Discrimination in the Large and Small *

Ian Chadd[†], Mike Shor[‡]

This version: October 26, 2026

Click here for the most recent version.

Abstract

Hiring managers routinely choose from applicant pools ranging from a handful of candidates to hundreds. Yet, most experiments on discrimination examine choices between just two candidates, leaving unexplored the question of how the size of the choice set affects discrimination. In a pre-registered experiment, subjects act as managers, choosing candidates from pools of either four or sixteen. Candidate gender is displayed directly while ability is available only through a signal that is harder to process. Across all subjects, increasing the number of candidates shifts the female hiring rate from 52 to 49 percent, a pro-female bias in small pools that disappears in large ones. A pro-male bias emerges in larger pools among male managers, managers with high search costs, and managers with hostile attitudes toward women. Click-level search data shows managers reveal men's and women's ability signals equally often and in roughly the same order in both pool sizes. Instead, discrimination arises in how managers use the revealed information. The probability that a manager hires a woman when she is the best considered candidate decreases in large choice pools. Sequential search, satisficing, and rational inattention with a gender-biased prior each predict the direction of the main effect, but are inconsistent with the search data. The evidence is consistent with managers switching decision heuristics as pools grow, relying on a pro-male bias only in large choice sets.

JEL Codes: J71, J16, C91, D83, D91
Keywords: gender discrimination; choice overload; cognitive load; heuristics

*Declarations of interest: none.

[†]Department of Economics, Rensselaer Polytechnic Institute, Email: chaddi@rpi.edu
[‡]Department of Economics, University of Connecticut, Email: mike.shor@uconn.edu

1 Introduction

Applicant pools range from a handful to thousands of applicants. How do hiring decisions change as the size of the pool changes? In common models of costly choice, a manager who searches (Weitzman, 1979), satisfices (Simon, 1955), or rations attention (Matějka and McKay, 2015) may examine only a sample of a larger pool, and may examine candidates less closely, relying more on prior beliefs in larger choice sets. Any bias in that prior is present at every pool size but carries more weight as the pool size grows. The psychology of stereotyping (Fiske and Neuberg, 1990; Bodenhausen and Lichtenstein, 1987) and research on choice overload (Iyengar and Lepper, 2000; Payne et al., 1988; 1993; Besedeš et al., 2012b) point to a different mechanism. As cognitive load increases, decision makers rely on simpler rules that favor whatever information is easiest to process. In hiring, the easiest information may be demographic. Therefore, a larger pool can change who gets hired in two ways: by changing which candidates a manager considers, or by changing how a candidate is chosen from among those considered. Either can introduce an effect of choice-set size on discrimination.

We ask whether gender discrimination in hiring changes when managers choose from larger pools, and whether such changes are driven by which candidates are considered or by how they are evaluated. The question matters because applicant pool sizes vary enormously, from one to over four thousand applicants per posting in the LinkedIn data of Gee (2019), while the experimental literature on discrimination rarely asks subjects to choose among more than two or a handful of candidates (Neumark, 2018). Field evidence suggests that gender matters more at early stages of hiring, where pools of applicants are large. Fernandez-Mateo and King (2011) document that candidate gender matters more in initial screening than in later stages among temporary staffing agencies. Fernandez and Campero (2017) find that gender gaps in tech hiring are concentrated in large applicant pools and disappear by the interview stage. Botelho and Abraham (2017) find that investment recommendations made by women are less likely to be viewed than those made by men, particularly when there are many recommenders to choose from. Jiang et al. (2024) report more discrimination against women when the ratio of job-seekers to positions is high. These studies rely on naturally occurring variation in pool size, and thus cannot separate its effect from everything else that differs across hiring stages, markets, and business cycles. Additionally, pool size in the field is endogenous to anticipatory sorting by candidates, to a firm's choice of stated qualifications, and to how many candidates managers advance to each stage. We isolate the effect of pool size experimentally.

In a pre-registered online experiment, participants act as hiring managers who choose one candidate from a pool of applicants and are paid according to the chosen candidate's ability. Each manager makes twenty hiring decisions, some "small decisions" from pools of four candidates and some "large decisions" from pools of sixteen. We design the information available about each candidate to mirror two features of the field. Hiring managers have only limited information about an applicant's ability. The information that they do have differs in how easy it is to process. A demographic characteristic like gender can be quickly inferred from a name, but ability is more difficult to assess and must be inferred from experience, recommendations, writing samples, or test scores. It is the fact that relevant information is both complex and noisy that may lead managers to rely on easily observed but irrelevant demographic data when a large pool makes evaluation cognitively costly.

Accordingly, a candidate profile contains a randomly generated ID number and only two char-

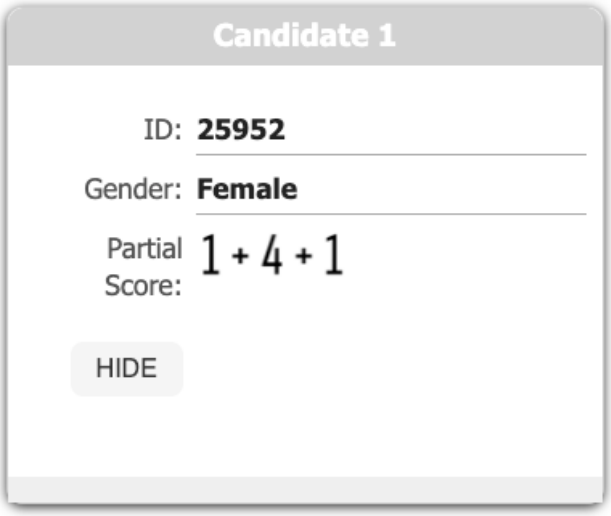


Figure 1: Example candidate profile

Notes: A candidate profile as shown to managers, with the Partial Score revealed. Each profile has a randomly generated five-digit ID, the candidate's gender, and a Partial Score: the number correct among 10 randomly chosen questions from the candidate's 20-question math and science quiz, shown as an unsolved arithmetic expression (here, $1 + 4 + 1 = 6$). Managers are paid for the candidate's Total Score on all 20 questions, so the Partial Score is a noisy signal of ability. It is hidden until the manager clicks to reveal it, and only one can be displayed at a time.

acteristics (Figure 1). Gender is stated directly as "male" or "female." Ability, defined as a candidate's score on a 20-question math and science quiz (Exley and Kessler, 2022), is conveyed by a noisy signal. Managers see a "Partial Score," the number correct among 10 randomly chosen questions on the quiz, presented as an unsolved arithmetic expression. A manager must solve "$1+4+1$" to assess that the candidate in Figure 1 has a Partial Score of 6, imposing non-trivial cognitive costs (as in Caplin et al., 2011; Chadd et al., 2025). The Partial Score is hidden until the manager clicks to reveal it, and only one candidate's Partial Score can be displayed at a time. We observe which candidates a manager examines, in what order, and for how long. Decision problems are constructed so that a female candidate has the highest Partial Score in exactly half of them, both overall and by pool size, so that any gender-blind decision rule hires female candidates half the time. Deviations from fifty percent measure gender bias.

We present results on both decisions made and the process to make them.

First, across all participants, female candidates are hired in 52.2 percent of small decisions, significantly above parity, and in 49.0 percent of large ones. This roughly six percent reduction is robust to relaxing our exclusion criteria, to demand-effect controls, and to conditional-logit specifications with choice-set fixed effects. The effect is larger for male managers. In small decisions, male and female managers alike hire female candidates about 52 percent of the time. In large decisions, female managers hire female candidates at parity (50.7 percent) while male managers hire them significantly below it (47.3 percent, a ten percent reduction). The reduction is also larger for managers with high search costs measured by an incentivized arithmetic task (10 percent) and managers with hostile attitudes toward women (20 percent). Beliefs about gender differences in ability do not moderate the effect.

Second, the effect does not run through which candidates are considered. Managers consider

more candidates in large decisions and spend more time in total, but less time per candidate. Large decisions also produce more decisions with no candidates' scores revealed and more with incomplete search. None of this differs by candidate gender. The probability that a candidate's Partial Score is revealed is the same for male and female candidates in small and large decisions. Female candidates are considered slightly earlier than male ones in large decisions, but controlling for this does not shrink the effect. Instead, it runs through use of candidate information. Conditional on the candidates a manager examined and their Partial Scores, a female candidate is more likely to be hired in small decisions and less likely in large ones.

Section 2 discusses what these patterns mean for theory. Sequential search, satisficing, and rational inattention each become models of discrimination once the manager's prior over ability favors men, and each then predicts that larger pools produce more male hires. Each also has implications for the choice process that are difficult to square with our data. More fundamentally, all three models predict a consistent bias that grows with pool size, not a bias that reverses from pro-female to pro-male, as it does among male managers. Our results are consistent with the choice-overload literature (Payne et al., 1988; 1993; Besedeš et al., 2012a;b; Lenton and Francesconi, 2010) which suggests that decision makers facing more options do not apply the same rule less thoroughly but switch to a simpler rule.

To our knowledge, we are the first to experimentally vary choice set size to investigate whether discrimination is related to choice overload, in contrast to the single-candidate vignettes, matched pairs, and small choice sets of most discrimination experiments (see, e.g., Neumark, 2018). We contribute to a small but growing literature on the relationship between discrimination, search, and cognitive load (Bartoš et al., 2016; Lane et al., 2024; Abel and Burger, forthcoming). While Bartoš et al. (2016) find discrimination in what subjects consider, we find "equal exposure, unequal use," where large choice sets change not what candidates are examined but which of the examined candidates is selected.

The rest of the paper is organized as follows. Section 2 positions the paper relative to the experimental literature on discrimination and derives the process predictions of choice models. Section 3 describes the design. Section 4 presents the results on the hiring rate, its moderators, and patterns of search and use. Section 5 concludes and discusses implications for theory, experiments, and practice.

2 Conceptual Framework

2.1 Choice-set size in experiments on discrimination

Most experimental studies of discrimination do not observe the size of the choice set. The literature surveyed by Neumark (2018) relies heavily on methods such as correspondence studies, in which the characteristics of the applicant pool are not observed. Some observational studies do observe naturally occurring choice sets and find discrimination against women (Moreno et al., 2012; Goldin and Rouse, 2000). Closer to our question, Lane et al. (2024) show that recruiters for a tech training program are more likely to contact male prospects as their workload increases, but cannot separate the number of prospects from time pressure, and recruiters are rewarded for converting prospects into applicants and enrollees rather than for identifying the best one. Controlled experiments that do observe the choice set, however, typically ask participants to choose among very few candidates, often just two (e.g., Correll et al., 2007; Krings et al., 2011; Reuben et al., 2014; Hedegaard and Tyran, 2018). This remains true in recent work (e.g., Mourelatos, 2023; Campos-Mercade and

Mengel, 2024; Fabregas and Zanoni, 2025; Barron et al., 2025; Abel and Burger, forthcoming).[1]

A few experiments use larger choice sets, but their designs differ from ours. Charness et al. (2020) use two-participant "firms" choosing from five candidates and find no substantial discrimination against women. Fonseca and McCrea (2025) have participants choose from ten candidates, but participants determine the gender composition of the pool. Kessler et al. (2019; 2026) present employers with forty candidates but ask for ratings of each rather than a hiring decision. We are unaware of a controlled hiring experiment in which participants choose from sets as large as our 16-candidate decisions, even though applicant pools in the field are routinely far larger.

