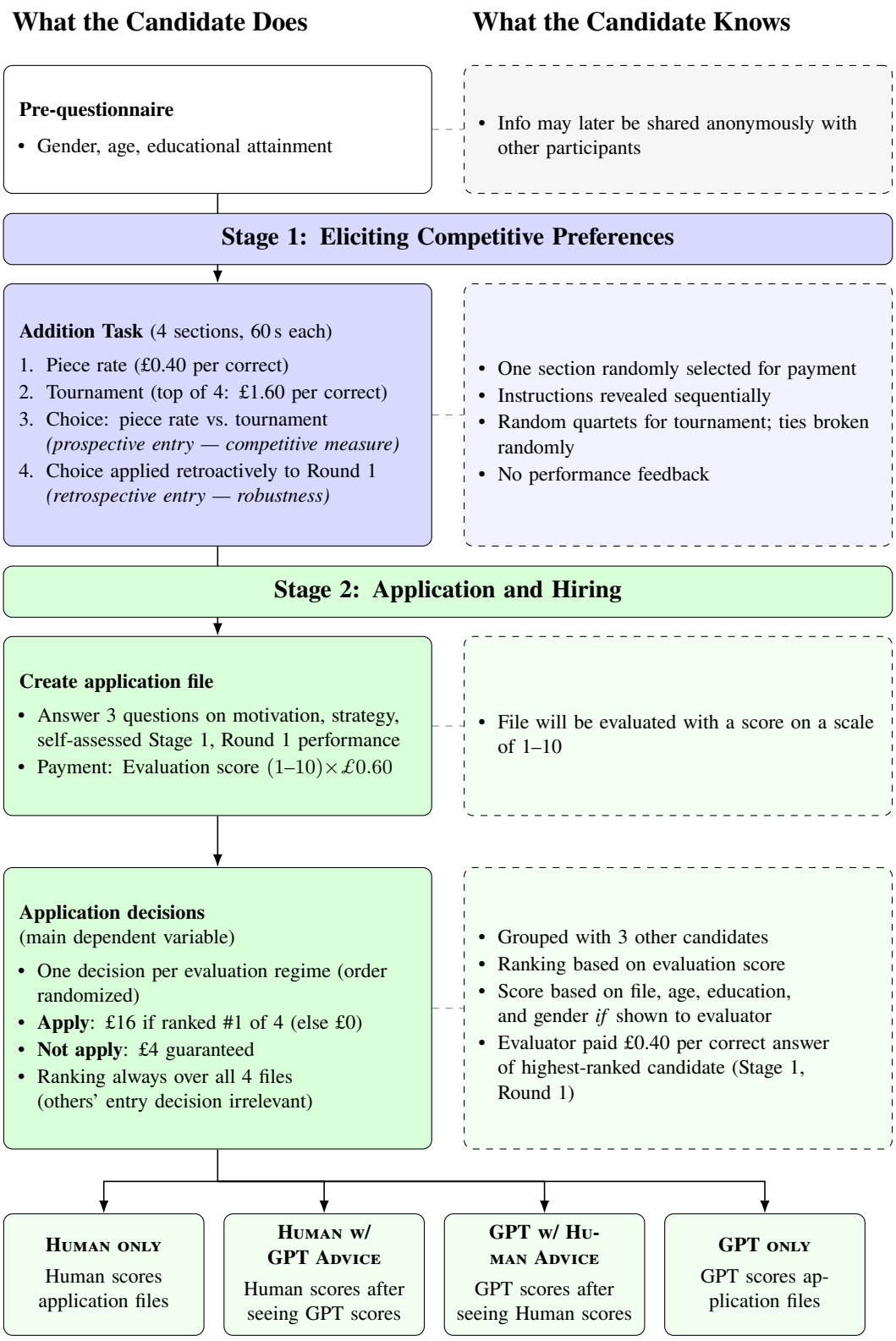
3.1 Stage 1: Eliciting Competitive Preferences

Stage 1 elicits competitive preferences using the tournament-entry paradigm introduced by Niederle and Vesterlund (2007), adapted to an online environment. In this paradigm, participants perform the same real-effort task under both non-competitive and competitive incentive schemes, and willingness to choose tournament incentives over a piece-rate is interpreted as a measure of competitive preference. The key advantage of this design is that it separates willingness to compete from task ability. Participants worked on a mathematical real-effort task in which they repeatedly added two two-digit numbers without a calculator. Each section lasted 60 seconds. To reduce cheating in the online setting, every answer had to be entered within 10 seconds.

Stage 1 comprised four sections. In Section 1, participants were paid on a piece-rate basis, earning £0.40 per correct answer. This provides a baseline measure of individual performance under non-competitive incentives. In Section 2, participants performed the same task under tournament incentives. They were randomly assigned to quartets, and only the highest-performing participant in each quartet received £1.60 per correct answer; all other participants earned zero. Ties for first place were broken randomly. This section exposes all participants to competitive incentives and yields a measure of performance under competitive pressure. In Section 3, participants chose ex ante between a piece-rate and a tournament payment scheme, then completed the task once more. This prospective tournament-entry choice is the standard Niederle–Vesterlund measure of competitive preference and serves as our primary measure of competitiveness. If the non-competitive payment was chosen, participants earned £0.40 per correct answer. Following Niederle and Vesterlund (2007), participants who chose the tournament were matched against the recorded Section 2 performances of three randomly selected other participants. This matching rule ensures that tournament outcomes in the choice section are not mechanically affected by contemporaneous selection into competition. Participants earned £1.60 per correct answer, if they solved strictly more problems than the other three participants; otherwise they earned £0 (ties were resolved by random draw). In Section 4, participants did not solve additional math problems. Instead, they chose which payment scheme should be applied retroactively to their performance in Section 1 (the non-competitive piece-rate round). If the non-competitive payment was chosen, participants earned £0.40 per correct answer from Section 1. If the competitive payment was chosen, participants were randomly

[2]The exact wording and instructions of the experiment can be found in Appendix B.1.

Figure 1: Experimental design overview



matched with three new participants, and their Section 1 performance was compared with the Section 1 performances of the matched participants. Participants earned £1.60 per correct answer if they solved strictly more problems than the other three participants in Section 1; otherwise, they earned £0 (ties were resolved by random draw). We use the retrospective tournament-entry choice in Section 4 as a robustness measure.[3] If Stage 1 was randomly selected as payoff-relevant, a second independent random draw determined which of the four Stage 1 sections determined the participant's payment. This preserves incentive compatibility across all Stage 1 decisions and limits hedging across sections.

3.2 Stage 2: Application and Hiring

Stage 2 operationalizes the hiring environment and consists of two parts: (a) preparation of an application file and (b) application decisions under four evaluation regimes. In Stage 2a, all participants prepared a short written application file by answering three standardized questions about (i) their motivation to perform well in the math task, (ii) their strategy for solving many tasks correctly, and (iii) their self-assessed performance in Section 1 of the Niederle and Vesterlund (2007) math task. Participants were instructed to respond as if applying for a job, creating a stylized but ecologically meaningful analogue of written application materials, and were informed that their payoff would be determined by multiplying their evaluation score (1–10) by 60 pence.[4] All application files were evaluated irrespective of whether participants later chose to apply in Stage 2b, and participants received no feedback before making their application decisions.

In Stage 2b, participants made application decisions for four separate hiring situations differing only in the evaluation regime. They were grouped with three other candidates, with rankings based on evaluation scores; each application file contained written responses, age, and educational attainment, with gender shown or withheld depending on treatment assignment.

The four evaluation regimes vary in the degree and direction of AI involvement. In the HUMAN ONLY regime, a human employer evaluated and scored the four application files. In the GPT ONLY regime, ChatGPT did so independently. These initial scores define the HUMAN ONLY and GPT ONLY regimes. The initial ChatGPT score was then shown to the human employer as advice, after which the human assigned a potentially revised score, defining the HUMAN W/ GPT ADVICE regime; symmetrically, the initial human score was shown to ChatGPT, defining the GPT W/ HUMAN ADVICE regime. Within each group, the same employer and AI model evaluated the same candidates across paired regimes, ensuring that differences in application decisions reflect changes in evaluation architecture rather than evaluator quality or candidate composition. Each participant made four application decisions, one per regime, in fully randomized order. The payoff structure mirrors the tournament-versus-piece-rate choice in Niederle

[3] Appendix Table A.3 confirms that all main findings are robust to replacing the prospective measure with the retrospective choice from Section 4.

[4] Although incentives encourage writing quality, they do not rule out that the texts constitute a form of "cheap talk." To varying degrees, this is the case in all application procedures, i.e., cover or motivation letters are inherently a form of cheap talk. In our data, we observe a positive and highly significant correlation between evaluation scores and actual math-task performance under piece-rate (Stage 1, Section 1) for all regimes (Pearson's $\rho > 0.144, p < 0.001$), indicating that candidates conveyed signals that evaluators were able to detect from the application files.

and Vesterlund (2007): applying yields £16 if ranked first among all four candidates and £0 otherwise, while not applying yields a guaranteed £4. Crucially, all candidates were always ranked regardless of their entry decision, so the strategic attractiveness of applying depends solely on expected rank, not on others' choices.

3.3 Post-Experimental Belief Elicitation

After completing the application decisions, participants completed a post-experimental questionnaire on perceptions of fairness, objectivity, and potential bias in human and AI evaluation—partly inspired by standard digital literacy scales (Avinç and Doğan, 2024; Zeike et al., 2019). We also elicited beliefs about their expected ranking under the different evaluation regimes and their risk tolerance Falk et al. (2023).

3.4 Incentive Structure

Candidate incentives were designed to make all decisions payoff-relevant while limiting hedging across tasks. After completing both stages, a first random draw determined whether Stage 1 or Stage 2 was payoff-relevant. If Stage 1 was selected, a second draw determined which of the four sections was payoff-relevant. If Stage 2 was selected, a second draw determined whether Stage 2a or Stage 2b was payoff-relevant. For Stage 2b, a third draw determined which of the four evaluation regimes was payoff-relevant. Only one decision was ever realized for payment, ensuring that candidates could not hedge across tasks or regimes. Expected earnings were calibrated to be similar across stages to maintain attention throughout the experiment. Employers were incentivized to identify the most capable candidate. In the randomly selected evaluation regime, the employer earned £0.40 per correctly solved problem in the Stage 1 piece-rate task (Section 1) of the candidate they ranked highest, irrespective of whether that candidate applied in Stage 2b. This aligns employer incentives with underlying candidate ability and avoids confounding employer behavior with candidates' application choices.

3.5 Procedures and Sample

The study was approved by the Ethical Review Board of the University of Cologne (Approval No. 2400009B1, February 7, 2024). We recruited participants via Prolific and restricted eligibility to residents of the United Kingdom. We stratified recruitment by gender to obtain an approximately balanced candidate sample. The experiment was programmed in Qualtrics and conducted in English. We imposed no device restrictions. A total of 1,248 individuals started the candidate experiment, yielding a final sample of 1,191 candidates (49.5% female) after excluding non-consenting participants, failed attention checks, and non-completers. Candidates had a mean age of 44.0 years. The median completion time was 17 minutes, and average candidate earnings were £5.98. Payments were made after the employer-side evaluations had been collected; participants received feedback only on final earnings. Employer-side evaluations were collected in a separate but linked study using the same candidate application files and scoring instructions. We recruited 600 employers (300 male, 300 female), equally divided across the GENDER

INFO and GENDER NOINFO treatment arms. Each employer evaluated a group of four candidates on a 1–10 scale. Employers were also recruited via Prolific and received a base payment of £1. Participants who failed attention checks were excluded. We implemented GPT evaluations using OpenAI's GPT-4o model with temperature set to 0. Prompt instructions for GPT-4o were standardized to mirror the instructions provided to human employers as closely as possible, ensuring comparability across evaluation modes.

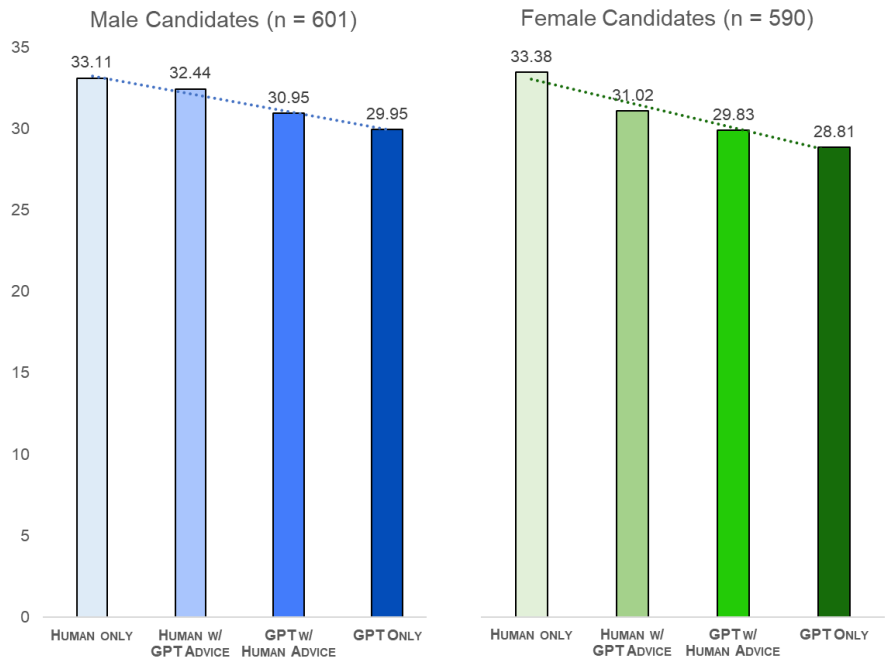
4 Results

We examine how evaluation regimes with different AI involvement (H1) and transparency about applicants' gender (H2) affect application decisions, and how competitiveness moderates these relationships. Panel regression analyses (Section 4.3 and Section 4.4).