2.2 Three benchmark models

A manager who must pick the best candidate when ability is costly to learn and gender is free to observe faces a problem economists have modeled in three ways. She may search sequentially, revealing scores one at a time and stopping when the best score so far exceeds the reservation value of the best unrevealed candidate (McCall, 1970; Weitzman, 1979). She may satisfice, accepting the first candidate whose score clears an aspiration level that rises with the size of the set (Simon, 1955; Caplin et al., 2011). Or she may allocate a limited budget of attention across candidates and choose on the basis of noisy, endogenously precise signals (Sims, 2003; Matějka and McKay, 2015; Caplin et al., 2019). Each becomes a model of discrimination once the manager's prior over ability favors men (Phelps, 1972; Arrow, 1973), and each then predicts that a larger pool produces more male hires: in search, because men's higher reservation values put them first in line and a larger pool makes it likelier that a man clears the bar before any woman is examined; in satisficing, because a rising bar is cleared more easily by candidates who receive a premium; in rational inattention, because a larger pool raises the opportunity cost of learning about any one candidate and shifts weight onto the free signal (Bartoš et al., 2016).

As the three models agree on the direction of the aggregate effect, the hiring rate alone cannot distinguish them. While the models disagree about the process through which discrimination occurs, the disagreements are testable with data on which candidates were considered, in what order, for how long, and who was ultimately chosen. Search models predict that male candidates are considered before female candidates and that the female share of the consideration set falls with pool size, but that, conditional on the consideration set, the manager picks the highest score without regard to gender. Satisficing predicts that the order of consideration is unrelated to gender, but that hired male candidates have lower scores than hired female candidates, because the threshold is lower for them. Rational inattention predicts that female candidates receive less attention as pool size grows and that any gender bias in choice shrinks with the time a manager considers a candidate. All three predict that a manager who considers no one picks a male candidate, at any pool size. And all three predict a bias in the same direction in small and large decisions, though of different magnitudes, but none predicts a reversal from preferring female candidates to preferring male candidates. Table 1 summarizes these predictions.

2.3 Heuristic switching

A fourth possibility that the models above exclude by construction is that the manager applies one rule in small decisions and a different rule in large ones. Two literatures suggest this hypothesis.

[1] Castillo and Petrie (2010) is an exception: participants submit an incentivized ranking of 19 candidates and exhibit racial discrimination when other information is absent.

Table 1: Predictions of the benchmark models

	Search	Satisficing	Rational inattention
Predictions for Choice Data			
1. Direction of gender bias	Favors male candidates; more so in large decisions	Favors male candidates; more so in large decisions	Favors male candidates; more so in large decisions
2. Choice conditional on the consideration set	Highest score; no gender effect	Favors male candidates, more so in large decisions	Favors male candidates; less so the longer a candidate is considered
3. Partial Score of hired candidates	Same for male and female candidates	Lower for male candidates	Lower for male candidates
4. Choice when no candidate is considered	Male candidate	Male candidate	Male candidate
Predictions for Choice Process Data			
5. Female share of the consideration set	Below one half; lower in large decisions	One half	Below one half; lower in large decisions
6. Order in which candidates are considered	Male before female	Independent of gender	Male candidates earlier; more so in large decisions

Notes: Qualitative predictions of each benchmark model for a manager whose prior over ability favors men, choosing from a set that is half male and half female candidates, with a costless gender signal and a costly ability signal. The consideration set is the set of candidates whose Partial Scores the manager views. Because a female candidate has the highest Partial Score in exactly half of decisions, a gender-blind manager hires women half the time. "Favors male candidates" means more male hires than a gender-blind manager with the same information would make. "More so in large decisions" is relative to small decisions. The models are described in Section 2.

The first is the psychology of stereotypes under cognitive load. Since at least the 1950s, psychologists have hypothesized that stereotypes emerge from the cognitive process of categorization, which helps people simplify complex information (Allport, 1954; Taylor et al., 1978; Fiske and Neuberg, 1990; Kahneman, 2011). Accordingly, stereotypes are more likely to be used when cognitive resources are strained. Increased reliance on stereotypes has been observed when multitasking (Gilbert and Hixon, 1991; Macrae et al., 1994), when lacking sleep (Ma et al., 2013), at a low point in the circadian rhythm (Bodenhausen, 1990), as task complexity increases (Bodenhausen and Lichtenstein, 1987), under memory load (Rothbart et al., 1978), under time pressure (Kruglanski and Freund, 1983; Pratto and Bargh, 1991), and following a separate demanding task (Govorun and Payne, 2006). Closest to our setting, Abel and Burger (forthcoming) find that time pressure in a two-candidate hiring choice increases reliance on race. The continuum model of Fiske and Neuberg (1990) suggests that under capacity pressure, people rely more on category-based than individuating information.

The second is the literature on choice overload and heuristic switching. An abundance of options can lead to worse choices, decision paralysis, and lower satisfaction (Iyengar and Lepper, 2000; Schwartz, 2004; Caplin et al., 2011; Chadd et al., 2021; Townsend and Kahn, 2014). The

adaptive-decision-maker framework of Payne et al. (1988; 1993) suggests that as the number of alternatives grows, people may switch to simpler decision rules that require less effort, including changing which attributes they consider or how they weight them. Besedeš et al. (2012a;b) show that such a switch reduces the quality of decisions from larger choice sets in a task modeled on prescription drug plan selection. Lenton and Stewart (2008) and Lenton and Francesconi (2010) document the same switch in mate choice, with larger choice sets shifting weight toward easily assessed attributes such as height. Cunow et al. (2021) find that voters are less likely to consider candidates' positions as the number of candidates on a ballot grows and are more likely to vote for the first name listed.

Heuristic switching is not a formal model in the sense of the three above. It does, however, have two testable implications. First, if switching reflects cognitive strain, its effects should be concentrated among managers for whom the large set is demanding. We measure each manager's cost of processing the ability signal in an incentivized arithmetic task. Second, cognitive load weakens the control people exercise to suppress their use of stereotypes (Devine, 1989; Govorun and Payne, 2006; Kunda and Spencer, 2003). Because the stereotypes that remain as control weakens are more pronounced among individuals with more hostile attitudes (Lepore and Brown, 1997), switching should have a larger effect for managers with such attitudes, and need not track their stated beliefs about ability. We measure both. This can also explain a reversal in the gender effect. In small decisions, managers may correct (and overcorrect) for stereotypes they do not endorse but be unable to do so under cognitive load. Beyond these predictions, heuristic switching does not specify the decision rules used in small and large choice sets. We therefore treat switching as the hypothesis that remains if the process data reject the benchmarks in Table 1, and test its two predictions.

3 Experiment

Our experiment is structured around a primary hiring task. In this stage of the experiment, participants act as "managers" who hire "workers" based on a limited amount of information contained in candidate profiles. Managers are provided with two pieces of information for each candidate: gender, which we conjecture is relatively easy to process, and a signal of the worker's ability that is more difficult to process. They are incentivized to select the highest-ability worker in each decision. The central treatment variation in the hiring experiment concerns the size of the choice set, that is, the number of applicants presented to a hiring manager in a given decision problem. Our main outcome of interest is the hiring rate of female versus male candidates in large versus small choice sets.

After completing the hiring task, subjects completed a math quiz to measure their ability to process worker performance information in candidate profiles. Finally, subjects completed a post-experiment survey that contained questions on demographics (e.g., gender and age) as well as on attitudes and beliefs about discrimination and gender stereotypes.

3.1 The Hiring Part

In the Hiring Part of our experiment, each participant takes on the role of a "manager" whose task is to choose a "worker" from a set of displayed candidates. They are incentivized to choose the worker they believe had the highest ability in a previously performed math and science task and must make this decision based on limited information contained in a candidate profile.

Candidate Profiles

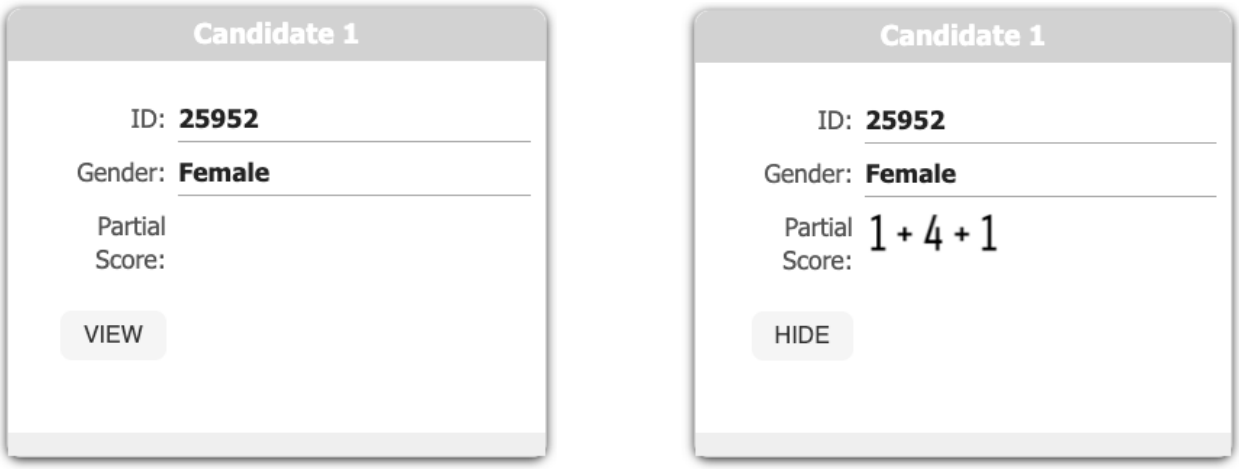
We carefully design candidate profiles to mirror elements of the relevant field setting while still retaining the simplicity and control afforded by an experiment. In the field, hiring managers choose among candidates equipped with only limited information regarding their fit. In our experiment, we operationalize this limited candidate information by sourcing worker characteristics using publicly available data from Exley and Kessler (2022). In their experiment, participants completed a math and science task modeled on the Armed Services Vocational Aptitude Battery, which is used to determine occupational eligibility in the United States Armed Forces. Participants answered 20 questions, had 30 seconds to respond to each, and were paid a piece rate for each correct answer. We treat the participant's score on this quiz as a measure of ability and pair it with their gender to create a simplified candidate profile for use in our experiment.

In the field, limited information such as this is often asymmetric in terms of its ease of processing: gender can be quickly inferred (e.g., from name) while ability is more difficult to assess (e.g., from experience, recommendations, writing samples, standardized assessments). It is exactly this tension that we conjecture may result in more discriminatory behavior in larger choice sets; choice overload may prompt managers to rely on gender and its associated stereotypes rather than information on ability, which is more pertinent but difficult to process.

We operationalize this in our experiment by presenting candidate gender and ability differently along two dimensions: noise and ease of processing. Gender is presented in a direct and easy-to-process manner (e.g., as a simple statement of "Male" or "Female"), as illustrated in Figure 2. In contrast, information about candidate ability is both noisy and more difficult to parse. As previously mentioned, we define candidate ability as their score in a 20-question math and science quiz, which we term their Total Score. Managers are incentivized to choose candidates they believe have the highest Total Score; they earn \$0.15 per correct answer completed by the candidate they select. However, to introduce noise in the assessment of candidate ability, the Total Score of a candidate is unobserved and a Partial Score is presented instead. Partial Scores are defined as the number correct among 10 randomly chosen questions on the quiz and so are positively correlated with, but do not perfectly predict, the Total Score. As such, this noisy signal of candidate ability leaves room for statistical discrimination to play a role: a manager may be less likely to choose a female candidate over a male with the same Partial Score if they believe that female candidates are lower-ability on average.

As we conjecture that managers may avoid difficult-to-process information, Partial Scores are presented in profiles as the result of a three-number arithmetic operation. For example, a candidate with a Partial Score of six may have it presented as "$1 + 4 + 1$," as is the case in Figure 2. This general approach is common in experiments wherein participants search through available options (e.g., Caplin et al., 2011; Chadd et al., 2025) and allows us to induce non-trivial cognitive search costs while maintaining control over the objective value thereof.

To enhance the naturalness of this environment, we pair candidate gender and Partial Score information with a randomized five-digit numeric ID included in each profile. As illustrated in Figure 2, candidate profiles are initially displayed with the Partial Score hidden. A manager can reveal a candidate's Partial Score by clicking a view button. This design choice is similar in fashion to Mouselab-style experiments (e.g., Gabaix et al., 2006) and enables us to trace what Partial Score information a manager views and when.



(a) Partial Score hidden

(b) Partial Score displayed

Figure 2: Candidate profile with the Partial Score hidden and displayed

Notes: Panel (a) shows a candidate profile as it first appears, with the Partial Score hidden. Panel (b) shows the same profile after the manager clicks the view button, which reveals the Partial Score as an unsolved arithmetic expression ($1 + 4 + 1 = 6$). Only one Partial Score can be displayed at a time.

Decision Problems

Managers make hiring choices in each of 20 decision problems. As we are primarily interested in whether discriminatory treatment gets worse in larger choice sets, the central variation of interest concerns the number of candidate profiles shown per decision. In small decision problems, managers see 4 profiles; in large ones they see 16. Regardless of the size, managers must select one candidate in each problem in order to move on.

Managers are randomized into one of two treatment arms which vary based on the proportion of small and large decision problems they encounter. In the More Small treatment, managers see 14 small problems and six large ones. In the More Large treatment, this is reversed. Variation in this dimension is done both within- and between-subject, with decision problems separated into two Blocks presented sequentially. Block 1 consisted of 12 decision problems, six small and six large, and was presented to all managers. Block 2 consisted of either eight small or eight large decision problems, depending on the treatment. Decision problem order was randomized at the subject level within each block.

Our central outcome of interest is the hiring rate of female candidates across small and large decision problems. For ease of analysis and interpretation of our findings, we construct decision problems such that discrimination against female candidates can be measured as the difference between 50% and the hiring rate. To facilitate this, we populate decision problems with candidate profiles randomly with several constraints, the primary of which is that exactly half of the decision problems have a female candidate with the highest Partial Score. This constraint is applied to the overall set of decision problems, to small (large) ones only, to Block 1 (Block 2) only, as well as to the set of problems included in the More Small (More Large) treatment arms. As a result, regardless of the cut of the data, discrimination can be measured as distance from a hiring rate of 50%.

When constructing these decision problems, we were concerned that treatment differences may emerge based on spurious differences in the gender distribution of Partial Scores excluding the first-

best in each problem. For example, managers may be less likely to find the highest Partial Score in large problems, disproportionately choosing the second-best. Our constraints above were not applied to every rank of candidate profile, leaving open the possibility that one gender is overrepresented in "good but not first-best" profiles, perhaps differentially by choice set size. To control for this confound, within each treatment arm we randomize managers into one of two decision problem data files: a Baseline one constructed as described above, and a Gender-Swapped one that retains all candidate profile Partial Scores but swaps the candidate gender. Our aggregate results pool across these data files to mitigate concerns regarding any spurious gender differences in Partial Score distributions excluding the first-best. A more detailed account of our data generation process can be found in Appendix A.

3.2 The Math Part

Our chosen design has the advantage of allowing us to construct a measure of manager search costs based on their facility in solving the arithmetic problems used for the presentation of Partial Scores. This serves two purposes. First, measuring the dependence of the extent and intensity of search (i.e., viewing Partial Scores) on math ability serves as a gut check to confirm that indeed large choice sets induce more cognitive load: managers who have an easier time processing Partial Score information should look at more of them. This is confirmed in Section 4.2. Second, it allows us to investigate the extent to which search cost heterogeneity moderates any treatment effect of large decisions, with the conjecture being that managers with lower search costs should be better able to navigate larger choice sets without exhibiting additional discriminatory behavior.

As such, after completing the Hiring Part of the experiment, managers completed an incentivized math quiz. They were asked to solve as many three-number addition and subtraction problems as possible within a 60-second window. Each correct response submitted within the time limit earned a bonus payment of $0.08.

3.3 Implementation

Managers were paid for both parts of the experiment. At the end of the experiment, one of the 20 decision problems in the Hiring Part was chosen at random and the manager was paid $0.15 times the Total Score of the candidate chosen in that problem. The bonus payment of $0.08 per correct response in the Math Part of the experiment was added to this sum, along with a $2.50 completion fee.

After pre-registering our experiment on AsPredicted.org (#289442), we collected responses from 509 subjects on Prolific in May 2026. The experiment took a median of 23 minutes to complete and participants were paid on average $5.43. Our treatments are balanced on a host of demographic variables (see Appendix Table C.1). Following our pre-registered criteria, 64 subjects are excluded from our analysis for failing attention checks or moving through the experiment too quickly. Nevertheless, as discussed in the following section, our primary finding is robust to the inclusion of these subjects.

4 Results

We begin by considering whether the hiring rate of female candidates depends on the size of the choice set. We find that it does: female candidates are chosen approximately six percent less frequently in large choice sets than in small choice sets. This "Large Decision effect" is concentrated among male hiring managers, who choose female candidates approximately ten percent less

frequently in large choice sets. We further show that this is robust to a number of alternative specifications, including alternative cuts of the data. We then investigate the extent to which the Large Decision effect is moderated by exogenous factors, namely participant characteristics measured independently of the choice task. We find that search costs and attitudes moderate it, but not beliefs. Finally, we consider whether large decisions induce endogenous changes to search behavior and whether these changes subsequently affect the propensity to hire female candidates. We find that while large decisions do induce more extensive but less intensive search, these patterns do not absorb the Large Decision effect on female candidates.

We further document that this effect emerges through "equal exposure, unequal use" of information regarding available candidates. While managers avail themselves of information equally across male and female candidates, they disproportionately pass over best-viewed females in large decisions to choose lower quality candidates. We find that this happens regardless of the cost of doing so, as measured by the gap in candidate Partial Scores between the best-viewed candidate and the best-viewed candidate of the other gender.

4.1 Choices

Are women hired less frequently in large decisions?
In this section, we investigate whether large decisions affect the propensity to hire female candidates, overall and separately for male and female managers. We find that this is the case, particularly for male managers. This Large Decision effect is robust to controlling for features of the sample (e.g., attentiveness) and alternative estimation procedures (e.g., conditional and mixed logistic regression).

Figure 3 presents the proportion of decision problems in which a female candidate was hired. The means are presented separately for small and large decisions, using the entire sample (Pooled) and separately for Female and Male managers. The overall proportion across all decisions is 50.60% (indicated by the horizontal dashed line in Figure 3) and the proportion is everywhere higher in small decisions than in large ones. The Large Decision effect is highest in magnitude among Male managers (4.87pp) and smallest among Female managers (1.58pp).

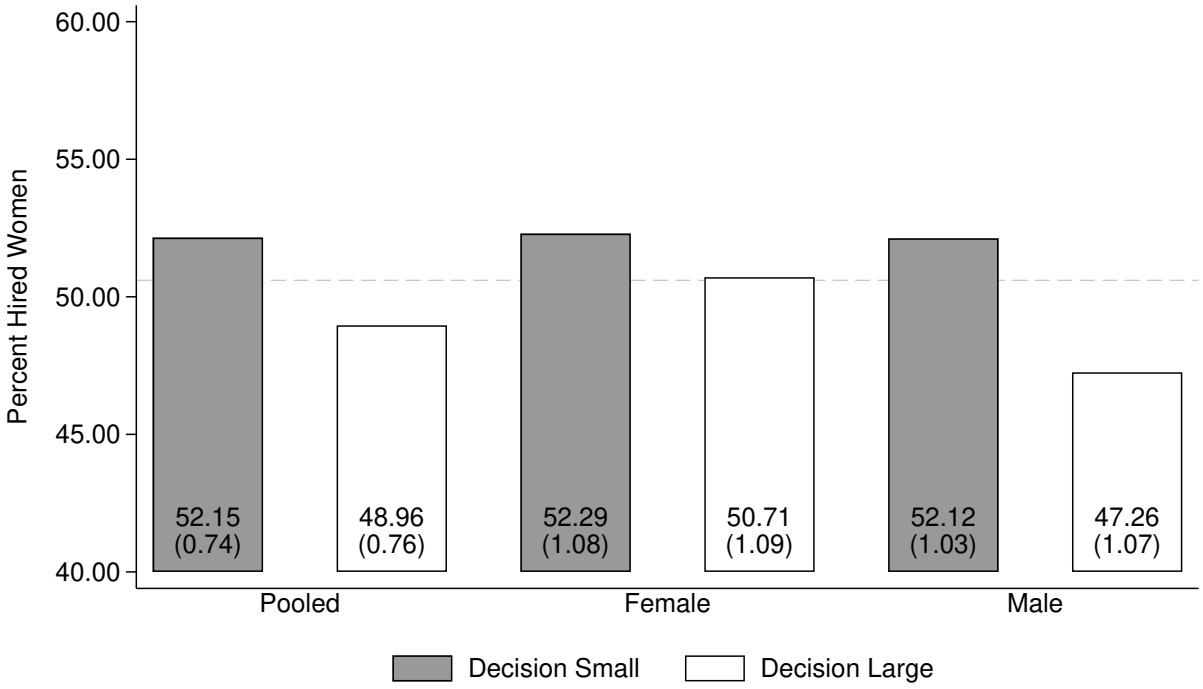


Figure 3: Female candidate raw hiring rates are lower in large decisions

Notes: Percent of decisions in which a female candidate was hired, for small decisions (4 candidates; grey bars) and large decisions (16 candidates; white bars), pooling all managers (Pooled) and separately for female and male managers. Labels give each mean with its standard error in parentheses. The sample is 8,900 decisions (4,558 small, 4,342 large) from 445 managers. By design, a gender-blind manager hires a female candidate in 50% of decisions. The dashed line marks the overall rate, 50.60%.

Given that we elicited multiple observations ($n = 20$) per manager, we test for the significance of the Large Decision effect via logistic regression with robust standard errors clustered at the subject level. The marginal effects are presented in Figure 4 for the Pooled sample, as well as separately for Female and Male managers.[2] As illustrated in Figure 4 these margins are significantly different from zero for the Pooled ($p < 0.001$) and Male ($p < 0.001$) samples, but not the Female sample ($p = 0.248$).[3] Taken together, these results indicate that indeed large decisions affect the propensity to hire female candidates, predominantly for male respondents.[4]

[2]Table C.2 in the Appendix reports the log-odds coefficients for these and other supporting logistic regressions.

[3]As a check that does not rely on decision-level regressions, we use only Block 1 (the first 12 decisions, six small and six large for every manager) and compute each manager's share of female hires in small and in large decisions. Both a paired t-test and an exact Wilcoxon signed-rank test of the within-manager difference between the two shares give $p < 0.001$ for the Pooled and Male samples, and $p = 0.084$ and $p = 0.076$, respectively, for the Female sample. The small-decision share is significantly above 0.50 in all three samples (t-test, signed-rank: Pooled $p = 0.002$, $p < 0.001$; Female $p = 0.024$, $p = 0.005$; Male $p = 0.021$, $p < 0.001$). The large-decision share is below 0.50 by t-test for the Pooled ($p = 0.033$) and Male ($p = 0.006$) samples but not the Female sample, and by signed-rank test in none.