4.1 Application Rates by Evaluation Regimes

We begin by examining how varying levels of AI involvement affect application decisions of women (H1a) and men (H1b), exploiting our within-subject design to isolate how the same individual adjusts application decisions across HUMAN ONLY, GPT ONLY, and hybrid evaluation regimes.

Figure 2: Application rates by evaluation regime



Notes: Application rates denote the share of candidates who apply under each evaluation regime. Each candidate makes four application decisions (within-subject design). The left panel shows male candidates (blue tones), the right panel female candidates (green tones). HUMAN ONLY and GPT ONLY denote purely human and purely AI evaluation; HUMAN W/ GPT ADVICE and GPT W/ HUMAN ADVICE denote hybrid regimes.

Figure 2 displays application rates (in percent) for the full sample (n=1,191), separated by male

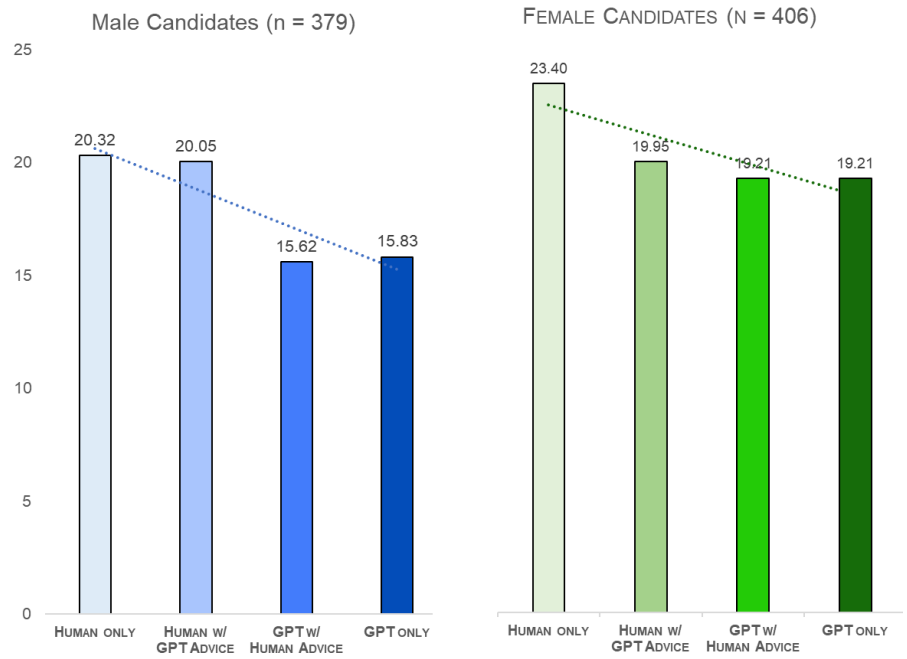
(left panel) and female candidates (right panel). We report exact p-values of two-sided non-parametric matched-pairs tests in the subsequent analyses. Comparing HUMAN ONLY to GPT ONLY evaluation, application rates drop (weakly) significantly for both male (by 3.16 percentage points, from 33.11% to 29.95%, $p=0.048$) and female candidates (by 4.58 percentage points, from 33.39% to 28.81%, $p=0.010$), clearly rejecting the null hypothesis of no regime differences. Human-in-the-loop approaches mitigate AI-induced deterrence, though their effectiveness tends to vary by decision authority and gender. When humans retain final decision authority (HUMAN w/ GPT ADVICE), male candidates' application rates remain virtually unchanged (32.44%, $p=0.738$), whereas female candidates already show an insignificant decline (31.02%, $p=0.175$). This pattern is more pronounced when GPT assumes final decision authority (GPT w/ HUMAN ADVICE): male rates decline modestly (30.95%) ($p=0.193$), whereas female rates (29.83%) drop significantly below HUMAN ONLY ($p=0.038$).

Result 1: *GPT evaluation crowds out male and female candidates, with stronger effects for women. Human-in-the-loop partially mitigates this, but effectiveness varies by decision authority and gender.*

4.1.1 Competitive Preferences: Application Rates of Non-Competitive Candidates

We turn to the role of competitive preferences. Corroborating Niederle and Vesterlund (2007), men showed significantly higher rates of competitiveness (37%) than women (31%, two-sided Fisher's exact test, $p=0.038$), and competitiveness correlates significantly with application willingness (Spearman's $\rho=0.360$, $p<0.001$), confirming that it predicts job market entry (Buser et al., 2014, 2024).

Figure 3: Application rates of non-competitive candidates by evaluation regime



Notes: Sample restricted to candidates who did not enter the tournament in the Niederle and Vesterlund Task.

Figure 3 displays application rates (in percent) for non-competitive male and female candidates, i.e., participants that did not choose to enter the competitive tournament setting of Niederle and Vesterlund (2007) in stage 1 of our experiment.

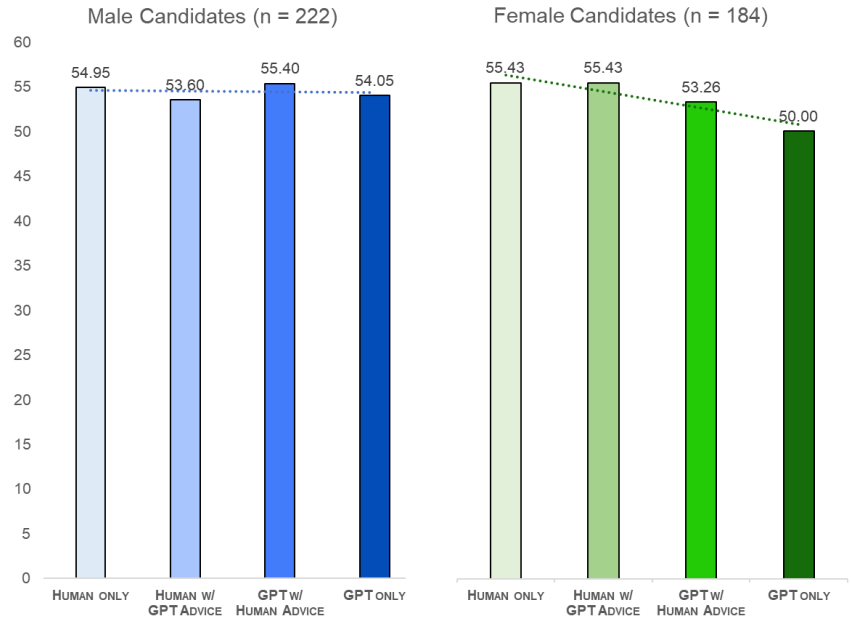
Both genders demonstrate lower willingness to apply to GPT ONLY than to HUMAN ONLY evaluations (male candidates: 20.32% vs. 15.83%, p=0.021; female candidates: 23.40% vs. 19.21%, p=0.057). Examining hybrid evaluation forms reveals that human-in-the-loop approaches are largely ineffective at mitigating AI-induced deterrence among non-competitive candidates. For women and men, in three of four hybrid conditions, application rates remain significantly below the human baseline, reinforcing our findings from the aggregate data. Female candidates are consistently crowded out by AI integration, applying (weakly) significantly less often than to HUMAN ONLY evaluation in both hybrid conditions (HUMAN W/ GPT ADVICE: 19.95%, p=0.087; GPT W/ HUMAN ADVICE: 19.21%, p=0.027). Men experience similar crowding out under GPT decision authority (GPT W/ HUMAN ADVICE: 16.62%, p=0.054). An exception is male candidates under HUMAN W/ GPT ADVICE (20.05%, p=1.000), where human decision authority almost completely restores application rates to HUMAN ONLY.

Result 2: *The deterrence effects from Result 1 are mainly driven by non-competitive candidates. Hybrid systems largely fail to restore application rates, except for men when humans retain decision authority.*

4.1.2 Competitive Preferences: Application Rates of Competitive Candidates

Figure 4 displays application rates of male candidates (left panel, n=222) and female candidates (right panel, n=184) classified as competitive.

Figure 4: Application rates of competitive candidates by evaluation regime



Notes: Sample restricted to candidates who entered the tournament in the Niederle and Vesterlund Task.

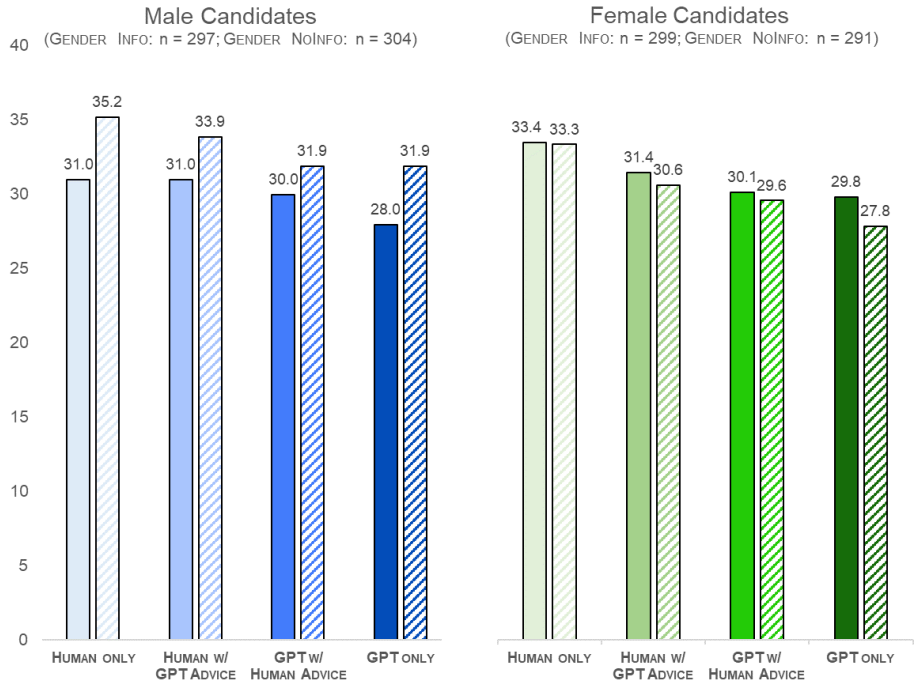
The figure reveals a different pattern for competitive candidates. Male competitive candidates maintain consistently high application rates regardless of the evaluation system, with no significant decline relative to HUMAN ONLY in any regime (all $p>0.711$). Female competitive candidates show some sensitivity to pure AI evaluation: application rates decline by 5.43 percentage points from HUMAN ONLY (55.43%) to GPT ONLY (50.00%), though this difference does not reach conventional significance levels ($p=0.110$). Hybrid systems fully counteract this displacement, maintaining rates statistically indistinguishable from the human baseline (HUMAN W/ GPT ADVICE: 55.43%, $p=1.000$; GPT W/ HUMAN ADVICE: 53.26%, $p=0.636$).

Result 3: *While competitive men maintain high application rates across all evaluation systems, competitive women show a weak and insignificant tendency toward lower application rates under pure GPT evaluation. Hybrid systems tend to restore application rates for competitive women.*

4.2 Gender Transparency Effects

Next, we turn to the between-subject variation, focusing on our second question: how transparency about applicants' gender influences application decisions of female (H2a) and male applicants (H2b). Figure 5 displays application rates across the four evaluation regimes, separately for gender-disclosed (solid bars) and gender-concealed conditions (striped bars).

Figure 5: Application rates by evaluation regime and gender transparency



Notes: Gender transparency effects – Filled bars indicate evaluation with gender disclosed to the evaluator (GENDER INFO); striped bars indicate evaluation with gender concealed (GENDER NOINFO).

We find that application rates are remarkably similar between evaluation systems. If anything, male candidates show a slight tendency toward higher application rates when gender is concealed, whereas no such pattern emerges for women. We find no statistically significant effects of gender transparency on application decisions, neither for male (all two-sided Fisher's exact tests, $p>0.298$) nor for female candidates (all two-sided Fisher's exact tests, $p>0.649$). Thus, we fail to reject the null hypothesis of no transparency effects and find no evidence of gender-specific responses.

These patterns hold when we disaggregate by competitive preferences (see Figures A.1 and A.2 in the appendix), though competitive male candidates show a somewhat more visible tendency toward higher application rates when gender is concealed—a pattern not observed for any other subgroup.[5] Taken together, these findings provide no evidence that gender transparency systematically influences application decisions. We summarize these findings as follows.