[4]This analysis includes all decisions, importantly those where the manager chose the candidate with the highest Partial Score (62.92% of all observations), among which there will be no gender bias by design. Figure C.1 in the Appendix reports the raw means and average marginal effects conditional on sub-optimal choice. These results are directionally similar, though higher in magnitude (e.g., -5.69pp and -8.35pp for the Pooled and Male samples,

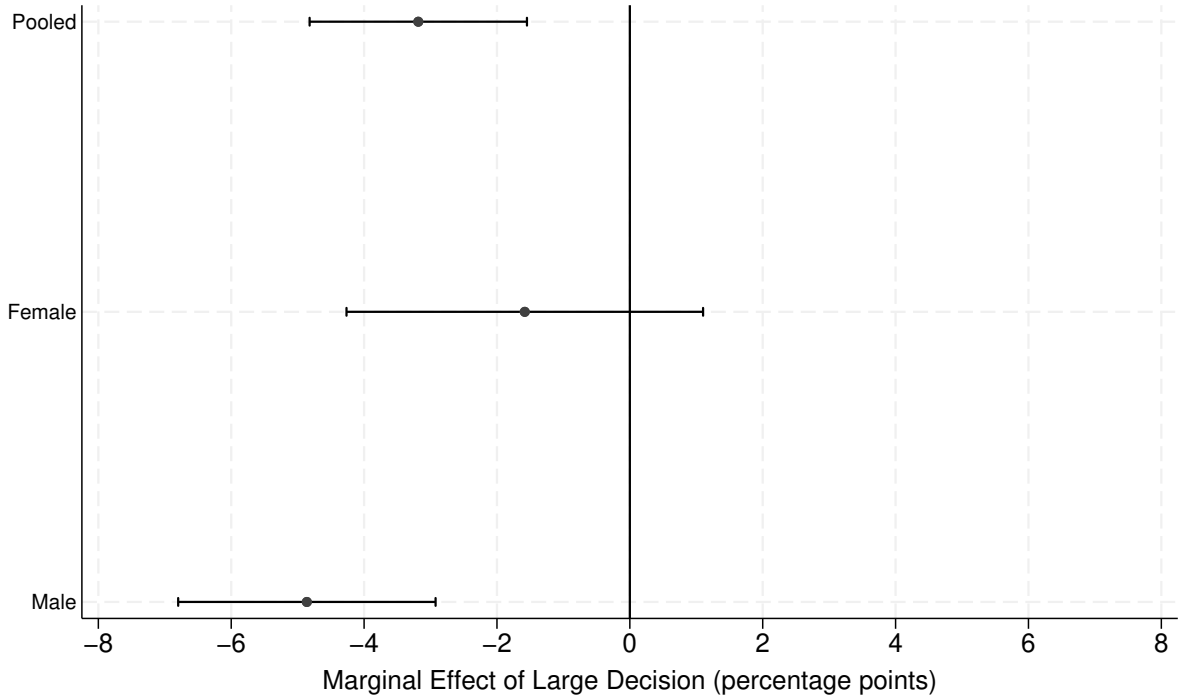


Figure 4: Regression-adjusted Large Decision effect present for Pooled and Male samples

Notes: Average marginal effect of a large decision (16 candidates) relative to a small decision (4 candidates) on the probability of hiring a female candidate, in percentage points, with 95% confidence intervals. Estimates come from logistic regressions of *Choice Female* (one if a female candidate was hired, zero otherwise) on *Large Decision* only, with standard errors clustered by subject, for all managers (Pooled) and separately for female and male managers. The Pooled sample is 8,900 decisions from 445 managers; the Female and Male samples exclude six managers who identified as neither (8,780 decisions in total). Log-odds coefficients are reported in Appendix Table C.2.

We confirm the robustness of this finding in several ways. First, we consider whether the results from the logistic regressions presented above are sensitive to participant attention, several features of the environment, and possible experimenter demand or social desirability bias indicators. Supportive estimated coefficients from logistic regressions are reported in Appendix Table C.3. In terms of participant attention, we first confirm the pooled effect with the full sample of managers (i.e., including previously excluded respondents as per our pre-registered criteria). We also utilize several post-experiment survey items that measure the extent to which a participant took the study seriously (e.g., "I made decisions randomly") to construct an index (*Agree*) and interact it with the treatment (*Large Decision*). Neither of these approaches meaningfully attenuates the Large Decision effect (estimates reported in columns one and two of Appendix Table C.3).

With respect to features of the environment, we find that the pooled effect is not attenuated by i) the decision problem data file (Baseline versus Gender-Swapped), ii) the assigned treatment arm (More Small and More Large, which vary based on the proportion of small versus large decisions presented to managers), nor iii) the decision problem block (Block 1 with within-subject variation of small versus large decisions or Block 2 with between-subject variation in the same), with esti-

respectively).

mated coefficients reported in columns three through five in Appendix Table C.3. Additionally, given our stylized environment that emphasizes gender as a candidate characteristic, it is possible that participants were aware of the nature of the study and changed their behavior relative to the counterfactual where they were naïve to its intent. To investigate whether this is a substantial concern, we asked participants what they thought the study was about in a free-text response in the post-experiment questionnaire. We find that the pooled effect is not attenuated by whether the participant mentioned anything about gender (e.g., "gender," "female") in their response, as reported in column six of Appendix Table C.3.

Additionally, given that the between-subject (Block 2) treatment variation applied to the final 8 decision problems, only after managers had experienced both small and large decisions (in Block 1), we re-run our primary logistic regression using only round one decisions. Given the structure of the Block 1 decision problems, each subject was equally likely to be presented with a small or large decision with no prior experience in either, making this specification the cleanest between-subject comparison we have at our disposal. The estimated coefficients in column seven of Appendix Table C.3 illustrate that, if anything, the Large Decision effect is greater in magnitude in the first round (log-odds coefficient of -0.391 as compared to -0.128 using all observations, as shown in Appendix Table C.2).

So far, our primary analysis has relied on conventional logistic regression. Alternatively, we could test for the differential effect of a candidate profile being female in large decision problems via conditional logit, treating candidate profiles as the objects of choice. Doing so allows us to control for additional features of a candidate profile that ex-ante may affect its probability of being chosen. These could be rational features pertinent for payoff maximization (e.g., Partial Score) or randomized features of the choice environment (e.g., profile position on screen). Panel A of Appendix Table C.4 presents the estimated coefficients of five such specifications. The first four use all observations and iteratively control for the profile's Partial Score, position on screen, and a proxy for Partial Score computational complexity (number of subtraction operations in the arithmetic problem). The final specification restricts attention to the within-decision-problem consideration set, including only candidate profiles whose Partial Scores were viewed by the manager. Across all these specifications, the Large Decision effect persists, confirming the robustness of our central finding.

One could also imagine that there is considerable heterogeneity across hiring managers with respect to the response to large decisions—indeed, our results in the following section on the moderators of the Large Decision effect confirm this. To account for this heterogeneity, we additionally run mixed logistic regression analogues of the conditional logistic specifications just discussed (i.e., iteratively controlling for additional candidate profile features). Estimated mean coefficients and standard deviations are presented in Panel B of Appendix Table C.4 and further confirm the robustness of our central finding: female candidates are chosen less frequently in large decisions.

Before turning to additional analysis, we now highlight the connection between these results and our benchmark model predictions as presented in Section 2. Search, satisficing, and rational inattention models, assuming that managers carry a prior that favors men, all predict that male candidates will be favored and that this bias is exacerbated in large decisions. In contrast, we find that, if anything, female candidates are favored in small decisions, but not large ones; a bias against female candidates emerges for male managers in large decisions. These models diverge when we restrict attention to choices conditional on the consideration set. Search models predict no effect of gender, whereas satisficing and rational inattention models predict the same bias in favor of males

as before. None of these are borne out in the data.

What manager characteristics moderate the Large Decision effect?

We now turn to investigating the contours of the Large Decision effect with respect to a set of manager characteristics, finding evidence that it is moderated by search costs and attitudes in several relevant dimensions (e.g., hostile sexism). In contrast, beliefs do not moderate the effect. Table 2 reports the estimated coefficients from logistic regressions testing for the effects of a host of moderators.[5]

Table 2: Search costs and attitudes moderate the Large Decision effect; beliefs do not

	Search		Beliefs		Attitudes		
	Math Ability	Cognition	Direct	Stereotype	Hostile	Denial	Accept
Large Decision	−0.132***	−0.128***	−0.120***	−0.156***	−0.131***	−0.130***	−0.131***
	(0.034)	(0.034)	(0.034)	(0.039)	(0.033)	(0.033)	(0.033)
Moderator	−0.026	−0.011	0.083*	0.047	−0.010	−0.003	0.028
	(0.020)	(0.028)	(0.049)	(0.037)	(0.024)	(0.023)	(0.028)
Large Decision × Moderator	0.085***	−0.011	0.043	−0.053	−0.108***	−0.092***	−0.098***
	(0.031)	(0.034)	(0.056)	(0.042)	(0.034)	(0.032)	(0.037)
Constant	0.088***	0.086***	0.094***	0.111***	0.086***	0.086***	0.087***
	(0.023)	(0.023)	(0.023)	(0.032)	(0.023)	(0.023)	(0.023)

Notes: Log-odds coefficients from logistic regressions of *Choice Female* on *Large Decision*, a manager-level moderator, and their interaction; each column uses a different moderator. Standard errors clustered by subject in parentheses. $N = 8{,}900$ decisions from 445 managers in every model. *Math Ability* is the number of arithmetic problems solved correctly in the 60-second Math Part; *Cognition*, *Hostile*, *Denial*, and *Accept* are averages of three survey items each (see Appendix B). *Direct* is the manager's own belief about whether men or women are better at math and science tasks, and *Stereotype* is their belief about what most people in the United States think; both run from -2 (men "significantly better") to $+2$ (women "significantly better"), with 0 meaning "equally good." Search and Attitudes moderators are standardized, so the *Large Decision* coefficient is the effect at the sample mean and the interaction is the change per standard deviation. For Beliefs, the *Large Decision* coefficient is the effect for a manager who believes men and women are equally good. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Search: If the direct effect of large decision problems is to induce cognitive strain, then it stands to reason that managers may exhibit heterogeneous responses based on their ability and willingness to process relevant information. As such, we test for the effects of "search" moderators in several ways. First, we consider heterogeneous responses based on search costs, which we measure with a proxy based on the manager's ability to solve the arithmetic problems required to process the Partial Score of a candidate. Recall that each manager was tasked with solving as many of these arithmetic problems as possible in 60 seconds in the Math Part of the experiment. Those who could solve more problems in the allotted time may be less affected by large decisions since it is less costly for them to search through additional candidate profiles; those worse at solving them may disproportionately rely on free signal (gender). This conjecture is consistent with the predictions of a heuristic switching model, as we describe in Section 2.

Our results are consistent with this conjecture, as the Large Decision effect is moderated by the manager math ability. Table 2 reports the estimated coefficients from a logistic regression inclusive of a measure of math ability (*Math Ability*), specified as the number of math problems solved correctly in the Math Part of the experiment, standardized using the sample mean and standard

[5] Appendix Table C.5 reports the direct effect of each moderator (i.e., without the inclusion of *Large Decision*) on the propensity to choose a female candidate. Only Direct Beliefs, as well as Hostile and Denial Attitudes, have a direct effect on the propensity to hire a female candidate.

deviation. Managers with mean math ability exhibit a Large Decision effect similar in magnitude to the pooled effect documented previously (coefficient of -0.132 on *Large Decision* in the *Math Ability* model of Table 2 versus -0.128 in model one of Appendix Table C.2). The positive and statistically significant coefficient on the interaction term (*Large Decision* $\times$ *Moderator*) indicates that this effect reduces in magnitude with math ability. To further illustrate this point, we investigate the Large Decision effect for high versus low math ability participants. Figure 5 reports the estimated marginal effects of large decisions on the probability of choosing a female candidate for participants with low (below median) and high (above median) math ability, both pooled and for Female and Male managers separately. Interestingly, math ability attenuates the gender difference we documented in the previous section: low ability female participants are affected by large decisions similarly to low ability male participants, reducing the rate at which they choose female candidates. The effect of large decision problems disappears for high ability participants when pooled and for females, but not males.

In addition to utilizing this measure of a manager's *ability* to search, we also investigate the moderating effects of their *preference* to search. To do this, we construct an index of the manager's "need for cognition" (*Cognition* in Table 2) using three validated survey instruments (Cacioppo and Petty, 1982, see Appendix B for details). We find that this measure does not moderate the Large Decision effect.

Beliefs: It is possible that the effect is largest for those with the strongest beliefs regarding gender differences in math and science ability. To test this, we asked two questions in the post-experiment questionnaire: one regarding the manager's direct beliefs and another regarding what they thought "most people in the United States" believed. Responses were on a 5-point Likert scale, with extremes of "Women (men) are significantly better than men (women)" and a midpoint of "Women and men are equally good." For regression analysis, responses were recoded taking integer values from negative two to positive two, with a midpoint at zero. Neither the manager's direct beliefs (*Direct* in Table 2), nor their belief regarding what most people in the US think (*Stereotype* in Table 2) moderate the Large Decision effect, as indicated by the insignificant estimated coefficients on the interaction term. One may be tempted to view this as stemming from the use of unincentivized beliefs rather than an incentivized belief elicitation task. While we recognize that the estimated coefficients would certainly be different with incentivized beliefs, we contend that our unincentivized measures are meaningful in that they capture reasonable behavior with respect to the direct effect of beliefs on the propensity to choose a female candidate. Participants who believe that women are better at the task relative to men indeed choose them more frequently, as evidenced by the positive and statistically significant coefficient on the direct belief in Table 2 ($p < 0.10$), as well as the same in Appendix Table C.5 ($p < 0.01$).

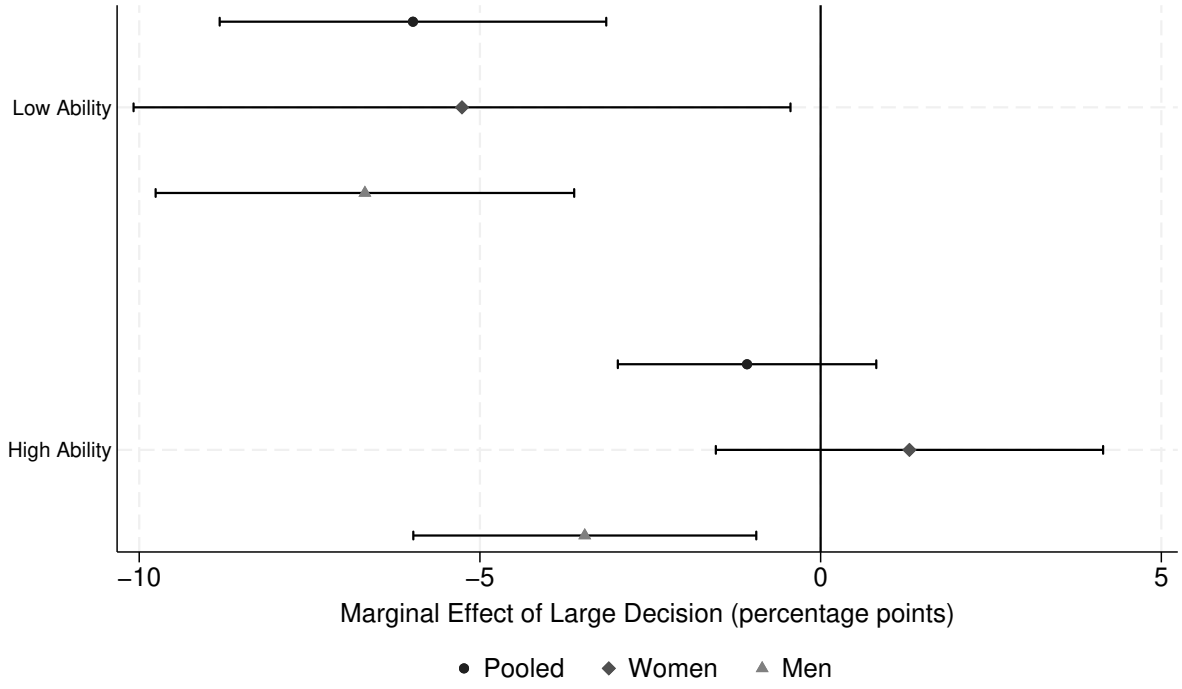


Figure 5: Math ability moderates the Large Decision effect

Notes: Average marginal effect of a large decision (16 candidates) relative to a small decision (4 candidates) on the probability of hiring a female candidate, in percentage points, with 95% confidence intervals, for managers below (Low Ability) and above (High Ability) the sample median score on the 60-second arithmetic task in the Math Part. Estimates come from logistic regressions of *Choice Female* on *Large Decision*, a High Ability indicator, and their interaction, with standard errors clustered by subject, estimated separately for all managers (Pooled), female managers, and male managers. Samples are as in Figure 4: 8,900 decisions from 445 managers (Pooled) and 8,780 decisions in the gender-specific samples.

Attitudes: Finally, we explore the extent to which the effect is moderated by a host of attitudes. First, we measure hostile sexism using a three-item index drawn from the hostile sexism subscale of the Ambivalent Sexism Inventory (e.g., Glick and Fiske, 1996; Rollero et al., 2014, see Appendix B for more information). Managers rate agreement on a five-point Likert scale with three statements (e.g., "women seek to gain power by getting control over men"). We average responses into a single hostile sexism index (*Hostile* in Table 2).

Among managers with mean levels of hostile sexism, the Large Decision effect is similar to the pooled effect reported above (coefficient of -0.131 on *Large Decision* in the *Hostile* model of Table 2 versus -0.128 in model one of Appendix Table C.2). The interaction between the large decision indicator and the hostile sexism index is negative and significant, indicating that more hostile managers exhibit a larger anti-female bias in large decisions. Figure 6 shows the corresponding marginal effects: the least hostile managers show no significant effect, with confidence intervals including zero for the pooled sample as well as for men and women separately. The most hostile managers (*Hostile* = 5) show an approximately twenty percent reduction in the female hiring rate. This pattern also helps explain the gender difference in our main results: point estimates are similar

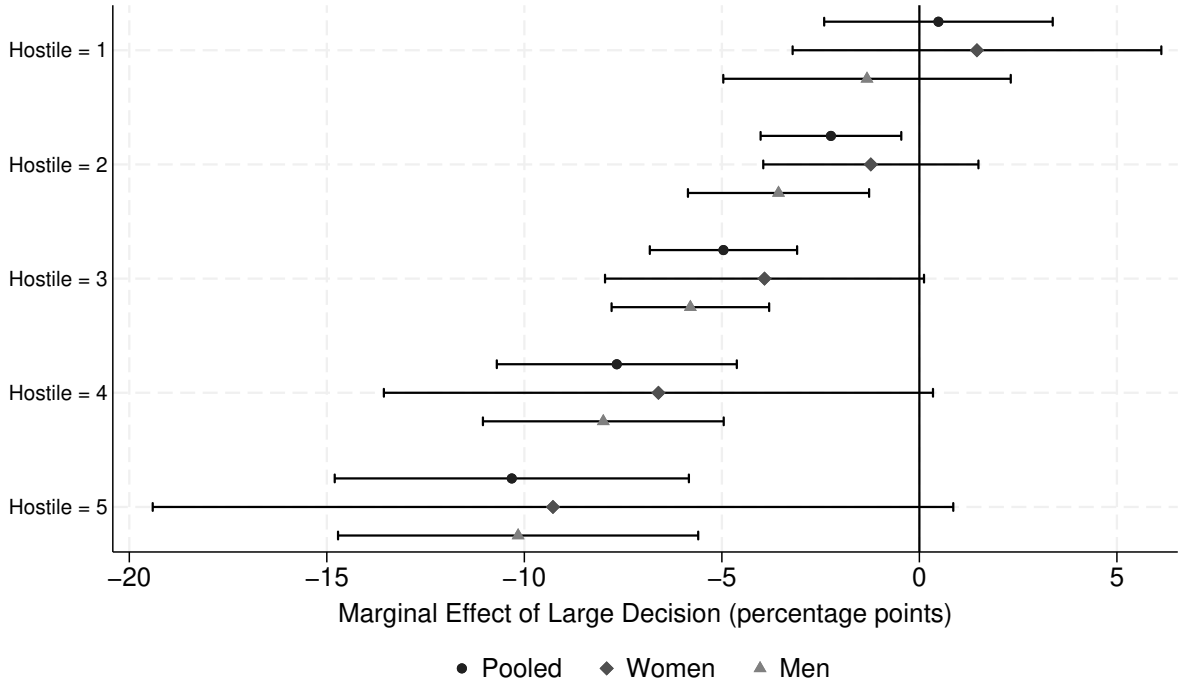


Figure 6: Hostile sexism moderates the Large Decision effect

Notes: Average marginal effect of a large decision (16 candidates) relative to a small decision (4 candidates) on the probability of hiring a female candidate, in percentage points, with 95% confidence intervals, evaluated at each integer value of the hostile sexism index (*Hostile*). *Hostile* is the average of three agreement items from the hostile sexism subscale of the Ambivalent Sexism Inventory (e.g., Glick and Fiske, 1996; Rollero et al., 2014), ranging from one (least hostile) to five (most hostile); see Appendix B. Estimates come from logistic regressions of *Choice Female* on *Large Decision*, unstandardized *Hostile*, and their interaction, with standard errors clustered by subject, estimated separately for all managers (Pooled), female managers, and male managers. Samples are as in Figure 4.

across gender within hostile sexism bins, and the null effect for women overall is driven in part by wide confidence intervals at high hostile sexism levels, where there are relatively few female managers. For example, approximately 25% of male managers have *Hostile* scores greater than three, compared to only about 10% of females.

We construct indexes that capture the extent to which the manager denies sex-based discrimination exists (*Denial*) and the extent to which the manager finds discriminatory hiring practices acceptable (*Accept*) in a similar fashion (see Appendix B for more detailed information on these indexes). These attitudes moderate the Large Decision effect in a manner similar to hostile sexism (as seen in Table 2; marginal effects reported in Appendix Figures C.2 and C.3).[6]

Do managers use different standards when choosing men versus women?

While our benchmark models carry identical predictions with respect to the gender bias in hiring decisions, they differ in the standard applied to considered male and female candidates. Search

[6]Given that both *Math Ability* and these attitudes moderate the Large Decision effect, it is natural to wonder whether each effect survives when they are jointly estimated. We additionally run logistic regressions including the direct effect of *Math Ability* and each attitude scale, along with both interactions with *Large Decision*. We find that both sets of effects survive at magnitudes similar to those estimated in Table 2.

models predict that hired men and women will be of equal quality given a single applied reservation utility. In contrast, satisficing and rational inattention models predict that female candidates will be held to a higher standard and thus the (observable to managers) ability of chosen female candidates should be higher than that of males. In Table 3 we present the estimated coefficients from OLS models crafted to detect these differences. Model one treats the Partial Score of the chosen candidate (*Chosen Partial*) as the dependent variable, whereas model two uses the gap between the chosen Partial Score and the maximum available in the decision problem (*Chosen Gap*). In each, the sample is restricted to only those observations where the chosen candidate was viewed by the manager and thus the Partial Score was in their information set. We find that hired candidate quality does not depend on gender: the coefficient on *Female Candidate* is statistically indistinguishable from zero in both models.

Table 3: No detectable gender gap in hired candidate quality

	Chosen Partial (1)	Chosen Gap (2)
Female Candidate	−0.045	0.010
	(0.039)	(0.028)
Large Decision	0.563***	0.592***
	(0.059)	(0.055)
Female × Large	−0.002	0.055
	(0.062)	(0.054)
Constant	6.623***	0.496***
	(0.041)	(0.038)
Observations	8,260	8,260
Subjects	440	440

Notes: OLS coefficients from regressions of the quality of the hired candidate on *Female Candidate* (one if the hired candidate is female), *Large Decision*, and their interaction; the unit of observation is a decision. *Chosen Partial* is the Partial Score of the hired candidate (model one). *Chosen Gap* is the highest Partial Score among all candidates in the decision minus *Chosen Partial* (model two). Standard errors clustered by subject in parentheses. The sample is decisions in which the hired candidate's Partial Score was viewed: $N = 8{,}260$ decisions from 440 managers in both models. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

4.2 Choice Process

Our chosen design has the advantage of allowing us to reconstruct consideration sets by observing which candidates were considered (i.e., their Partial Score was viewed), for how long, and in what order. In this section, we first turn to preliminaries and confirm that indeed large decisions generate larger consideration sets (i.e., that the size of the choice set in small problems binds on

average), less frequent complete consideration sets, and more frequent empty consideration sets. They also generate less intensive search (i.e., less time spent per profile). We then confirm that the Large Decision effect is robust to the inclusion of the direct effect on consideration behavior, highlighting a conflict with the benchmark model predictions as summarized in Table 1: the Large Decision effect is significant and of relatively large magnitude, even for observations with empty consideration sets.