Result 4: *Gender transparency in evaluation does not significantly affect overall application decisions for either male or female candidates.*

To quantify the null effects of gender transparency, we compute regime-specific treatment differences in application rates ($\hat{p}_{\text{Gender Info}} - \hat{p}_{\text{Gender NoInfo}}$), separately for men and women under Human only, Human w/ GPT Advice, GPT w/ Human Advice, and GPT only. We construct 95% confidence intervals for the difference using the Newcombe/Wilson method for differences in two independent proportions. For men, point estimates range from -2.0 to -4.7 percentage points, with 95% confidence intervals spanning roughly [-11.9, 2.6] to [-9.1, 5.1] percentage points. For women, point estimates range from 0.1 to 2.4 percentage points, with confidence intervals spanning roughly [-7.1, 7.4] to [-4.5, 9.4] percentage points. All intervals include zero, indicating no statistically detectable transparency effect in any regime.[6]

4.3 Panel Regressions

To assess the robustness of our non-parametric findings, we estimate generalized estimating equation (GEE) panel regressions, treating the four application decisions for each regime as different waves. The panel regressions jointly account for evaluation-regime effects, gender-transparency conditions, and participant heterogeneity. In particular, we examine whether the evaluation regime effects documented in Results 1–3 remain stable when controlling for gender disclosure and individual-level covariates in a unified framework. This approach models each participant's four application decisions using a binomial family with logit link and an exchangeable correlation structure; for all GEE logit models, we report

[5] Among non-competitive participants, gender transparency does not meaningfully affect application rates in any evaluation regime (all two-sided Fisher's exact tests, $p>0.409$ for men; $p>0.559$ for women). Among competitive female candidates, differences are similarly absent ($p>0.298$). For competitive male candidates, a somewhat more discernible tendency toward higher application rates under gender concealment emerges (see Figures A.1 and A.2 in the appendix), most notably in the GPT w/ Human Advice regime ($p=0.060$).

[6] In Figure A.3 in the Appendix, we display differences and confidence intervals for all treatments. We follow Hawkins and Samuels (2021), who argue that confidence intervals provide a richer and more appropriate basis for interpreting nonsignificant findings than post hoc power calculations.

average marginal effects to facilitate comparison across specifications and subsamples and for direct comparability with our non-parametric results.[7]

Table 1 presents GEE Logit panel regressions to examine application decisions using the full sample. The dependent variable is binary, capturing the four application decisions of a participant to each of the evaluation regimes. Model (1) presents results for all participants. Models (2) and (3) disaggregate by competitive preferences. Models (4)–(7) separate the analysis by gender and competitiveness. The baseline category for regime comparisons is the HUMAN ONLY regime evaluation. We control for whether gender information was disclosed to evaluators (GENDER INFO treatment), risk tolerance, and math-task performance in the task of Niederle and Vesterlund (2007) under piece rate (Stage 1, Section 1). Standard errors are clustered at the participant level to account for the within-subject design.[8]

Table 1: Application Decisions: Average Marginal Effects from GEE Logit Panel Regressions

	Combined Gender			Male Candidates		Female Candidates	
	(1) Full Sample	(2) Non-Comp	(3) Comp	(4) Non-Comp	(5) Comp	(6) Non-Comp	(7) Comp
HUMAN W/ GPT ADVICE	-0.015	-0.018	-0.007	-0.002	-0.013	-0.033^{*}	0.000
	(0.011)	(0.013)	(0.020)	(0.017)	(0.025)	(0.019)	(0.032)
GPT W/ HUMAN ADVICE	-0.028^{**}	-0.039^{***}	-0.007	-0.037^{**}	0.005	-0.041^{**}	-0.022
	(0.011)	(0.013)	(0.020)	(0.017)	(0.025)	(0.019)	(0.032)
GPT ONLY	-0.039^{***}	-0.043^{***}	-0.029	-0.045^{***}	-0.009	-0.041^{**}	-0.054^{*}
	(0.011)	(0.013)	(0.020)	(0.018)	(0.025)	(0.019)	(0.032)
GENDER INFO	-0.023	0.006	-0.074^{*}	0.003	-0.105^{*}	0.007	-0.037
	(0.020)	(0.022)	(0.040)	(0.031)	(0.055)	(0.031)	(0.059)
Female	0.059^{***}	0.052^{**}	0.069				
	(0.020)	(0.022)	(0.042)				
Competitive	0.219^{***}						
	(0.018)						
Risk Tolerance	0.049^{***}	0.046^{***}	0.055^{***}	0.049^{***}	0.053^{***}	0.044^{***}	0.058^{***}
	(0.004)	(0.005)	(0.009)	(0.007)	(0.012)	(0.006)	(0.012)
Math-task Performance	0.020^{***}	0.011^{**}	0.035^{***}	0.005	0.033^{***}	0.016^{**}	0.038^{***}
	(0.004)	(0.005)	(0.008)	(0.007)	(0.010)	(0.006)	(0.012)
Observations	4,764	3,140	1,624	1,516	888	1,624	736
Groups	1,191	785	406	379	222	406	184
Wald χ^2	287.34	96.66	45.97	50.60	24.14	49.04	24.16

Notes: Average marginal effects (percentage point changes) from logit GEE. In the panel regressions, we treat the four application decisions for each regime as different waves. Binomial family with logit link and exchangeable correlation structure. Robust standard errors in parentheses. Significance indicators: *** $p<0.01$, ** $p<0.05$, * $p<0.1$

The regression results support our previous non-parametric findings, demonstrating that they are robust to controlling for gender information disclosure and important individual characteristics. Consistent with the matched-pairs tests and in line with algorithm aversion documented by Dargnies et al. (2026),

[7] Further robustness checks with sequence fixed effects confirm that our findings are robust to different orders (Appendix Table A.1).

[8] Candidates' self-assessed math-task performance—elicited as part of the application file—is heavily right-skewed (median = 9, IQR: 7–12) with extreme values up to 100, making it unsuitable as a linear control. We instead control for actual math-task performance under piece rate. The behaviorally relevant beliefs for application decisions are candidates' expected evaluation scores under each regime, which we analyze in Table 2.

we find systematic displacement effects from evaluation institutions with AI integration across the full sample (Model 1). The decline is most pronounced under pure GPT ONLY evaluation (-3.9 percentage points, $p<0.001$). Overall, we confirm that human-in-the-loop approaches can mitigate deterrence effects, though their effectiveness depends on the degree of AI involvement. When humans retain primary decision authority with AI advice (HUMAN w/ GPT ADVICE), the application decline is modest and insignificant (-1.5 percentage points, $p=0.175$). When AI assumes primary decision authority with human advice (GPT w/ HUMAN ADVICE), the deterrence effect is pronounced and significant (-2.8 percentage points, $p=0.010$). These effects remain robust when controlling for GENDER INFO, Risk Tolerance, and Math-task Performance.

The regressions also corroborate the substantial heterogeneity by competitive preferences documented in our non-parametric analysis. Among non-competitive candidates (models 2, 4, 6), AI involvement consistently reduces application rates across all regimes. Importantly, human-in-the-loop approaches prove considerably less effective at mitigating deterrence effects for non-competitive candidates. For non-competitive male candidates (model 4), the HUMAN w/ GPT ADVICE treatment shows no significant deterrence effect (-0.2 percentage points, $p=0.883$). By contrast, for non-competitive female candidates (model 6), human-in-the-loop approaches fail entirely—both HUMAN w/ GPT ADVICE (-3.3 percentage points, $p=0.074$) and GPT w/ HUMAN ADVICE (-4.1 percentage points, $p=0.029$) show (weakly) significant deterrence effects.

Competitive candidates (models 3, 5, 7) show some resilience to AI evaluation, with most coefficients small and statistically insignificant. However, this resilience masks gender differences. For competitive male candidates (model 5), the evaluation regime seems to be largely irrelevant to their application behavior—all marginal effects are close to zero and highly insignificant. These candidates consistently apply regardless of AI involvement. The marginally significant GENDER INFO coefficient in the pooled competitive sample (model 3: -7.4 percentage points, $p=0.074$) appears to be primarily driven by competitive male candidates (model 5: -10.5 percentage points, $p=0.055$), consistent with the weak tendency toward higher application rates under gender concealment documented in the previous section and with Dargnies et al. (2026), who find that candidates exhibit stronger preferences for algorithmic evaluation when explicitly informed that the algorithm is gender-blind. By contrast, competitive female candidates (model 7) exhibit a different pattern: while they show greater resilience than non-competitive women, they still display weak but notable aversion to pure AI evaluation (GPT ONLY: -5.4 percentage points, $p=0.089$). However, human-in-the-loop approaches effectively eliminate this aversion, with both hybrid systems showing negligible and insignificant effects. In sum, the analysis shows that evaluation regime effects remain robust across all specifications regardless of whether we control for gender transparency.

Appendix Table A.1 replicates all seven specifications with 47 full sequence fixed effects absorbing the complete presentation-order variation; evaluation-regime coefficients remain virtually unchanged, confirming that results are not driven by presentation order effects.

Appendix Table A.2 shows that the above presented regression results are also robust when including Competitive $\times$ regime interactions. The interaction terms reveal that the moderating role of competitive-

ness is regime-specific primarily for men: competitive male candidates show attenuated deterrence under both regimes where GPT decides relative to the HUMAN ONLY baseline (both $p<0.10$), a pattern that is visible but weaker in the pooled sample (Model 1). For female candidates, competitiveness operates as a level effect that increases application rates uniformly across regimes without significantly moderating regime-specific deterrence.

Appendix Table A.3 replaces the prospective Niederle–Vesterlund competitiveness measure with the retrospective choice from Section 4 and yields qualitatively identical results across all models.

4.4 Behavioral Correlates of Competitive Resilience

Our results reveal an empirical puzzle: competitive candidates appear largely immune to AI-induced deterrence. We explore whether observable differences in beliefs and attitudes toward AI evaluation are associated with this resilience. To systematically assess candidates' attitudes toward AI involvement in hiring, we conducted a post-experimental survey measuring perceived objectivity of evaluation systems, concerns about stereotyping and bias, perceptions of potential harm, and broader attitudes toward AI including trust and appreciation of AI characteristics. We employ principal component analysis to reduce these survey items into six standardized psychological constructs: (i) perceived objectivity (PC_{obj}), (ii) stereotyping concerns (PC_{stereo}), (iii) harm concerns (PC_{hurt}), (iv) trust in AI evaluation ($PC_{gpt-trust}$), (v) AI adoption willingness ($PC_{gpt-adoption}$), and (vi) trust in human evaluation ($PC_{employer-trust}$).[9]

Descriptively, competitive candidates report more favorable beliefs and attitudes toward AI evaluation than non-competitive candidates. These differences motivate the subsequent analysis examining whether such beliefs and attitudes are systematically associated with application behavior. Two-sample t-tests reveal that competitive candidates hold significantly higher beliefs about their expected evaluation scores across all regimes (largest difference for pure GPT: $t=11.84$, $p<0.001$), perceive AI as more objective ($t=7.15$, $p<0.001$), and express greater trust in AI systems ($t=5.87$, $p<0.001$) than non-competitive candidates. However, concerns about algorithmic harm do not differ significantly between groups ($t=0.42$, $p=0.674$). We now examine whether these more favorable beliefs and attitudes can account for competitive candidates' immunity to AI-induced deterrence effects.

To investigate whether competitive candidates' immunity to AI evaluation stems from observable differences in beliefs or attitudes, we estimate GEE models analogous to Table 1, augmented with beliefs about expected evaluation scores and six principal components measuring psychological attitudes toward AI and human evaluation (Table 2). Model (1) presents results for the full sample; Models (2)–(5) disaggregate by gender and competitive preferences to examine heterogeneity in how beliefs and psychological attitudes relate to application behavior across demographic groups. The baseline category for regime comparisons is the HUMAN ONLY evaluation.

[9]The first PCA yields PC_{obj} (eigenvalue = 2.65, 44.2% variance) and PC_{stereo} (eigenvalue = 1.45, 24.1%). The second PCA produces PC_{hurt} (eigenvalue = 3.15, 78.8%). The third PCA generates $PC_{gpt-trust}$ (eigenvalue = 3.86, 42.9%), $PC_{gpt-adoption}$ (eigenvalue = 1.15, 12.7%), and $PC_{employer-trust}$ (eigenvalue = 1.04, 11.6%). All components are standardized to mean 0, SD 1.