We then turn to other features of the process data that connect directly to these predictions. The female share of the consideration set is roughly 50% and does not depend on the size of the choice set. The order of consideration is slightly biased in favor of female candidates (i.e., they are viewed earlier in search) in large decisions, but this difference is economically insignificant and yields no explanatory power for hiring choices. The choice conditional on the consideration set is gender-biased, favoring females in small decisions and males in large. None of our benchmark models is consistent with the totality of these process results.

Do large decisions generate more search?

Table 4 reports the estimated coefficients for a number of specifications intended to test for the effect of large decisions on search overall. We find that participants search more overall, but less intensively in large decisions.

Table 4: Large decisions increase search overall but reduce time viewed per candidate

	Candidates Viewed			Time Viewed (sec.)	
	N Viewed	Viewed None	Viewed All	Total	Per Candidate
	(1)	(2)	(3)	(4)	(5)
Large Decision	6.786***	0.500***	−1.881***	23.243***	−0.663***
	(0.240)	(0.162)	(0.107)	(1.336)	(0.180)
Constant	3.597***	−2.953***	1.767***	16.356***	4.612***
	(0.043)	(0.182)	(0.108)	(0.638)	(0.190)
Model	OLS	Logit	Logit	OLS	OLS
Observations	8,900	8,900	8,900	8,895	8,325
Subjects	445	445	445	445	444

Notes: Coefficients on *Large Decision* from regressions of search behavior. Standard errors clustered by subject in parentheses. *N Viewed* is the number of candidates whose Partial Scores the manager viewed. *Viewed None* and *Viewed All* equal one if the manager viewed no Partial Scores, or all of them, in the decision, and zero otherwise. *Total* is the time in seconds spent viewing Partial Scores in the decision, and *Per Candidate* is the time spent per candidate viewed. *Total* excludes five decisions with missing time data; *Per Candidate* additionally excludes 570 decisions in which no Partial Score was viewed. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

We measure the extent of search with two alternative approaches: using a binary indicator for whether a Partial Score was viewed and the view time (in seconds) of the Partial Score. The first three models in Table 4 take the first approach and indicate that participants view more candidates' Partial Scores overall (i.e., *N Viewed* is increasing in *Large Decision*), but that large decisions lead to a higher incidence of no search (*Viewed None*) and lower incidence of complete search (*Viewed All*). The remaining models consider the time spent viewing Partial Score information,

both overall (*Total*) and per candidate profile viewed (*Per Candidate*). While participants spend more time overall viewing Partial Score information (i.e., *Total* is increasing in *Large Decision*), they spend less time per viewed candidate in large decisions, consistent with an interpretation that search is less intensive. [7]

How does endogenous search moderate the Large Decision effect?

It is natural to consider whether the Large Decision effect runs through changes in endogenous search. At one extreme, a manager who chooses without searching any profiles has no relevant information upon which to base their decision other than gender. It is possible that such a subject is unaffected by large decisions: random choice would result in no effect, as would many choice rules based on statistical or taste-based discrimination (e.g., "always choose a man"). At the other extreme, a manager who considers all profiles may similarly be unaffected by large decisions: optimal choice would require equal propensity to choose a female candidate given the structure of the decision problem data. Nevertheless, as we conjecture in our discussion of heuristic switching in Section 2, it is possible that managers use different choice rules between small and large decisions such that endogenous consideration does not fully attenuate the Large Decision effect.

Our results indicate that while hiring decisions are moderated by search behavior, controlling for differential search patterns does not absorb the Large Decision effect. Table 5 reports the estimated coefficients for several specifications of logistic regressions controlling for the search variables we previously investigated as moderators. Each model treats the binary *Choice Female* as the dependent variable and includes *Large Decision*, a relevant search moderator (*Moderator*, with column labels indicating which moderator is used), and their interaction as independent variables.

The extensive margin of search weakly moderates the effect, but does not eliminate it. The interaction terms in Table 5 show that the Large Decision effect is larger when the manager viewed no candidates (column 2), and smaller when the manager viewed all of them (column 3), those these effects are only marginally significant ($p < 0.10$). The intensive margin, reflected by time spent per profile or in total, does not moderate it (columns 4–5). The levels are displayed in Table 6. With no search, female candidates are hired in 58.4 percent of small decisions and 47.4 percent of large ones, an eleven percentage point reduction. The reduction is 5 percentage points when managers search some but not all candidates, and 2 percentage points when all candidates are viewed. In short, the Large Decision effect is strongest when the manager has nothing but gender and screen position to go on, reduces with search, but remains even when every Partial Score is viewed.

These results indicate that while large decisions affect search, and search patterns can in turn moderate the propensity to choose female candidates in the same, the Large Decision effect survives when we control for this phenomenon. Moreover, the substantial and statistically significant impact even with zero consideration is inconsistent with the predictions of any of our benchmark models as summarized in Table 1.

Does the female share of the consideration set depend on set size?

Our benchmark models differ in their predictions regarding the proportion of female candidates in the consideration set. Search models predict that the female share falls in large available sets, as would a model based on Rational Inattention. Satisficing models, however, generate no such

[7] We use a similar set of specifications to confirm that math ability behaves as a measure of search cost (Appendix Table C.6). A one standard deviation increase in *Math Ability* is associated with viewing 0.24 more candidates in small decisions and 1.65 more in large decisions (interaction $p < 0.001$), a lower probability of viewing no Partial Scores and a higher probability of viewing all of them, with no detectable difference by decision size, and less time per candidate viewed (1.35 fewer seconds in small decisions and 0.87 fewer in large decisions).

Table 5: Search can weakly moderate the Large Decision effect, but does not eliminate it

	Candidates Viewed			Time Viewed	
	N Viewed	Viewed None	Viewed All	Total	Per Candidate
	(1)	(2)	(3)	(4)	(5)
Large Decision	−0.316***	−0.109***	−0.275***	−0.143***	−0.120***
	(0.118)	(0.033)	(0.091)	(0.049)	(0.040)
Moderator	−0.042	0.267*	−0.143	0.000	0.001
	(0.030)	(0.160)	(0.095)	(0.001)	(0.003)
Large × Moderator	0.046	−0.335*	0.196*	0.000	0.003
	(0.029)	(0.198)	(0.101)	(0.001)	(0.005)
Constant	0.237**	0.073***	0.208**	0.083***	0.070**
	(0.118)	(0.022)	(0.091)	(0.031)	(0.028)
Observations	8,900	8,900	8,900	8,895	8,325
Subjects	445	445	445	445	444

Notes: Log-odds coefficients from logistic regressions of *Choice Female* on *Large Decision*, a search moderator, and their interaction, one per column. Standard errors clustered by subject in parentheses. Moderators are unstandardized and defined as in Table 4 (number of candidates viewed, indicators for viewing none or all, and time in seconds), so the *Large Decision* coefficient is the effect when the moderator equals zero. *Total* excludes five decisions with missing time data; *Per Candidate* additionally excludes 570 decisions in which no Partial Score was viewed. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 6: Large Decision effect highest in magnitude when no Partial Scores viewed; remains when all viewed

	Small	Large	Difference (pp)	p
Viewed None	58.4% (226)	47.4% (344)	−11.0 (4.8)	0.023
Partial search	53.5% (439)	48.5% (1,950)	− 5.0 (2.8)	0.077
Viewed All	51.6% (3,893)	49.7% (2,048)	− 1.9 (0.9)	0.036
Viewed ≥ 4	51.6% (3,893)	49.1% (3,385)	− 2.5 (0.9)	0.003

Notes: Share of decisions in which a female candidate was hired, by the number of Partial Scores the manager viewed, with the number of decisions in parentheses. *Viewed None* and *Viewed All* mean the manager viewed no candidates' Partial Scores or all of them; *Partial search* means at least one but not all. These three rows partition the 8,900 decisions. *Viewed* ≥ 4 restricts large decisions to those with at least four candidates viewed, the number in a full small-decision search; for small decisions it coincides with *Viewed All*. Difference is the large-decision share minus the small-decision share, in percentage points. The standard error (in parentheses) and p-value come from a logistic regression of *Choice Female* on *Large Decision* within each row, with standard errors clustered by subject.

dependence. To distinguish between them, we test whether the proportion of binary Partial Score views and viewing time dedicated to female profiles depends on the size of the choice set.

Table 7 presents the estimated coefficients from a number of OLS models intended to test for this

effect. In Models 1 through 5, the dependent variable is *Female Share: Considered*, which takes values between zero and one and indicates the proportion of profiles whose Partial Scores were viewed that were female; models 6 through 10 consider the viewing time (in seconds) analogue *Female Share: Time Viewed*. *Female Share: Considered* is 50.2% in small decision problems ($p = 0.463$ for $H_0 : Constant = 0.5$ in model 1) and *Female Share: Time Viewed* is slightly above this at 50.5% ($p = 0.068$ for $H_0 : Constant = 0.5$ in model 6). The estimated coefficients on *Large Decision* in models 1 and 6 are inconsistent with the prediction of the search and rational inattention models: the female share of the consideration set does not depend on the size of the choice set.

Despite the fact that the female share does not depend on the size of the choice set *overall*, it may for the managers for whom the treatment effect is largest. Recall from the previous section that these included managers who were relatively low in math ability, more sexist, more accepting of discriminatory hiring practices, and who exhibited more denial of discrimination. Models 2 through 5 and 7 through 10 include both these moderators and the interaction with large decisions and indicate that this is not the case: none of the previously identified moderators has a statistically significant impact on the female share nor on the impact of large decisions on the same.

Table 7: No detectable difference in female share of consideration set by choice set size

	Female Share: Considered					Female Share: Time Viewed				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Large Decision	0.002	0.002	0.002	0.002	0.002	−0.002	−0.003	−0.002	−0.002	−0.002
	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Moderator		0.001	−0.004	−0.003	−0.002		−0.002	−0.005	−0.003	−0.003
		(0.002)	(0.003)	(0.003)	(0.003)		(0.003)	(0.003)	(0.003)	(0.003)
Large × Moderator		0.002	0.000	−0.003	0.000		0.005	0.001	0.000	0.002
		(0.003)	(0.004)	(0.004)	(0.004)		(0.004)	(0.005)	(0.005)	(0.005)
Constant	0.502***	0.502***	0.501***	0.502***	0.502***	0.505***	0.506***	0.505***	0.505***	0.505***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Moderator		Math Ability	Hostile	Denial	Accept		Math Ability	Hostile	Denial	Accept
Observations	8,330	8,330	8,330	8,330	8,330	8,325	8,325	8,325	8,325	8,325
Subjects	444	444	444	444	444	444	444	444	444	444

Notes: OLS coefficients, with standard errors clustered by subject in parentheses. In models 1 to 5, the dependent variable *Female Share: Considered* is the proportion of the candidates whose Partial Scores the manager viewed in the decision who were female. In models 6 to 10, *Female Share: Time Viewed* is the proportion of the total time spent viewing Partial Scores that was spent on female candidates. Regressors are *Large Decision* and, in models 2 to 5 and 7 to 10, a standardized manager-level moderator and its interaction with *Large Decision*. The moderators are *Math Ability* and the attitude scales as defined in Appendix B, standardized. The constant is the female share in small decisions (at the moderator mean in models 2 to 5 and 7 to 10). Models 1 to 5 exclude 570 decisions in which no Partial Score was viewed; models 6 to 10 additionally exclude five decisions with missing time data. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Does the order of consideration of male and female candidates depend on set size?

Our benchmark models also contrast in their predictions for the order in which male and female candidates are considered. Independent of the relative shares of female/male candidates in the consideration set, Search and Rational Inattention models predict that male candidates are considered earlier; Satisficing models predict that the order is gender-independent. Table 8 presents the estimated coefficients of several OLS models used to test these predictions. In each model, the unit of observation is a candidate profile viewed in some decision problem (i.e., the analysis is restricted only to those profiles that were actually viewed). The dependent variable is the normalized consideration order taking values from zero (first considered) to one (last considered) of the profile (*Normalized Consideration Order*) and constructed as $\frac{\text{Considered Order}-1}{\text{Total Considered}-1}$ in order to facilitate comparison between small decisions, where considered order ranges from one (first) to four (last), and large ones with a range up to 16 (last). In contrast to all of the benchmark model predictions,

female candidates appear to be considered marginally earlier, but only in large decision problems (coefficient on *Female* $\times$ *Large* being negative and significant in Table 8, model one). Following our earlier approach, we also interact this effect with the effect of our manager-level moderators in models two through five, finding no significant effect.

While the effect of large decisions on consideration order is significant in a statistical sense, we argue that the effect on female candidates is not of meaningful economic magnitude. To illustrate this, consider that a candidate profile can minimally be considered one position earlier in some search, corresponding to a change in *Normalized Consideration Order* of roughly -0.33 if in a small decision or -0.066 if in a large one (at minimum, with full consideration) with a fixed consideration set size. In this context, a coefficient of -0.014 on *Female* $\times$ *Large*, as in model one of Table 8, is obviously of relatively small economic significance. This interpretation is reinforced by observing that the impact of candidate gender on the likelihood of being chosen in large versus small problems is of similar magnitude when controlling for this small magnitude difference in consideration order. The log-odds coefficient *Female* $\times$ *Large Decision* goes from -0.128 to -0.117 after the inclusion of normalized consideration order in conditional logistic regression, a difference that is not statistically significant ($p = 0.192$ via the triple interaction term). As such, while we can precisely estimate even a small magnitude consideration order change in large choice sets, this exercise yields no explanatory power when linked to choice data.

Equal Exposure, Unequal Use

The results above show that hiring choices respond to the size of the choice set, and that this response is moderated by search costs and by managers' attitudes toward women. Taken together, these patterns are hard to reconcile with benchmark models of search, satisficing, and limited attention, all of which locate the source of choice-set-size effects in what information managers acquire. Yet the differential search patterns predicted by those models do not appear in the data: as shown above, men and women receive roughly equal attention, in both small and large choice sets. If the departure from the benchmark predictions is not explained by unequal exposure to candidate information, it must instead lie in how managers use the information they do acquire. In this section, we investigate that possibility directly.

We first consider the possibility that gender affects the perception or use of Partial Score information such that the returns to Partial Scores with respect to the impact on choice probability differ by candidate gender. If managers systematically perceive identical Partial Scores as less informative for female candidates, more intensive patterns of this bias in large decision problems would generate our estimated treatment effects. We find no evidence of this in our data: additional conditional logistic regressions that interact the Partial Score with candidate profile gender reveal that the returns to higher Partial Scores do not differ by candidate gender, regardless of choice set size (see Appendix Table C.7).

If managers do not discount female candidates' ability in this manner, do they nonetheless pass over them when they are the strongest candidate viewed? Table 9 presents the frequency with which a manager chooses the best-viewed candidate in the decision problem, measured as a choice of the candidate with the highest Partial Score within the consideration set. These means are presented separately by best-viewed candidate gender and decision problem size. We first note that managers choose the best-viewed candidate less frequently in large decisions. Critically, best-viewed female candidates suffer more from this effect: when the best-viewed candidate is female, she is passed over 11.2pp more frequently in large decisions, compared to 4.1pp for males (diff-in-diff $p < 0.001$

Table 8: Female candidates considered marginally earlier in large decisions

	Outcome: Normalized Consideration Order				
	(1)	(2)	(3)	(4)	(5)
Female Candidate	0.011	0.010	0.009	0.010	0.010
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Large Decision	0.007**	0.007*	0.007*	0.007**	0.007*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Female × Large	−0.014**	−0.014*	−0.014*	−0.014**	−0.014*
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Moderator		−0.002	0.006*	0.006*	0.003
		(0.003)	(0.003)	(0.003)	(0.004)
Female × Moderator		0.005	−0.012*	−0.012*	−0.007
		(0.006)	(0.007)	(0.007)	(0.007)
Large × Moderator		0.003	−0.004	−0.004	−0.003
		(0.003)	(0.004)	(0.004)	(0.004)
Female × Large × Moderator		−0.005	0.007	0.009	0.007
		(0.007)	(0.007)	(0.007)	(0.007)
Constant	0.495***	0.495***	0.495***	0.495***	0.495***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Moderator		Math Ability	Hostile	Denial	Accept

Notes: OLS coefficients, with standard errors clustered by subject in parentheses. The unit of observation is a candidate profile whose Partial Score was viewed. The dependent variable, *Normalized Consideration Order*, equals $\frac{\text{Considered Order}-1}{\text{Total Considered}-1}$, where Considered Order is the position in which the profile's Partial Score was viewed and Total Considered is the number of profiles viewed in the decision. It runs from zero (first viewed) to one (last viewed), so negative coefficients mean earlier consideration. Regressors are *Large Decision*, *Female Candidate* and, in models 2 through 5, a standardized manager-level moderator and its interactions. The moderators are *Math Ability* and the attitude scales as defined in Appendix B, standardized. The sample is 61,073 profiles from 370 managers: the 61,470 viewed profiles (of 87,704 shown) less 397 profiles from decisions in which only one profile was viewed, for which the normalized order is undefined. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

per logistic regression).

Having shown that large decisions lead managers to disproportionately pass over the best-viewed candidate when she is a woman, we next ask whether this pattern is sensitive to the cost of doing so. For each decision problem, we construct a measure of the cost of passing over the best-viewed candidate, defined as the Partial Score gap between them and the best (highest Partial Score) opposite-gender candidate that was viewed (*Lead*). In Table 10, we disaggregate the data presented above by *Lead*, restricting attention to observations where there is a non-trivial cost to discriminating against the best-viewed candidate, i.e., where *Lead* is greater than zero, and where at least one female and one male candidate are viewed. As expected, the frequency with which the best-viewed candidate is hired increases with *Lead*: as it becomes more expensive to pass over a best-viewed candidate, managers do so less frequently. Moreover, we observe that for all costs, best-viewed female candidates are passed over by a larger magnitude in large decisions as compared to males. For example, when the best-viewed candidate only leads by one point in the Partial

Table 9: Best-viewed female candidates passed over more frequently in large decisions

	Small Decisions	Large Decisions	Diff. (Large − Small)
Best-Viewed: Male	0.781	0.740	−0.041***
	(0.013)	(0.014)	
Best-Viewed: Female	0.814	0.702	−0.112***
	(0.013)	(0.016)	
Diff. (Female − Male)	0.033***	−0.038***	−0.070***

Notes: Adjusted predictions of the probability that the manager hires the best-viewed candidate, the candidate with the highest Partial Score among those whose Partial Scores the manager viewed, by choice set size (small: 4 candidates; large: 16) and by the gender of the best-viewed candidate. Standard errors in parentheses. Differences are between the adjusted predictions (large minus small decisions; female minus male best-viewed candidate). Predictions come from a logistic regression of an indicator for hiring the best-viewed candidate on an indicator for large decisions, an indicator for the best-viewed candidate being female, and their interaction, with standard errors clustered by subject. The sample is the $N = 8{,}112$ decisions in which at least one Partial Score was viewed, excluding decisions in which two or more viewed candidates were tied for the highest Partial Score. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Score compared to the best-viewed opposite-gender candidate, best-viewed females suffer from a 16pp penalty in large decisions, as compared to 8pp for males.

Table 10: Best-viewed candidate hired less often in large decisions at every lead; larger effect when the best candidate is female

Lead	Best Hired in Small	Best Hired in Large	Difference	N (Small/Large)
When best is female				
1	80%	64%	−16	(1392/632)
2	83%	72%	−11	(316/850)
≥3	85%	74%	−11	(338/257)
When best is male				
1	76%	68%	−8	(1410/672)
2	80%	75%	−5	(311/841)
≥3	85%	79%	−6	(319/256)

Notes: Share of decisions in which the manager hires the best-viewed candidate, the candidate with the highest Partial Score among those viewed, by decision size (small: 4 candidates; large: 16), the gender of the best-viewed candidate, and *Lead*. *Lead* is the Partial Score of the best-viewed candidate minus the highest Partial Score among viewed candidates of the other gender, with leads of three or more pooled. The sample is decisions in which at least one male and at least one female candidate were viewed and *Lead* is greater than zero. Difference is the large-decision share minus the small-decision share, in percentage points. Differences are descriptive; Table 11 reports the corresponding regression.

Table 11 reports analogous results from a logistic regression used to investigate whether con-

trolling for this cost absorbs the Large Decision effect. In this regression, the dependent variable is an indicator for whether the best-viewed candidate was chosen (*Best Viewed Candidate Hired)* and includes as independent variables indicators for large decisions (*Large Decision*) and whether the best-viewed candidate was female (*Best is Female*) alongside *Lead* and the interaction terms. Consistent with the results above, Table 11 indicates that best-viewed candidate is passed over more frequently in large decisions, with this penalty applying more strongly to female candidates. Moreover, the best-viewed candidate is passed over more frequently when the cost of doing so to hire a candidate of the opposite gender is low (positive and significant coefficient on *Lead*). This "cost of discrimination" effect appears to differentially affect neither best-viewed female candidates (insignificant coefficient on *Best is Female x Lead*) nor best-viewed candidates in large decisions (insignificant coefficient on *Large Decision x Lead*). Moreover, controlling for this effect does not absorb the Large Decision effect, consistent with an interpretation that our treatment effects are not driven by mechanically different distributions in *Lead* across large and small decisions.

In summary, from this additional analysis we glean that while managers appear to pursue equal exposure to male and female candidates in larger pools, they do not pursue equal use of this information. Instead, they pass over best-viewed female candidates more frequently, even when the cost of doing so is high. Taken together, these results are difficult to square with the predictions of benchmark attention models applied to our setting and instead suggest that managers rely on some coarse heuristic (e.g., "choose more men") in large decisions.

Table 11: Probability of hiring best-viewed candidate falls more in large decisions when she is female, controlling for Partial Score lead

	Best Viewed Candidate Hired
Large Decision	−0.293*
	(0.154)
Best is Female	0.393***
	(0.140)
Lead	0.241***
	(0.058)
Large Decision × Best is Female	−0.588***
	(0.196)
Large Decision × Lead	−0.025
	(0.088)
Best is Female × Lead	−0.120
	(0.078)
Large Decision × Best is Female × Lead	0.130
	(0.107)
Constant	0.889***
	(0.109)

Notes: Log-odds coefficients from a logistic regression of an indicator that the manager hired the best-viewed candidate on *Large Decision*, *Best is Female* (one if the best-viewed candidate is female), *Lead*, and all of their interactions. *Lead* is the Partial Score of the best-viewed candidate minus the highest Partial Score among viewed candidates of the other gender, in points (not pooled as in Table 10); it is the cost of passing over the best-viewed candidate. Main-effect coefficients are evaluated at *Lead* $= 0$ and a male best-viewed candidate. The sample is $N = 7{,}594$ decisions from 436 managers in which at least one male and at least one female candidate were viewed and *Lead* is greater than zero. Standard errors clustered by subject in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

5 Discussion and Conclusion

Inspired by the variation in applicant pool size encountered by hiring practitioners in the field, we report the results of an experiment designed to causally identify the impact of choice set size on gender-biased decision making when hiring for a math and science task. We prompt participant "managers" to hire a single candidate based on limited information, including the candidate's gender and a noisy and difficult-to-parse signal of their math and science ability. They repeatedly choose from controlled sets of applicants that vary in size, allowing us to assess the impact of choice set size on patterns of bias.