Table 2: Beliefs and Psychological Channels (Average Marginal Effects)

	Combined	Male Candidates		Female Candidates	
	(1) Full Sample	(2) Non-Comp	(3) Comp	(4) Non-Comp	(5) Comp
HUMAN W/ GPT ADVICE	-0.015	-0.002	-0.014	-0.033^{*}	0.000
	(0.011)	(0.018)	(0.027)	(0.020)	(0.031)
GPT W/ HUMAN ADVICE	-0.028^{**}	-0.037^{**}	0.005	-0.041^{**}	-0.022
	(0.011)	(0.019)	(0.027)	(0.020)	(0.031)
GPT ONLY	-0.039^{***}	-0.045^{**}	-0.009	-0.041^{**}	-0.055^{*}
	(0.011)	(0.019)	(0.027)	(0.020)	(0.031)
GENDER INFO	-0.020	0.001	-0.096^{*}	0.017	-0.020
	(0.018)	(0.027)	(0.050)	(0.028)	(0.056)
Female	0.054^{***}				
	(0.019)				
Competitive	0.190^{***}				
	(0.017)				
Risk Tolerance	0.037^{***}	0.037^{***}	0.031^{**}	0.038^{***}	0.032^{**}
	(0.004)	(0.007)	(0.013)	(0.006)	(0.013)
Math-task Performance	0.012^{***}	-0.004	0.033^{***}	0.011^{*}	0.023^{*}
	(0.004)	(0.006)	(0.010)	(0.006)	(0.012)
Beliefs about Evaluation Scores					
Belief: HUMAN W/ GPT ADVICE	0.009	0.066^{***}	-0.020	-0.021	0.002
	(0.010)	(0.015)	(0.024)	(0.016)	(0.028)
Belief: GPT W/ HUMAN ADVICE	0.022^{**}	-0.003	0.003	0.049^{***}	0.025
	(0.010)	(0.016)	(0.026)	(0.015)	(0.027)
Belief: GPT ONLY	0.024^{***}	0.027^{**}	0.038^{*}	0.013	0.021
	(0.008)	(0.013)	(0.022)	(0.012)	(0.024)
Belief: Human	0.020^{**}	-0.001	0.042^{**}	0.024^{*}	0.021
	(0.008)	(0.012)	(0.021)	(0.012)	(0.022)
Psychological Attitudes (PCA, standardized)					
PC_{obj}	0.000	-0.049^{**}	0.072^{**}	-0.010	0.049
	(0.013)	(0.020)	(0.036)	(0.019)	(0.032)
PC_{stereo}	-0.002	0.020	0.019	-0.021	-0.054^{*}
	(0.010)	(0.014)	(0.025)	(0.015)	(0.029)
PC_{hurt}	-0.001	-0.005	0.067^{***}	-0.010	-0.058^{**}
	(0.010)	(0.014)	(0.025)	(0.015)	(0.029)
$\text{PC}_{gpt-trust}$	-0.001	-0.019	-0.004	-0.000	0.033
	(0.010)	(0.015)	(0.029)	(0.015)	(0.031)
$\text{PC}_{gpt-adoption}$	-0.009	-0.022	0.008	-0.025	0.041
	(0.010)	(0.015)	(0.030)	(0.015)	(0.032)
$\text{PC}_{employer-trust}$	-0.015	-0.020	-0.006	-0.015	0.019
	(0.010)	(0.014)	(0.025)	(0.016)	(0.027)
Observations	4,764	1,516	888	1,624	736
Groups	1,191	379	222	406	184

Notes: Entries report average marginal effects (percentage point changes) from logit GEE models with exchangeable correlation structure. Robust standard errors in parentheses. Continuous variables are standardized; marginal effects for beliefs and PCA components refer to a one-standard-deviation increase. Baseline evaluation regime is HUMAN ONLY. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

In Model 1, all evaluation-regime coefficients and the competitiveness gradient remain virtually identical to Table 1. Beliefs about expected evaluation scores are positively associated with willingness to apply, while none of the six psychological attitude components reach significance (all $p>0.10$).

Disaggregating by gender and competitive preferences reveals where these associations concentrate. Non-competitive male and female candidates (Models 2 and 4) continue to exhibit substantial deterrence from AI-involved evaluation regimes even after conditioning on beliefs and psychological attitudes, whereas competitive candidates (Models 3 and 5) display markedly greater stability in their application behavior. For competitive male candidates (Model 3), all evaluation-regime coefficients remain close to zero and statistically insignificant, mirroring the pattern observed in Table 1. Among the psychological attitude components, perceived objectivity (PC_{obj}) and harm concerns (PC_{hurt}) are significantly associated with application behavior, suggesting that competitive men who perceive AI as more objective and who are more concerned about potential harm from evaluation systems are paradoxically more likely to apply. However, these associations do not attenuate the core competitive resilience pattern—evaluation-regime coefficients remain unchanged—indicating that they reflect individual heterogeneity within the competitive male subgroup rather than a mechanism that explains their immunity to AI-induced deterrence. Competitive female candidates (Model 5) exhibit a similar pattern: while they display some sensitivity to pure GPT evaluation, associations between beliefs or attitudes and application decisions are weak and do not materially attenuate the competitiveness gradient. Notably, stereotyping concerns (PC_{stereo}) and harm concerns (PC_{hurt}) are negatively associated with application rates among competitive women, suggesting that those who worry more about AI bias or rejection are less likely to apply—yet these associations do not explain the regime-level deterrence effects.

Taken together, these results indicate that competitive resilience reflects a stable behavioral trait rather than differences in observable beliefs or attitudes toward AI evaluation. Beliefs about expected evaluation scores—rather than psychological attitudes—emerge as the more consistent predictors of application behavior. We emphasize that beliefs and attitudes are measured after application decisions and are therefore interpreted as correlates rather than causal mechanisms.

4.5 Candidate Decision Quality

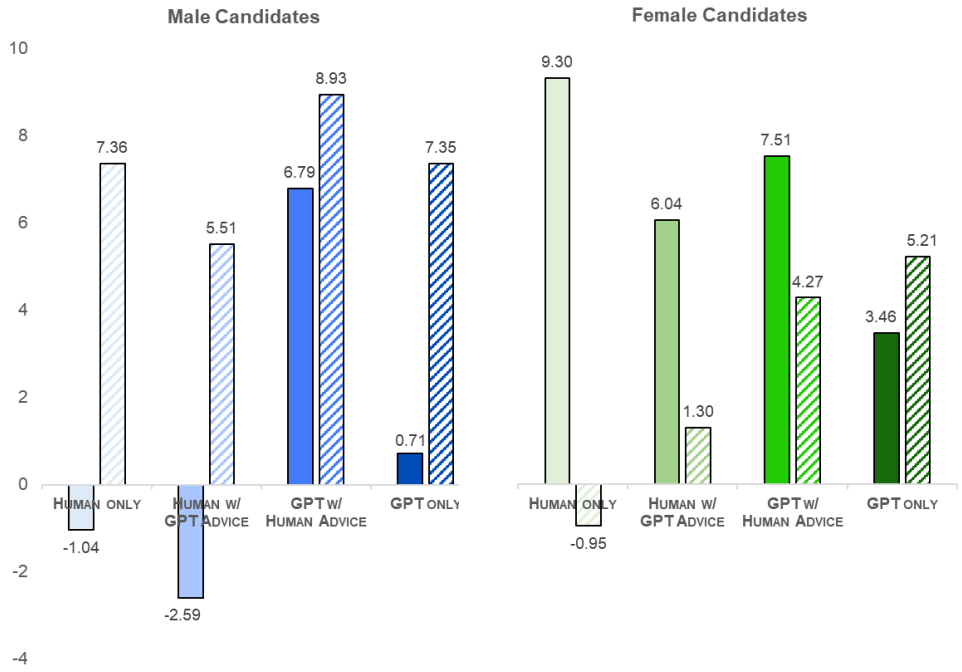
Given that competitive candidates apply persistently across all evaluation regimes, while non-competitive candidates refrain from AI-involved evaluation, two critical questions emerge. *Are these divergent application patterns rational? Are these patterns driven by accurate self-assessment, or by overconfidence among competitive candidates and excessive caution among non-competitive ones?*

To assess whether candidates make rational application decisions, we examine the relationship between application decisions and candidates' objective probability of getting the job when applying. For each candidate and evaluation regime, we compute the winning probability as the average probability of achieving the highest evaluation score across all possible groups of four candidates drawn from the same evaluation regime, accounting for ties via a fair random draw. This measure directly translates each candidate's actual evaluation score into their expected hiring probability under that regime, providing an objective benchmark for rational self-selection. Under rational behavior, candidates who correctly anticipate their relative standing should apply if and only if their winning probability is sufficiently high to justify the risk relative to the safe outside option. This implies that applicants should on average hold

higher winning probabilities than non-applicants within each group and regime. Moreover, larger positive differences indicate better-calibrated self-selection — candidates are more accurately sorting themselves based on their objective hiring prospects. Negative differences, by contrast, signal adverse selection: candidates with below-average winning probabilities disproportionately choose to enter the competition. To test this prediction, we compare the mean winning probability of applicants versus non-applicants within each gender-competitiveness-regime combination.

Figure 6 displays these differences (applicants' mean winning probability minus non-applicants' mean winning probability) across evaluation regimes, disaggregated by gender (left panel: male candidates; right panel: female candidates) and competitiveness. Competitive (non-competitive) candidates are represented by filled (stripe) areas in the bars. Positive values indicate that applicants have a higher objective probability of winning the hiring tournament than non-applicants, consistent with rational self-selection, while negative values indicate adverse selection, whereby candidates with lower winning probabilities disproportionately choose to apply.

Figure 6: Differences in mean winning probability between applicants and non-applicants by gender, competitiveness, and evaluation regime: Filled (striped) bars denote competitive (non-competitive) candidates.



Notes: Bars show differences in mean winning probability between applicants and non-applicants (applicants minus non-applicants) within each gender-competitiveness group and evaluation regime. Winning probabilities are computed as each candidate's objective probability of achieving the highest evaluation score in a randomly drawn group of four from the same treatment pool, averaged over all possible such groups. Positive values indicate positive selection — applicants have a higher objective hiring probability than non-applicants; negative values indicate adverse selection. Filled (striped) bars denote competitive (non-competitive) candidates. Sample sizes vary slightly across regimes ($n \approx 575$–590 per group) due to data collection issues affecting a small number of employer sessions that could not be matched to candidate groups.

Differences are assessed using two-sided Mann-Whitney tests comparing the winning probabilities of applicants versus non-applicants within each group-regime combination. The figure shows that most differences are positive, indicating that applicants tend to hold higher objective hiring probabilities than non-applicants, but this pattern masks heterogeneity across gender and competitive preferences. Among male candidates, the pattern differs by competitiveness. Competitive men do not hold significantly higher winning probabilities than non-applicants in most evaluation regimes. Specifically, the differences are negative and insignificant under the regimes where a human decision-maker decides ($p=0.475$ for HUMAN ONLY; $p=0.541$ for HUMAN W/ GPT ADVICE), indicating adverse selection that is statistically indistinguishable from random entry. Under the regimes where GPT decides, the pattern reverses directionally, marginally so under the hybrid regime ($p=0.085$ for GPT W/ HUMAN ADVICE) but not under pure GPT evaluation ($p=0.538$ for GPT ONLY). Non-competitive men show significant positive selection under HUMAN ONLY ($p=0.013$) and GPT ONLY ($p=0.033$), with insignificant differences in the hybrid regimes.

Among female candidates, the pattern is more uniform across competitiveness levels. Competitive women exhibit consistently positive and largely significant differences across regimes ($p=0.053$ for HUMAN ONLY; $p=0.083$ for HUMAN W/ GPT ADVICE; $p=0.009$ for GPT W/ HUMAN ADVICE; $p=0.205$ for GPT ONLY). Non-competitive women show no significant positive selection in human-dominated regimes (all $p>0.629$ for HUMAN ONLY and HUMAN W/ GPT ADVICE), while differences trend towards significance under GPT-decision regimes ($p=0.070$ for GPT W/ HUMAN ADVICE; $p=0.111$ for GPT ONLY).

To formally test the patterns in the figure above, we estimate in Table 3 OLS regressions of winning probabilities on an application indicator, separately for each of the four gender-competitiveness groups. Each regression includes all four evaluation regimes jointly via regime dummies and their interactions with the application indicator. For ease of presentation, we do not report the raw regression coefficients but instead present linear combinations of the apply indicator coefficient and the relevant apply $\times$ regime interaction term, directly mapping each table entry to the corresponding bar in the figure above: for HUMAN ONLY — the baseline regime — the entry corresponds to the apply indicator coefficient alone; for the remaining regimes, it is the sum of the apply indicator and the relevant interaction term, recovering the regime-specific applied effect in a single interpretable quantity. Each entry thus captures the conditional difference in winning probability between applicants and non-applicants within a given regime and subgroup, controlling for gender information treatment, risk tolerance, and math-task performance under piece rate. A positive entry indicates rational self-selection; a negative entry indicates adverse selection.

Table 3: Applicant Selection Quality by Gender, Competitiveness, and Evaluation Regime

	Male Candidates		Female Candidates	
	(1) Non-Comp	(2) Competitive	(3) Non-Comp	(4) Competitive
HUMAN ONLY	0.063	−0.010	−0.010	0.093**
	(0.040)	(0.037)	(0.031)	(0.042)
HUMAN W/ GPT ADVICE	0.039	−0.026	0.013	0.060
	(0.039)	(0.037)	(0.036)	(0.043)
GPT W/ HUMAN ADVICE	0.071	0.068*	0.043	0.075*
	(0.044)	(0.040)	(0.034)	(0.038)
GPT ONLY	0.055	0.007	0.052	0.035
	(0.040)	(0.035)	(0.036)	(0.042)
Observations	1,488	860	1,584	720
Subjects	372	215	396	180
Controls	Yes	Yes	Yes	Yes

Notes: Robust standard errors in parentheses, clustered at the subject level. Winning probabilities reflect each candidate's objective probability of achieving the highest evaluation score in a randomly drawn group of four from the same treatment pool, averaged over all possible such groups. Significance indicators: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

The regression results corroborate the non-parametric findings. Among competitive women, applicants hold significantly higher winning probabilities than non-applicants under HUMAN ONLY evaluation ($\beta=0.093$, $p=0.029$), and the difference remains marginally significant under GPT W/ HUMAN ADVICE ($\beta=0.075$, $p=0.052$), while coefficients are positive but insignificant in the remaining regimes, consistent with rational self-selection. Among competitive men, applied effects are negative under both human-decision regimes and close to zero under GPT ONLY, though none reach conventional significance levels. The sole exception is GPT W/ HUMAN ADVICE, where the applied effect turns positive and marginally significant ($\beta=0.068$, $p=0.094$), suggesting that adverse selection attenuates when GPT holds primary decision authority. Among non-competitive candidates, applied effects are uniformly positive across regimes and genders but fall short of significance in all cases.

What explains the adverse selection among competitive men? We turn to belief calibration to examine whether systematic overconfidence can account for this pattern. Table 4 compares believed versus actual evaluation scores across groups and regimes.

Table 4: Belief Calibration: Expected vs. Actual Evaluation Scores by Gender and Competitiveness

		Human only	Human w/ GPT Advice	GPT w/ Human Advice	GPT only
Panel A: Male Candidates					
	Believed Evaluation Score	7.43 (1.71)	7.17 (1.86)	7.00 (1.88)	6.76 (2.08)
Competitive	Actual Evaluation Score	6.58 (1.68)	5.88 (2.13)	6.54 (2.51)	4.71 (3.41)
	2-sided matched pairs test	$p<0.001$	$p<0.001$	$p=0.051$	$p<0.001$
	Believed Evaluation Score	6.63 (1.90)	6.33 (1.91)	6.23 (1.87)	5.94 (2.11)
Non-Competitive	Actual Evaluation Score	6.78 (1.78)	6.05 (2.14)	6.49 (2.51)	4.86 (3.48)
	2-sided matched pairs test	$p=0.182$	$p=0.032$	$p=0.104$	$p<0.001$
Panel B: Female Candidates					
	Believed Evaluation Score	7.14 (1.92)	6.87 (1.99)	6.82 (2.01)	6.51 (2.07)
Competitive	Actual Evaluation Score	6.66 (1.75)	6.41 (2.10)	6.12 (2.48)	6.16 (3.50)
	2-sided matched pairs test	$p=0.005$	$p=0.013$	$p=0.005$	$p=0.288$
	Believed Evaluation Score	6.61 (1.81)	6.37 (1.77)	6.17 (1.84)	5.86 (1.99)
Non-Competitive	Actual Evaluation Score	6.70 (1.68)	6.59 (2.00)	6.19 (2.45)	6.24 (3.43)
	2-sided matched pairs test	$p=0.509$	$p=0.101$	$p=0.849$	$p=0.041$

Note: Believed evaluation scores are elicited from the post-experimental survey; actual evaluation scores are assigned by human employers or GPT. Standard deviations in parentheses. p-values are from two-sided matched pairs tests comparing believed and actual evaluation scores within each group and evaluation regime.

The table reveals systematic overconfidence among competitive candidates of both genders: they hold high performance expectations across all evaluation regimes, yet receive substantially lower actual evaluation scores. In what follows, we present two-sided matched pairs tests. For competitive men, this overconfidence is particularly pronounced in human-evaluation contexts (Human only and Human w/ GPT Advice), both highly significant ($p<0.001$). Competitive women also exhibit significant overconfidence in human-evaluation contexts ($p=0.005$ for both Human only and Human w/ GPT Advice), though their belief-actual gaps are considerably smaller (0.48 and 0.46 points, compared to 0.85 and 1.29 points for competitive men). Strikingly, this overconfidence disappears under pure AI evaluation: competitive women show no significant miscalibration under GPT only evaluation ($p=0.288$), while competitive men remain significantly overconfident even there ($p<0.001$), consistent with our finding that competitive preferences provide stronger protection against AI-induced deterrence for men than for women.

The picture differs markedly for non-competitive candidates. Non-competitive men are largely well-calibrated, with significant overconfidence emerging only under Human w/ GPT Advice and GPT onlyevaluation. Non-competitive women, by contrast, are well-calibrated or even significantly underconfident: under pure GPT only evaluation, they believe they will score only 5.86 on average, yet actually receive 6.24—-a gap that is statistically significant ($p=0.041$). This underconfidence under GPT evaluation is particularly striking given that non-competitive women actually receive the highest actual evaluation scores under GPT only among all male subgroups, significantly outperforming both competitive men ($p=0.094$, two-sided Mann-Whitney test) and non-competitive men ($p=0.086$, two-sided Mann-Whitney test), while scoring comparably to competitive women ($p=0.808$, two-sided Mann-Whitney test). Their underconfidence thus directly mirrors the finding that non-competitive women

exhibit the largest application rate declines under GPT ONLY evaluation—pessimistic beliefs suppress applications even among those who are objectively well-suited. Taken together, male overconfidence drives adverse selection in human-evaluation contexts, while female underconfidence suppresses participation under AI evaluation.

Result 5: *Competitive men exhibit adverse selection under regimes with human decision-making, applying despite holding lower objective winning probabilities than non-applicants, driven by systematic overconfidence. Competitive women show the opposite pattern, with applicants holding significantly higher winning probabilities than non-applicants. Non-competitive women are significantly underconfident under GPT evaluation despite receiving the strongest objective evaluations of any subgroup.*

5 Conclusion

In this paper, we examined how AI involvement in hiring affects candidates' willingness to apply for competitive positions. In our pre-registered online experiment, participants made real application decisions across four evaluation regimes: pure human evaluation, pure AI evaluation, and two human-in-the-loop configurations. We measured competitive preferences using the Niederle Vesterlund (2007) paradigm and manipulated gender transparency to evaluators. We find systematic reluctance to be evaluated by AI systems. Compared to human-only evaluation, application rates decline by 3.2 percentage points for men and 4.6 percentage points for women under pure AI evaluation. Human-in-the-loop systems reduce but do not eliminate this deterrence, with effectiveness varying by decision authority and gender. These aggregate patterns mask substantial heterogeneity.

Consistent with prior evidence documenting the importance of competitive preferences for job market behavior (Buser et al., 2014, 2024; Henning et al., 2026; Reuben et al., 2024), we find that competitiveness fundamentally moderates how candidates respond to algorithmic evaluation. Underlying this heterogeneity, we replicate established gender differences in competitive preferences: men show higher competitiveness than women (37% vs. 31%), mirroring robust patterns from Niederle and Vesterlund (2007) and confirmed across diverse contexts (Buser et al., 2021; Hauge et al., 2023; Markowsky and Beblo, 2022). Non-competitive candidates, particularly non-competitive women, face the strongest deterrence from AI-involved hiring. Strikingly, these are also the candidates who receive the strongest objective evaluations under pure AI assessment — yet their systematic underconfidence prevents them from capitalizing on precisely the institution that evaluates them most favorably, risking a deepening of existing gender gaps as AI hiring spreads. Human-in-the-loop approaches largely fail to restore their application rates. In stark contrast, competitive individuals—especially competitive men—maintain stable application rates regardless of AI involvement. Since competitive individuals are disproportionately male and non-competitive individuals withdraw from AI evaluation, AI adoption may widen gender gaps through differential self-selection rather than algorithmic bias. Our analysis of behavioral channels reveals that the resilience of competitive candidates operates as a stable trait, rather than reflecting more favorable beliefs about AI evaluation. This suggests that competitiveness—as measured by tournament

entry in the Niederle-Vesterlund paradigm—captures an intrinsic preference for competitive selection processes that operates independently of institutional features or evaluator characteristics.

A critical question is whether these differential application patterns reflect rational self-selection. Our analysis reveals that competitive preferences affect selection quality differently for men and women. Among male candidates, competitiveness undermines rational self-selection specifically in human-evaluation contexts: competitive men exhibit significantly worse selection quality than non-competitive men when humans make hiring decisions, applying despite holding lower objective winning probabilities. This adverse selection disappears when AI assumes decision authority. We find systematic male overconfidence driving this pattern: competitive men hold inflated performance expectations across all evaluation regimes, particularly in human-evaluation contexts. In contrast, competitive women are well-calibrated under AI evaluation and exhibit positive self-selection across regimes, with applicants holding higher objective winning probabilities than non-applicants. This pattern echoes Barber and Odean's (2001) finding that overconfident male investors trade excessively to their detriment, and extends Niederle and Vesterlund's (2007) evidence that men "compete too much" in tournament entry decisions to a job market setting: in our data, competitive men persistently enter the hiring tournament despite systematically lower evaluation scores, replicating the over-entry pattern in a context where the stakes are an actual job. Notably, among female candidates, competitiveness operates beneficially: competitive women show resilience to AI crowding out across all evaluation systems while maintaining positive selection quality. Unlike competitive men, they exhibit no evidence of irrational over-application or adverse selection. This gender asymmetry suggests that competitive preferences help women navigate AI hiring systems without the overconfidence costs observed among men.

Our findings contribute to understanding how technological change in hiring practices intersects with behavioral gender differences to shape labor market equity. First, we focus on the participation margin by embedding application decisions in an incentivized competitive hiring environment following the Niederle and Vesterlund (2007) tournament-entry framework, where candidates face the fundamental choice between applying with the risk of rejection and taking a safe outside option. This enables us to examine algorithm aversion at the participation margin—whether candidates enter the competition at all—complementing Dargnies et al. (2026), who study expressed evaluator preferences in a design where all participants submit applications by construction. While they show that gender-blind algorithms increase workers' preferences for algorithmic evaluation, our results suggest this may be insufficient: the more consequential barrier is that AI involvement suppresses entry in the first place, particularly among non-competitive women for whom neither hybrid configurations nor human-in-the-loop designs restore participation. Understanding the participation decision is important given the growing role of intermediaries in modern hiring: while Cowgill and Perkowski (2024) show that third-party recruiters balance employer and candidate preferences when screening applications, our results demonstrate that self-selection into the applicant pool already varies systematically by competitive preferences and evaluation technology before such intermediation occurs. Second, we reveal substantial heterogeneity masked by aggregate effects: competitive preferences fundamentally moderate responses to algorithmic hiring, with differential effects on selection quality by gender. Competitive men apply persistently but irra-

tionally due to systematic overconfidence, while competitive women combine resilience with accurate self-assessment under AI evaluation, exhibiting smaller overconfidence gaps than competitive men in human-evaluation contexts. Third, we provide policy-relevant evidence on human-in-the-loop configurations mandated by regulations like the EU AI Act: while such systems can mitigate deterrence for some demographic groups, they prove insufficient for others, particularly non-competitive women who continue to withdraw from AI-involved processes.

The differential effectiveness of human-in-the-loop requirements across demographic groups carries important policy implications. Policymakers designing fairness interventions must recognize that one-size-fits-all approaches inadequately address heterogeneous responses to algorithmic systems. Organizations adopting AI hiring should monitor not only algorithmic bias in evaluation but also demographic shifts in applicant pools driven by differential self-selection. Our null findings on gender transparency suggest that disclosure requirements alone may not eliminate deterrence effects. A more fundamental reconsideration of how AI hiring systems are designed, implemented, and communicated may be necessary to prevent algorithmic hiring from widening existing gender gaps.

References

Alexander, L., Song, Q. C., Hickman, L., and Shin, H. J. (2025). Sourcing algorithms: Rethinking fairness in hiring in the era of algorithmic recruitment. *International Journal of Selection and Assessment*, 33(1):e12499.

Avery, M., Leibbrandt, A., and Vecci, J. (2024). Does artificial intelligence help or hurt gender diversity? Evidence from two field experiments on recruitment in tech. *CESifo Working Paper*, (10996).

Avinç, E. and Doğan, F. (2024). Digital literacy scale: Validity and reliability study with the rasch model. *Education and information Technologies*, 29(17):22895–22941.

Awad, E., Balafoutas, L., Chen, L., Ip, E., and Vecci, J. (2023). Artificial intelligence and debiasing in hiring: Impact on applicant quality and gender diversity. *SSRN Working Paper*, (4626059).

Barber, B. M. and Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292.

Bohnet, I. (2016). *What works: Gender equality by design*. Harvard University Press, Cambridge, MA.

Buser, T., Cappelen, A., Gneezy, U., Hoffman, M., and Tungodden, B. (2021). Competitiveness, gender and handedness. *Economics & Human Biology*, 43:101037.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129(3):1409–1447.

Buser, T., Niederle, M., and Oosterbeek, H. (2024). Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures. *Review of Economics and Statistics*.

Charness, G., Cobo-Reyes, R., Meraglia, S., and Sánchez, Á. (2020). Anticipated discrimination, choices, and performance: Experimental evidence. *European Economic Review*, 127:103473.

Cowgill, B. and Perkowski, P. (2024). Delegation in hiring: Evidence from a two-sided audit. *Journal of Political Economy: Microeconomics*, 2(4):852–882.

Dargnies, M., Hakimov, R., and Kübler, D. (2026). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*, 72(1):285–301.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

European Parliament and Council (2024). Artificial intelligence act. Regulation (EU) 2024/1689. Official Journal of the European Union, L 2024/1689.

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

Fumagalli, E., Rezaei, S., and Salomons, A. (2022). OK computer: Worker perceptions of algorithmic recruitment. *Research Policy*, 51(2):104420.

Gan, C., Zhang, Q., and Mori, T. (2024). Application of LLM agents in recruitment: A novel framework for automated resume screening. *Journal of Information Processing*, 32:881–893.

Gonzales, M. F., Liu, W., Shirase, L., Tomczak, D. L., Lobbe, C. E., Justenhoven, R., and Martin, N. R. (2022). Allying with AI? reactions toward human-based, AI/ML-based, and augmented hiring processes. *Computers in Human Behavior*, 130:107179.

Hauge, K. E., Kotsadam, A., and Riege, A. (2023). Culture and gender differences in willingness to compete. *The Economic Journal*, 133(654):2403–2426.

Hawkins, A. T. and Samuels, L. R. (2021). Use of confidence intervals in interpreting nonstatistically significant results. *Jama*, 326(20):2068–2069.

Henning, J., Loth, N., Müller, S., Rau, H. A., and Wolff, M. (2026). Battle of the sexes? A cross-cultural analysis of gender differences in competitiveness and pay. Unpublished manuscript.

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

Hunkenschroer, A. L. and Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, 178(4):977–1007.

Ip, E. (2025). Fair AI in hiring: Experimental evidence on how biased hiring algorithms and different debiasing methods affect the quality and diversity of applicants. *Behavioral Science & Policy*, 11(1):44–54.

Jussupow, E., Benbasat, I., and Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, 48(4):1575–1590.

Koch-Bayram, I. F., Kaibel, C., Biemann, T., and Triana, M. d. C. (2023). Applicants' experiences with discrimination explain their reactions to algorithms in personnel selection. *International Journal of Selection and Assessment*, 31(2):252–266.

Lacroux, A. and Martin-Lacroux, C. (2022). Should I trust the artificial intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. *Frontiers in Psychology*, 13:895997.

Markowsky, E. and Beblo, M. (2022). When do we observe a gender gap in competition entry? A meta-analysis of the experimental literature. *Journal of Economic Behavior & Organization*, 198:139–163.

Newman, D. T., Fast, N. J., and Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160:149–167.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

OECD (2023). Reporting gender pay gaps in OECD countries: Guidance for pay transparency implementation, monitoring and reform. Technical report, OECD Publishing, Paris.

Pethig, F. and Kroenung, J. (2023). Biased humans, (un)biased algorithms? *Journal of Business Ethics*, 183(3):637–652.

Pisanelli, E. (2022). Your resume is your gatekeeper: Automated resume screening as a strategy to reduce gender gaps in hiring. *Economics Letters*, 221:110892.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481. Association for Computing Machinery.

Reuben, E., Sapienza, P., and Zingales, L. (2024). Overconfidence and preferences for competition. *The Journal of Finance*, 79(2):1087–1121.

Schulte Steinberg, A. L. and Hohenberger, C. (2023). Can AI close the gender gap in the job market? Individuals' preferences for AI evaluations. *Computers in Human Behavior Reports*, 10:100287.

Tambe, P., Cappelli, P., and Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15–42.

Tidio Editorial Team (2024). AI in recruitment: Benefits, use cases examples. `https://www.tidio.com/blog/ai-recruitment/`. Accessed: 2025-03-28.

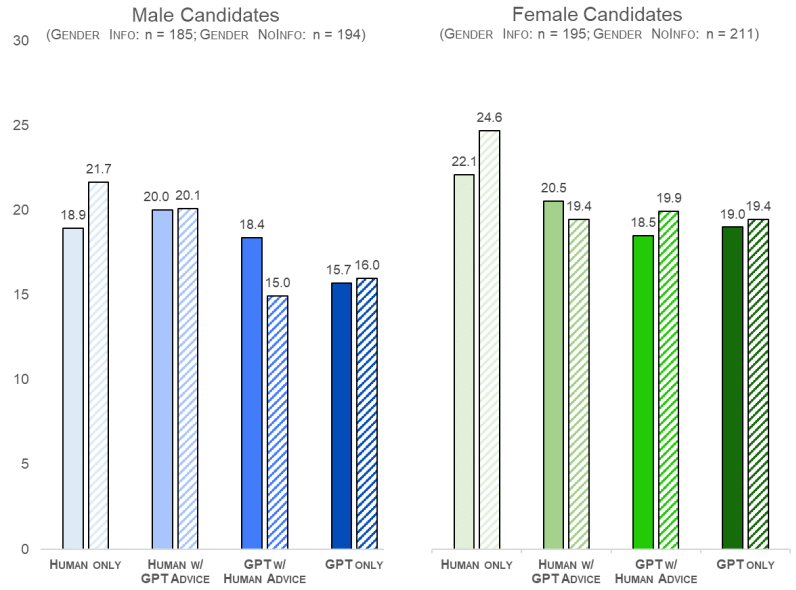
van Esch, P., Black, J. S., and Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, 90:2015–220.

Zeike, S., Bradbury, K., Lindert, L., and Pfaff, H. (2019). Digital leadership skills and associations with psychological well-being. *International Journal of Environmental Research and Public Health*, 16(14):2628.

Zhang, L. and Yencha, C. (2022). Examining perceptions towards hiring algorithms. *Technology in Society*, 68:101848.

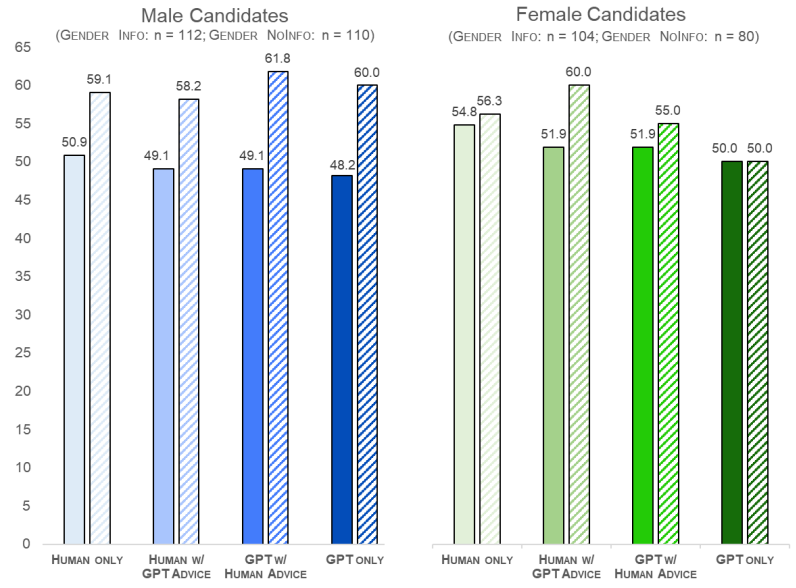
A Figures & Tables

Figure A.1: Application rates by evaluation regime and gender transparency (non-competitive candidates)



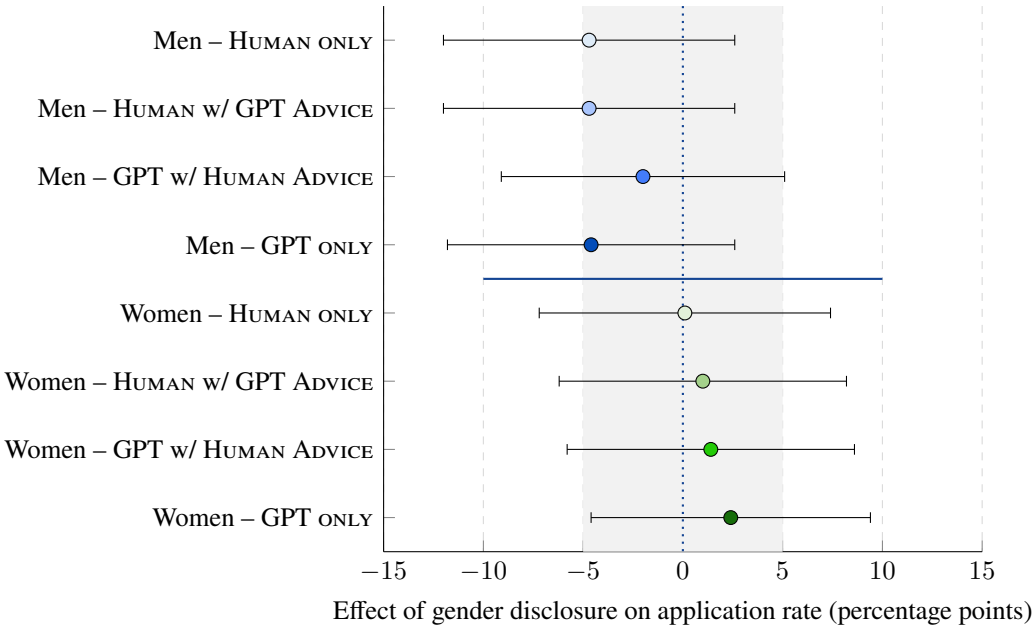
Notes: Gender transparency effects – Candidates who did not enter the tournament in the Niederle and Vesterlund Task. Filled bars indicate evaluation with gender disclosed to the evaluator (Gender Info); striped bars indicate evaluation with gender concealed (Gender NoInfo).

Figure A.2: Application rates by evaluation regime and gender transparency (competitive candidates)



Notes: Gender transparency effects – Candidates who entered the tournament in the Niederle and Vesterlund Task. Filled bars indicate evaluation with gender disclosed to the evaluator (Gender Info); striped bars indicate evaluation with gender concealed (Gender NoInfo).

Figure A.3: Gender transparency effects



Notes: Points show differences in application rates (Gender Info minus Gender NoInfo) with 95% confidence intervals (Newcombe/Wilson method). The shaded area indicates ±5 percentage points.

Table A.1: Average Marginal Effects from GEE Logit Regression on Application Rates (Robustness: Full Sequence Order Fixed Effects)

	Combined Gender			Male Candidates		Female Candidates	
	(1) Full Sample	(2) Non-Comp	(3) Comp	(4) Non-Comp	(5) Comp	(6) Non-Comp	(7) Comp
Human w/ GPT Advice	-0.015	-0.018	-0.007	-0.003	-0.014	-0.035^{*}	0.000
	(0.011)	(0.013)	(0.020)	(0.019)	(0.027)	(0.020)	(0.036)
GPT w/ Human Advice	-0.028^{**}	-0.039^{***}	-0.007	-0.039^{**}	0.005	-0.043^{**}	-0.024
	(0.011)	(0.013)	(0.020)	(0.020)	(0.027)	(0.020)	(0.036)
GPT only	-0.039^{***}	-0.043^{***}	-0.030	-0.048^{**}	-0.009	-0.043^{**}	-0.059^{*}
	(0.011)	(0.014)	(0.020)	(0.020)	(0.027)	(0.019)	(0.036)
Gender Info	0.146	0.132	0.156	0.078	-0.245	0.202	0.529^{*}
	(0.096)	(0.092)	(0.205)	(0.130)	(0.289)	(0.130)	(0.308)
Female	0.056^{***}	0.060^{***}	0.062				
	(0.020)	(0.022)	(0.043)				
Competitive (Sec. 3)	0.217^{***}						
	(0.018)						
Risk Tolerance	0.050^{***}	0.047^{***}	0.054^{***}	0.051^{***}	0.052^{***}	0.044^{***}	0.045^{***}
	(0.004)	(0.004)	(0.009)	(0.007)	(0.013)	(0.006)	(0.014)
Math-task Performance	0.020^{***}	0.012^{***}	0.041^{***}	0.005	0.050^{***}	0.020^{***}	0.042^{***}
	(0.004)	(0.005)	(0.008)	(0.007)	(0.010)	(0.006)	(0.013)
Order FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,760	3,108	1,620	1,440	864	1,552	680
Groups	1,190	777	405	360	216	388	170
Wald χ^2	302.67	129.34	65.95	75.43	48.74	77.97	42.90

Notes: Average marginal effects (percentage point changes) from logit GEE. Binomial family with logit link and exchangeable correlation structure. Robust standard errors in parentheses. Order FEs: 47 sequence fixed effects absorbing full presentation-order variation. Some order categories dropped due to perfect prediction in subgroup models. Significance Indicators: *** p<0.01, ** p<0.05, * p<0.1

Table A.2: Avg. Marginal Effects from GEE Logit Regression on Application Rates (Interactions)

	(1) Full Sample	(2) Male Candidates	(3) Female Candidates
Human w/ GPT Advice	-0.022	-0.003	-0.038^{*}
	(0.014)	(0.020)	(0.021)
GPT w/ Human Advice	-0.046^{***}	-0.045^{**}	-0.047^{**}
	(0.015)	(0.020)	(0.021)
GPT only	-0.051^{***}	-0.056^{***}	-0.047^{**}
	(0.015)	(0.020)	(0.021)
Competitive × Human w/ GPT Advice	0.016	-0.007	0.038
	(0.022)	(0.029)	(0.033)
Competitive × GPT w/ Human Advice	0.041^{*}	0.049^{*}	0.031
	(0.022)	(0.029)	(0.033)
Competitive × GPT only	0.029	0.049^{*}	0.006
	(0.022)	(0.029)	(0.033)
Gender Info	-0.024	-0.040	-0.008
	(0.020)	(0.028)	(0.028)
Female	0.058^{***}		
	(0.020)		
Competitive	0.198^{***}	0.200^{***}	0.196^{***}
	(0.022)	(0.031)	(0.032)
Risk Tolerance	0.049^{***}	0.051^{***}	0.048^{***}
	(0.004)	(0.006)	(0.006)
Math-task Performance	0.020^{***}	0.018^{***}	0.023^{***}
	(0.004)	(0.006)	(0.006)
Observations	4,764	2,404	2,360
Groups	1,191	601	590
Wald χ^2	288.46	149.43	141.33

Notes: Average marginal effects (percentage point changes) from logit GEE. Binomial family with logit link and exchangeable correlation structure. Robust standard errors in parentheses. Baseline evaluation regime is Human only. Competitive × Regime interactions test whether competitive candidates respond differently to AI-involved evaluation regimes relative to the human baseline. Significance indicators: *** p<0.01, ** p<0.05, * p<0.1.

Table A.3: Average Marginal Effects from GEE Logit Regression on Application Rates (Robustness: Alternative Competitiveness Measure)

	Combined Gender			Male Candidates		Female Candidates	
	(1) Full Sample	(2) Non-Comp	(3) Comp	(4) Non-Comp	(5) Comp	(6) Non-Comp	(7) Comp
Human w/ GPT Advice	-0.015	-0.011	-0.024	-0.002	-0.015	-0.018	-0.037
	(0.011)	(0.012)	(0.022)	(0.016)	(0.028)	(0.018)	(0.034)
GPT w/ Human Advice	-0.028^{**}	-0.026^{**}	-0.033	-0.020	-0.024	-0.032^{*}	-0.043
	(0.011)	(0.012)	(0.022)	(0.017)	(0.028)	(0.018)	(0.034)
GPT only	-0.038^{***}	-0.030^{**}	-0.057^{**}	-0.036^{**}	-0.024	-0.025	-0.097^{***}
	(0.011)	(0.012)	(0.022)	(0.017)	(0.028)	(0.018)	(0.034)
Gender Info	-0.006	0.022	-0.059	-0.012	-0.050	0.053^{*}	-0.069
	(0.020)	(0.022)	(0.042)	(0.031)	(0.057)	(0.030)	(0.061)
Female	0.060^{***}	0.051^{**}	0.090^{**}				
	(0.020)	(0.022)	(0.043)				
Competitive (Sec. 4)	0.229^{***}						
	(0.018)						
Risk Tolerance	0.049^{***}	0.048^{***}	0.054^{***}	0.054^{***}	0.055^{***}	0.043^{***}	0.053^{***}
	(0.004)	(0.005)	(0.009)	(0.007)	(0.012)	(0.006)	(0.012)
Math-task Performance	0.019^{***}	0.011^{**}	0.032^{***}	0.005	0.029^{***}	0.019^{***}	0.036^{***}
	(0.004)	(0.005)	(0.008)	(0.007)	(0.010)	(0.007)	(0.011)
Observations	4,764	3,284	1,480	1,580	824	1,704	656
Groups	1,191	821	370	395	206	426	164
Wald χ^2	294.36	97.80	43.64	54.99	20.37	47.50	25.02

Notes: All specifications replicate Table 1 replacing the prospective tournament-entry choice (Section 3) with the retrospective competitiveness measure elicited in Section 4, in which participants chose whether to apply tournament incentives retroactively to their Section 1 performance. Main treatment coefficients and the competitiveness gradient remain qualitatively unchanged. Significance Indicators: *** p<0.01, ** p<0.05, * p<0.1

B Appendix: Experimental Instructions

This appendix provides the verbatim wording of instructions presented to participants. We focus on the core experimental tasks, omitting routine elements (consent forms, demographic questions, attention checks).

B.1 Candidates Experiment

Sequence of the Experiment & Short Questionnaire

The study will consist of two consecutive parts (Part 1 and Part 2). Upon completion: A random selection by the computer will determine which of these two parts determine your final payment. After a few days you will be informed about the selected part and your corresponding payments.

Before the beginning of part 1, you are required to fill out a short questionnaire. The information provided may be shared anonymously with another participant during the experiment.

---------- NEW SCREEN ----------

Part 1

In part 1 of the experiment, you complete four sections. If at the end of the experiment, part 1 will be randomly selected to be payoff-relevant, we will randomly select one of the four sections that will determine your payoff.

---------- NEW SCREEN ----------

Section 1.1: Non-competitive payment

Task description: Your task is to calculate the correct sum of a series of two-digit numbers.

Time limit: Each calculation must be completed within 10 seconds.

Duration: The task lasts for a total of 60 seconds.

Payment: The payment is non-competitive in the sense that you will receive 40 pence for each correct answer independently of what other participants are doing.

Hint: Before you can enter an answer, you must click on the answer line with the mouse pointer.

---------- NEW SCREEN ----------

Section 1.2: Competitive payment

Task overview: You participate in the same 60-second mathematical task as in Section 1.1.

Tournament: You will be randomly paired with a group of three other participants to enter a tournament based on your task performance.

Payment: The payment is competitive in the sense that your success is determined by comparing the number of correctly solved math problems against that of the three randomly paired participants who did the task under identical conditions as you.

Winning: If you are the one who correctly solved the most problems in this group of four, you receive 160 pence per correct answer.

Losing: If not, you receive 0 pence per correct answer.

Tie: If there is more than one participant who correctly solved the most problems in the group of four, a random draw decides who is winning and losing among the participants with the most correctly solved problems.

---------- NEW SCREEN ----------

Section 1.3: Choice between payments

Task overview: You participate in the same 60-second mathematical task as before.

Choice (before the task starts):

Non-competitive payment: You can choose that you receive 40 pence for each correctly solved problem.

or:

Competitive payment: You can choose to enter a competition, where your performance will be measured against the performance that three other randomly paired participants achieved under the competitive payment scheme. You will be matched with different participants compared to those you were matched with under the competitive payment scheme in Section 1.2. You win, if you achieved the highest number of correctly solved math problems in this group.

Winning the Competition: You receive 160 pence for each correct answer.

Losing the Competition: You receive 0 pence for each correct answer. In case of a tie a random draw decides.

Make your choice

- I choose the non-competitive payment
- I choose the competitive payment

---------- NEW SCREEN ----------

Section 1.4. - Choice

Task overview: You do not have to add any numbers for the this task of the experiment. Instead you may be paid one more time for the number of problems you solved in the Section 1.1. – Non-competitive payment.

However, you now have to choose which payment scheme you want to be applied to the number of problems you solved. You can either choose to be paid according to the Non-competitive payment, or according to the competitive payment.

Payment: If the fourth task is the one selected for payment, then your earnings for this task are determined as follows.

If you choose the non-competitive payment you receive 40 pence per problem you solved in Section 1.1.

If you choose the competitive payment your will be randomly matched with three other new participants. Your performance will be evaluated relative to the performance of the other three participants of these other three participants in the Section 1.1. non-competitive payment.

If you correctly solved more problems in Section 1.1 than they did then you receive 160 pence per problem you solved in Section 1.1.

You will receive no earnings for this task if you choose the competitive payment and did not solve more problems correctly in Section 1.1. than the other members of your group. In case of a tie a random draw decides.

I want to be paid according to the ...

- Non-competitive payment
- Competitive payment

---------- NEW SCREEN ----------

Part 2

Part 2 of the experiment consists of two sections. If in the end of the experiment, part 2 will be randomly selected to be payoff-relevant, we will pay you for one section which will be randomly chosen from the two.

---------- NEW SCREEN ----------

Section 2.1 - Application file

In section 2.1 you participate in a questionnaire based on three questions related to your:

- Motivation,
- Skills, and
- Guess of Own Task Performance

regarding the math task of part 1 in the non-competitive payment scheme.

When writing your answers assume that they are used as an application file for a job that builds on the task. This application file will be evaluated (details see below) with a score on a 1-10 scale (1 = completely unsuitable for a job based on this task; 10 = completely suitable for a job based on this task).

Payment details:

For Section 2.1, your payment is based on your evaluation score, calculated as: Score x 60 pence.

Please provide answers to the following three questions as if you apply for a job with this task and need to convince your potential employer to hire you from a pool of four potential applicants. This will constitute your application file.

The higher the score of your application file, the higher your payment.

Questions:

1. *Motivation:* Please briefly describe your motivation to perform well in the calculation task.

2. *Skills:* Please briefly describe your strategy to be as successful as possible in correctly solving many math tasks (e.g., your focus: speed vs. accuracy during the task, etc.).?

3. *Guess of Own Task Performance:* Please guess how many of the sums you have calculated correctly in the math task of part 1 in the non-competitive payment scheme?

---------- NEW SCREEN ----------

Section 2.2. - Application Decision

In Section 2.2., again, you will be paired with three new randomly-selected participants distinct from the ones you were previously paired with. Like you, these participants have also taken part in this study.

There will be rankings between you and the three participants you are paired with. The rankings will be built from the evaluation scores your application file and the application files from the others will receive.

You will make a decision whether to apply or not, i.e., submit your application file.

If you apply

- . . . and are ranked first, you will receive a payment of £16.
- . . . and are not ranked first, you will receive a payment of £0.
- . . . and there is a tie between the first ranking positions a random draw will decide.

If you do not apply

- . . . you receive a payment of £4.

A randomly chosen new participant takes the role of an employer.

There are four different cases:

- ChatGPT decides
- Employer decides
- ChatGPT decides with Employer advice
- Employer decides with ChatGPT advice

In each of these four cases there will be a ranking of the four participants. The employer will receive a payment of 40 pence for each math problem that the participant ranked first in the ranking correctly solved in "Section 1.1 - Non-competitive payment".

Your application decision

You will decide whether or not you want to apply for a position in each of the four different cases separately. The case relevant for you to determine the ranking will be selected randomly after you made your decisions for all cases. Your payment depends on the ranking in the randomly selected case.

Note: Participants were randomly assigned to one of two conditions. In the "Gender Information" condition, participants were told that gender, age, and education level would be shared with evaluators. In the "No Gender Information" condition, only age and education level would be shared. Below we present the Gender Information version; the No Gender version was identical except omitting references to gender.

---------- NEW SCREEN ----------

Case: Employer decides

An employer will determine a ranking based on the application files, gender, age, and education level of you and the others. There is a 50% chance you will be matched to a male or a female employer.

- The employer scores each participant from 1 to 10 (1 = completely unsuitable, 10 = completely suitable).
- If the highest-scoring candidate in the employers final ranking applies, this candidate will get the job, i.e., receive £16.
- If there is a tie among the highest scores, the job is assigned randomly among those with the highest score who applied.

Do you want to apply?

- Yes, I would like to apply (winning £16 if I rank first, £0 otherwise).
- No, I do not want to apply (winning £4 for sure, independent of my ranking).

---------- NEW SCREEN ----------

Case: ChatGPT decides

ChatGPT, an artificial intelligence (AI), will determine a ranking based on the application files, gender, age, and education level of you and the others.

- ChatGPT scores each participant from 1 to 10 (1 = completely unsuitable, 10 = completely suitable).
- If the highest-scoring candidate in ChatGPT's final ranking applies, this candidate will get the job, i.e., receive £16.

- If there is a tie among the highest scores, the job is assigned randomly among those with the highest score who applied.

Do you want to apply?

- Yes, I would like to apply (winning £16 if I rank first, £0 otherwise).
- No, I do not want to apply (winning £4 for sure, independent of my ranking).

---------- NEW SCREEN ----------

Case: Employer decides with ChatGPT advice

An employer will determine a ranking with advice from ChatGPT, an artificial intelligence (AI). ChatGPT will determine its ranking first based on the application files, gender, age, and education level of you and the others. This ranking serves as an advice for the employer. Subsequently, the employer determines a ranking based on the advice, the application files, gender, age, and education level of you and the others. There is a 50% chance that you will be matched to a male or a female employer.

- ChatGPT scores each participant from 1 to 10 (1 = completely unsuitable, 10 = completely suitable).
- The employer reviews ChatGPT's ranking and then scores each participant from 1 to 10.
- If the highest-scoring candidate in the employers final ranking applies, this candidate will get the job, i.e., receive £16.
- If there is a tie among the highest scores, the job is assigned randomly among those with the highest score who applied.

Do you want to apply?

- Yes, I would like to apply (winning £16 if I rank first, £0 otherwise).
- No, I do not want to apply (winning £4 for sure, independent of my ranking).

---------- NEW SCREEN ----------

Case: ChatGPT decides with employer advice

ChatGPT, an artificial intelligence (AI), will determine a ranking with advice from an employer. The employer will determine their ranking first based on the application files, gender, age, and education level of you and the others. This ranking serves as an advice for ChatGPT. Subsequently, ChatGPT determines a ranking based on the advice, the application files, gender, age, and education level of you and the others. There is a 50% chance that you will be matched to a male or a female employer.

- The employer scores each participant from 1 to 10 (1 = completely unsuitable, 10 = completely suitable).
- ChatGPT reviews the employer's ranking and then scores each participant from 1 to 10.
- If the highest-scoring candidate in ChatGPT's final ranking applies, this candidate will get the job, i.e., receive £16.
- If there is a tie among the highest scores, the job is assigned randomly among those with the highest score who applied.

Do you want to apply?

- Yes, I would like to apply (winning £16 if I rank first, £0 otherwise).
- No, I do not want to apply (winning £4 for sure, independent of my ranking).

---------- NEW SCREEN ----------

Post-Experiment Survey

After completing all application decisions, participants answered questions about their beliefs and perceptions. Every question was presented on one single screen. The exact wording of the questions was as follows:

Expected Evaluation Scores: What evaluation score (1 = completely unsuitable, 10 = completely suitable) do you think you will receive when your application file is assessed by...? [ChatGPT / Employer / ChatGPT with employer advice / Employer with ChatGPT advice]

Perceived Objectivity: What do you think how objective will the evaluation score of an application file be when assessed by
ldots? (1 = not at all objective; 10 = very objective) [ChatGPT / Employer / ChatGPT with employer advice / Employer with ChatGPT advice]

Anticipated Harm: Assume in the following that you applied for the job under the same conditions as described in this experiment. How much would it hurt you (0 = not at all; 10 = very much) if you were not hired afterward because
ldots [ChatGPT decided against you? / Employer decided against you? / ChatGPT with employer advice decided against you? / Employer with ChatGPT advice decided against you?]

Stereotyping Concerns: To what extent do you think that ChatGPT is influenced by stereotypical ideas related to gender in recruiting decisions? (0 = not at all; 10 = very strongly)

To what extent do you think that the other participants of the experiment in the role of the employer are influenced by stereotypical ideas related to gender in recruiting decisions? (0 = not at all; 10 = very strongly)

Trust: How much trust do you place in ChatGPT regarding recruiting decisions (0 = no trust at all; 10 = very high level of trust)?

How much trust do you place in human employers regarding recruiting decisions (0 = no trust at all; 10 = very high level of trust)?

Risk Preferences: How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? (0 'not at all willing to take risks'
ldots10 'very willing to take risks')

AI Usage and Appreciation: How often do you use ChatGPT in everyday life (0 = never; 10 = very often)?

How much do you appreciate the use of ChatGPT in employment decisions because of
ldots (0 = not appreciate at all; 10 = appreciate very much) [its impartiality? / its lack of transparency in decision making? / its morality? / your confidence in algorithmic decisions? / your joy being judged by an algorithm?]

Digital Skills: How would you rate your general proficiency in digital skills? (0 = not proficient at all; 10 = highly proficient)

B.2 Employer Experiment

Participants in the employer role evaluated candidate applications to determine hiring outcomes for the candidate experiments.

Welcome

You are taking part in an academic study on economic decision-making.

Participants will receive a £1 base reward. A higher payment may result from bonuses. These bonuses result from your interactions and decisions, as well as those of other participants. Bonuses will be calculated and distributed after all participants have finished the experiment, which might take a few days. Participation is voluntary, and you may exit at any time; however, completion of the entire experiment is required for payment.

If you have any questions about this study or need further information, please feel free to contact us.

By agreeing and proceeding, you consent to participate and agree that your anonymized data may be used for research purposes. If you choose not to participate, please close the browser tab.

I agree to anonymized use of the data collected for research purposes.

- Yes, I agree

---------- NEW SCREEN ----------

Short Questionnaire

What is your age? [Open text field]

What is your gender?

- Female
- Male

What is the highest level of education you have completed?

- No formal education
- Primary education
- Secondary education / High school diploma
- Bachelor's degree

- Master's degree
- Doctoral degree
- Professional degree (e.g., MD, JD)

---------- NEW SCREEN ----------

Assessment and employment of applicants

In this experiment, you participate in the role of an employer. You will be matched with a group of four participants, called candidates, who have completed a math task.

The task involved adding combinations of two two-digit numbers (e.g., 64 + 27) within a 10-second limit, with a total duration of 60 seconds.

Each candidate earned 40 pence for each correct answer.

They also completed a summary questionnaire covering:

- Motivation
- Skills
- Guess of their own performance in the math task

The exact questions of the summary questionnaire were:

- *Motivation:* Please describe briefly your motivation to perform well in the math task.
- *Skills:* Please briefly describe your strategy to be as successful as possible in correctly solving many math tasks (e.g., your focus: speed vs. accuracy during the task, etc.).
- *Guess of Own Performance:* Please guess how many sums you have correctly calculated in the math task.

Additionally, candidates were informed to answer the questionnaire as if it was part of a job application file written to convince potential employers to hire them. They knew their application files would be rated and compared to the application files of three other candidates who participated in the same math task. Their application would be rated on a scale from 1 (completely unsuitable for a job based on this task) to 10 (completely suitable for a job based on this task).

They were told that they would be paid based on their evaluation score, calculated as: Score x 60 pence.

---------- NEW SCREEN ----------

Comprehension Check

To ensure you understand the task and evaluation process, please answer the following questions based on the instructions provided.

How much does a candidate earn for each correct answer in the math task?

- 16 pence
- 30 pence
- 40 pence
- Nothing

How is a candidate's payment calculated based on their evaluation score?

- Score x 10 pence
- Score x 60 pence
- Score x 14 pence
- There is no payment for a higher score

---------- NEW SCREEN ----------

How Candidates' Ranking Determines Who You Hire

The rating of the candidates determines their ranking among the four candidates.

The top-ranked candidate in the relevant ranking will be hired by you.

You will be paid a piece rate of 40 pence for each math problem that the hired candidate has solved correctly when they participated in the past task.

In total, four rankings will be created in the experiment.

- Two rankings are determined by ChatGPT, an artificial intelligence (AI), whereas
- two rankings are determined by you.

A random draw will determine which of the four rankings becomes the relevant ranking that will determine the candidate hired by you.

---------- NEW SCREEN ----------

Your Task: Creation of the First Ranking

In what follows, you will create two rankings. After submitting your second ranking, you will be informed which of the four rankings becomes decisive. This ranking will determine both the candidate who will be hired (i.e., the top-ranked one in this ranking) and your subsequent payment.

To create the first ranking, you evaluate each of the four candidates in the group. In this case, you receive information on their application files that describes their motivations and behavior for doing the math task and information on gender, age, and education level.

You will assess their suitability for doing this task and assign scores on a 1-10 scale for each candidate to create a ranking.

You are not allowed to assign your highest score to more than one candidate.

---------- NEW SCREEN ----------

Comprehension Check

To ensure you understand the task and evaluation process, please answer the following questions based on the instructions provided.

How much will you be paid for each math problem that the hired candidate solved correctly in the past task?

- 30 pence
- 40 pence
- 1 pound
- Nothing

How many candidates can receive the highest rating in your ranking?

- Only one candidate
- No limit on the number of candidates

---------- NEW SCREEN ----------

Note: For each of the four candidates, employers saw the following information and provided a 1-10 suitability rating. Gender information was included for employers in the Gender Information treatment and omitted for employers in the No Gender Information treatment.

Candidate #1

- Age:
- Gender:
- Education level:
- Motivation:
- Skills:
- Guess of Own Task Performance:

Please evaluate this candidate and their application file with regard to the suitability for a job based on the math task:

- 1 = completely unsuitable
- 2 through 9 (intermediate values)
- 10 = completely suitable

[This format repeated for Candidates #2, #3, and #4 with corresponding field variables]

---------- NEW SCREEN ----------

Your Task: Creation of the Second Ranking

Next, you again evaluate the same group of four candidates. In this case, you receive information on the evaluation score of the candidates that was made by ChatGPT and you again receive the information on their application files. You will assess their suitability for doing this task and assign scores on a 1-10 scale to create a ranking.

For each of the four candidates, employers saw:

Candidate #1

- Age:
- Gender:
- Education level:
- Motivation:
- Skills:
- Confidence:
- **Evaluation of ChatGPT:**

Please evaluate this candidate and their application file with regard to the suitability for a job based on the math task:

- 1 = completely unsuitable
- 2 through 9 (intermediate values)
- 10 = completely suitable

[This format repeated for Candidates #2, #3, and #4 with corresponding field variables including ChatGPT evaluation scores]

Post-Experiment Survey for Employers

After completing both rankings, employers answered the following survey questions:

Information Source Importance

Please rank the importance of the three information sources when you evaluated the candidates (where 1 is the most important and 3 is the least important):

- Candidates' Gender
- Candidates' Age
- Candidates' Education Level

Objectivity Perceptions

What do you think how objective will the evaluation score of an application file be when assessed by...? (0 = not at all objective; 10 = very objective)

- ChatGPT
- an employer
- ChatGPT with employer advice
- an employer with ChatGPT advice

Perceived Harm from Rejection

Assume in the following that you applied for the job under the same conditions described in this experiment. How much would it hurt you (1=not at all; 10 = very much) if you were not hired afterwards because...?

- an employer decided against you
- ChatGPT decided against you
- an employer with ChatGPT advice decided against you
- ChatGPT with an employer advice decided against you

Expectations About Information Source Importance

We will ask employers and ChatGPT to rank the importance of three information sources when evaluating candidates, where 1 is the most important and 3 is the least important: Candidates' Gender, Candidates' Age, Candidates' Education Level.

What do you think other participants in the role of employers will rank as the most important, second most important, and least important information sources?

What do you think ChatGPT will rank as the most important, second most important, and least important information sources?

Beliefs About Mean Evaluation Scores by Gender

What do you think is the mean score a female candidate of this experiment received when evaluated by an employer? [Scale 1-10]

What do you think is the mean score a male candidate of this experiment received when evaluated by an employer? [Scale 1-10]

What do you think is the mean score a male candidate of this experiment received when evaluated by ChatGPT? [Scale 1-10]

What do you think is the mean score a female candidate of this experiment received when evaluated by ChatGPT? [Scale 1-10]

What do you think is the mean score a female candidate of this experiment received when evaluated by an employer with ChatGPT advice? [Scale 1-10]

What do you think is the mean score a male candidate of this experiment received when evaluated by an employer with ChatGPT advice? [Scale 1-10]

What do you think is the mean score a female candidate of this experiment received when evaluated by ChatGPT with an employer advice? [Scale 1-10]

What do you think is the mean score a male candidate of this experiment received when evaluated by ChatGPT with an employer advice? [Scale 1-10]

Stereotyping

To what extent do you think that ChatGPT is influenced by stereotypical ideas related to gender in recruiting decisions? (0 = not at all; 10 = very strongly)

To what extent do you think that the other participants of the experiments are influenced by stereotypical ideas related to gender in recruiting decisions? (0 = not at all; 10 = very strongly)

Risk Preferences

How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? (0 'not at all willing to take risks' ...10 'very willing to take risks')

ChatGPT Usage and Trust

How often do you use ChatGPT in everyday life (0 = never; 10 = very often)?

How much trust do you place in ChatGPT regarding recruiting decisions (0 = no trust at all; 10 = very high level of trust)?

How much trust do you place in human employers regarding recruiting decisions (0 = no trust at all; 10 = very high level of trust)?

Appreciation of ChatGPT Use

How much do you appreciate the use of ChatGPT in employment decisions because of... (0 = not appreciate at all; 10 = appreciate very much)

- ... its impartiality?
- ... its lack of transparency in decision making?
- ... its morality?
- ... your confidence in algorithmic decisions?
- ... your joy being judged by an algorithm?

Digital Skills

How would you rate your general proficiency in digital skills? (0 = not proficient at all; 10 = highly proficient)

B.3 ChatGPT Evaluation Implementation

We used OpenAI's GPT-4o API (with temperature=0 for deterministic outputs) to evaluate candidate applications. The evaluation script processed candidates in groups of four, dynamically constructing prompts that included candidate application files and demographic information.

Script Structure The Python script performed the following operations:

1. Loaded candidate data from CSV files containing application responses, demographics, and group assignments
2. For each employer-group combination, extracted four candidates (skipping incomplete groups)
3. Ran two evaluation rounds per group:
    - **Round 1 (ChatGPT only):** Evaluated candidates based on application files and demographics without human ratings
    - **Round 2 (ChatGPT with employer advice):** Re-evaluated the same candidates including human employer ratings
4. Sent prompts to the OpenAI API and parsed the returned scores

The script included two experimental conditions varying the information provided: one where gender information was shown to ChatGPT (genderinfo=1) and one where it was withheld (genderinfo=0). Age and education level were always included.

Complete Evaluation Prompt The full prompt sent to ChatGPT consisted of three components: (1) task context and payment structure, (2) available candidate information fields, and (3) candidate profiles with response instructions. The exact prompt text was:

> `In this experiment, you participate in the role of an employer. You will be matched with a group of four participants, called candidates, who have completed a math task.`
>
> `The task involved adding combinations of two two-digit numbers (e.g., 64 + 27) within a 10-second limit, with a total duration of 60 seconds.`
>
> `Each candidate earned 40 pence for each correct answer.`
>
> `They also completed a summary questionnaire covering:`
>
> - `Motivation`
> - `Skills`
> - `Guess of their own performance in the math task`
>
> `The exact questions of the summary questionnaire were:`

1. Motivation: Please describe briefly your motivation to perform well in the math task.
2. Skills: Please briefly describe your strategy to be as successful as possible in correctly solving many math tasks (e.g., your focus: speed vs. accuracy during the task, etc.).
3. Guess of Own Performance: Please guess how many sums you have correctly calculated in the math task.

Additionally, candidates were informed to answer the questionnaire as if it was part of a job application file written to convince potential employers to hire them. They knew their application files would be rated and compared to the application files of three other candidates who participated in the same math task. Their application would be rated on a scale from 1 (completely unsuitable for a job based on this task) to 10 (completely suitable for a job based on this task).

They were told that they would be paid based on their evaluation score, calculated as: Score x 60 pence.

Additionally, they provided:

- Age
- Gender *[included only in gender information condition]*
- Human Rating *[included only in Round 2 with employer advice]*
- Education

Your task: Please read each candidate's application file and assign a score from 1 (completely unsuitable) to 10 (completely suitable).

Important Rules:

- Assign scores as whole numbers (integers) between 1 and 10.
- Ensure that no group has more than one candidate with the highest score.

[For each candidate A, B, C, D, the following profile was inserted:]

Candidate [Letter]:

Motivation: [motivation response]

Skills: [skills response]

Confidence (guessed correct answers): [confidence response]

Age: [age]

Gender: [gender] *[included only in gender information condition]*

Education: [education level]

Human Rating: [human employer score] *[included only in Round 2]*

Respond only with:

Candidate A: [score]

```
Candidate B: [score]
Candidate C: [score]
Candidate D: [score]
```

Implementation Notes The dual-evaluation design allowed us to create two distinct AI evaluation conditions. In Round 1, ChatGPT evaluated candidates independently based solely on application content and demographics, implementing the "ChatGPT decides" condition. In Round 2, ChatGPT received human employer ratings before making its assessment, implementing the "ChatGPT decides with employer advice" condition. This structure enabled direct comparison of AI decision-making with and without human input, while maintaining identical candidate information across both rounds. The gender information manipulation (present vs. absent) was crossed with these two rounds to examine whether demographic information availability affects AI evaluation patterns.