We find that female candidates are disadvantaged in larger applicant pools, particularly when assessed by male managers and those with higher search costs and sexist attitudes. Our design allows us to uncover that this difference emerges not as a result of changing search or attention patterns: female candidates are not viewed less frequently or for less time in larger pools. Instead,

managers appear to pass over high-quality women more frequently in large decisions, regardless of the cost of doing so. We interpret this as managers exhibiting "equal exposure, unequal use" of candidate information along gender lines, a finding that is difficult to square with several benchmark models (i.e., search, satisficing, rational inattention, as discussed in Section 2) applied to our setting.

By process of elimination, what remains as the leading interpretation of our findings is that managers switch from one choice rule to another based on the size of the available set, consistent with the psychology literature on the use of stereotypes under cognitive load and the economics literature on choice overload. Under both interpretations, managers under cognitive strain are less able to correct (or overcorrect) for stereotypes and rely more on category-based than on individuating information. Consistent with our findings, this effect would be concentrated among managers less able to process candidate information and those harboring more severe sexist attitudes.

Against this backdrop, our findings contribute to the nascent literature on decision bias moderated through attention and cognitive load (e.g., Bartoš et al., 2016). By showing that the effect is concentrated among managers with hostile attitudes, our work complements the findings of Abel and Burger (forthcoming). In a controlled experiment, they manipulate decision time and find that under time pressure, participants rely more heavily on race as a heuristic, using perceived racial identity as a shortcut for productivity per their prior beliefs, resulting in lower hiring rates for applicants with Black-sounding names. We complement these findings by showing that a similar effect emerges where choice set size is the catalyst for additional cognitive strain, but that managers rely on hostile attitudes rather than beliefs regarding ability. Taken together, these results imply that the continued development of formal models considering these channels may prove fruitful.

With respect to experiments, our findings suggest caution when designing studies with the aim of detecting the presence of labor market discrimination. Historically, most projects in this vein have been correspondence studies wherein a researcher submits controlled fictitious missives (e.g., emails, job inquiries) to firms that randomize along the dimension under study (e.g., gender or race as signaled through name). In these cases, researchers typically focus on the nature of any subsequent communication (e.g., callback rates, response content), sometimes pairing it with publicly available information on firm characteristics. In contrast, characteristics of the relevant "applicant pool" are unobserved. The remaining studies typically work with relatively small choice sets or do not involve hiring decisions at all. We are unaware of any controlled labor market experiments where participants make a hiring choice from sets as large as our 16-candidate decisions, despite the fact that applicant pools are frequently much larger in field settings. As a result, our results point to significant risk in overgeneralizing results from laboratory experiments on discrimination, particularly null results, and provide motivation for careful consideration of choice set size as a moderator of experimental findings.

Finally, this project yields implications for hiring practitioners. Our results point to a tension that firms weighing applicant-pool size have not had reason to consider. The standard case for casting a wide net is that a larger pool raises the odds of finding a strong candidate. This benefit of larger pools comes alongside known costs: larger pools take more manager time and money to process. Our results add a cost of a different kind, an equity cost. Existing practitioner and HR discourse has indeed already identified a version of this tradeoff, but has been focused almost entirely on the impact of potentially biased algorithmic pre-screening (e.g., Bogen and Rieke, 2018). Our results identify a separate channel whereby a larger pool can degrade equitable hiring through the cognitive demands of the task itself. This channel is additive to the debate surrounding biased algorithmic screening and suggests new equity gains to properly calibrating applicant pool size.

One may object that this channel is a secondary concern given that most firms at scale already use pre-screening procedures, but this neglects the fact that the intensity of screening is itself a firm choice. Manager discretion as to the intensity and structure of screening is at the heart of most Applicant Tracking Systems and there is emerging evidence that even the use of gender-balanced shortlisting procedures does not guarantee increased diversity in final hires (e.g., Peng et al., 2019; Sühr et al., 2021; Parasurama and Ipeirotis, 2026). Our results give firms a reason to consider screening intensity as an equity lever in addition to an efficiency one, since the size of the final choice set presented to human managers carries with it a gender-equity cost.

References

M. Abel and R. Burger. Name-based race discrimination: The role of heuristics. *Review of Economics and Statistics*, forthcoming.

G. W. Allport. *The Nature of Prejudice*. Addison-Wesley, Cambridge, MA, 1954.

K. J. Arrow. The theory of discrimination. In O. Ashenfelter and A. Rees, editors, *Discrimination in Labor Markets*, pages 3–33. Princeton University Press, Princeton, NJ, 1973.

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

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

G. S. Becker. *The Economics of Discrimination*. University of Chicago Press, Chicago, 1957.

T. Besedeš, C. Deck, S. Sarangi, and M. Shor. Decision-making strategies and performance among seniors. *Journal of economic behavior & organization*, 81(2):524–533, 2012a.

T. Besedeš, C. A. Deck, S. Sarangi, and M. Shor. Age effects and heuristics in decision making. *Review of Economics and Statistics*, 94(2):580–595, 2012b.

G. V. Bodenhausen. Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science*, 1(5):319–322, 1990.

G. V. Bodenhausen and M. Lichtenstein. Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology*, 52(5):871–880, 1987.

M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Technical report, Upturn, Dec. 2018. URL `https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf`.

J. K. Bosson, P. Jurek, J. A. Vandello, N. Kosakowska-Berezecka, M. Olech, T. Besta, M. Bender, V. Hoorens, M. Becker, A. Timur Sevincer, et al. Psychometric properties and correlates of precarious manhood beliefs in 62 nations. *Journal of Cross-Cultural Psychology*, 52(3):231–258, 2021.

T. L. Botelho and M. Abraham. Pursuing quality: How search costs and uncertainty magnify gender-based double standards in a multistage evaluation process. *Administrative Science Quarterly*, 62(4):696–730, 2017.

J. T. Cacioppo and R. E. Petty. The need for cognition. *Journal of Personality and Social Psychology*, 42(1):116–131, 1982.

J. T. Cacioppo, R. E. Petty, and C. F. Kao. The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48(3):306–307, 1984. doi: 10.1207/s15327752jpa4803_13.

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

A. Caplin, M. Dean, and D. Martin. Search and satisficing. *American Economic Review*, 101(7): 2899–2922, 2011.

A. Caplin, M. Dean, and J. Leahy. Rational inattention, optimal consideration sets, and stochastic choice. *Review of Economic Studies*, 86(3):1061–1094, 2019.

M. Castillo and R. Petrie. Discrimination in the lab: Does information trump appearance? *Games*

and Economic Behavior, 68(1):50–59, 2010.

I. Chadd, E. Filiz-Ozbay, and E. Y. Ozbay. The relevance of irrelevant information. *Experimental Economics*, 24(3):985–1018, 2021.

I. Chadd, E. Filiz-Ozbay, and E. Y. Ozbay. Choosing to search: Choice with a default option. *Experimental Economics*, 28(2):237–253, 2025.

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

S. J. Correll, S. Benard, and I. Paik. Getting a job: Is there a motherhood penalty? *American Journal of Sociology*, 112(5):1297–1339, 2007. doi: 10.1086/511799.

S. Cunow, S. Desposato, A. Janusz, and C. Sells. Less is more: The paradox of choice in voting behavior. *Electoral Studies*, 69:102230, 2021.

P. G. Devine. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56(1):5–18, 1989.

C. L. Exley and J. B. Kessler. The gender gap in self-promotion. *The Quarterly Journal of Economics*, 137(3):1345–1381, 2022.

R. Fabregas and W. Zanoni. The migrant penalty in latin america: Experimental evidence from job recruiters. *Journal of Public Economics*, 247:105393, 2025.

R. M. Fernandez and S. Campero. Gender sorting and the glass ceiling in high-tech firms. *ILR Review*, 70(1):73–104, 2017.

I. Fernandez-Mateo and Z. King. Anticipatory sorting and gender segregation in temporary employment. *Management Science*, 57(6):989–1008, 2011.

S. T. Fiske and S. L. Neuberg. A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 23, pages 1–74. Academic Press, New York, 1990.

M. A. Fonseca and A. McCrea. The role of experience in shifting gender beliefs on performance: Experimental evidence. *Journal of Economic Behavior & Organization*, 237:107115, 2025.

X. Gabaix, D. Laibson, G. Moloche, and S. Weinberg. Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, 96(4):1043–1068, 2006.

L. K. Gee. The more you know: Information effects on job application rates in a large field experiment. *Management Science*, 65(5):2077–2094, 2019.

D. T. Gilbert and J. G. Hixon. The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60(4):509–517, 1991.

P. Glick and S. T. Fiske. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, 70(3):491, 1996.

P. Glick and J. Whitehead. Hostility toward men and the perceived stability of male dominance. *Social Psychology*, 41(3):177, 2010.

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

O. Govorun and B. K. Payne. Ego—depletion and prejudice: Separating automatic and controlled components. *Social Cognition*, 24(2):111–136, 2006.

M. S. Hedegaard and J.-R. Tyran. The price of prejudice. *American Economic Journal: Applied Economics*, 10(1):40–63, 2018. doi: 10.1257/app.20150241.

S. S. Iyengar and M. R. Lepper. When choice is demotivating: Can one desire too much of a good

thing? *Journal of Personality and Social Psychology*, 79(6):995–1006, 2000.

Y. Jiang, M. Rossi, and V. K. Tuunainen. Breaking gender discrimination with task attention. In *Proceedings of the 15th Scandinavian Conference on Information Systems (SCIS)*. Association for Information Systems, 2024. URL `https://aisel.aisnet.org/scis2024/18/`.

D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

J. B. Kessler, C. Low, and C. D. Sullivan. Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review*, 109(11):3713–3744, 2019.

J. B. Kessler, C. Low, and X. Shan. Lowering the playing field: Discrimination through sequential spillover effects. *Review of Economics and Statistics*, 108(2):504–513, 2026.

F. Krings, S. Sczesny, and A. Kluge. Stereotypical inferences as mediators of age discrimination: The role of competence and warmth. *British Journal of Management*, 22(2):187–201, 2011. doi: 10.1111/j.1467-8551.2010.00721.x.

A. W. Kruglanski and T. Freund. The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology*, 19(5):448–468, 1983.

Z. Kunda and S. J. Spencer. When do stereotypes come to mind and when do they color judgment? a goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, 129(4):522–544, 2003. doi: 10.1037/0033-2909.129.4.522.

J. N. Lane, K. R. Lakhani, and R. M. Fernandez. Setting gendered expectations? recruiter outreach bias in online tech training programs. *Organization Science*, 35(3):911–927, 2024. Published online May 31, 2023.

A. P. Lenton and M. Francesconi. How humans cognitively manage an abundance of mate options. *Psychological Science*, 21(4):528–533, 2010.

A. P. Lenton and A. Stewart. Changing her ways: The number of options and mate-standard strength impact mate choice strategy and satisfaction. *Judgment and Decision Making*, 3(7):501–511, 2008.

L. Lepore and R. Brown. Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology*, 72(2):275–287, 1997. doi: 10.1037/0022-3514.72.2.275.

G. Lins de Holanda Coelho, P. H. P. Hanel, and L. J. Wolf. The very efficient assessment of need for cognition: Developing a six-item version. *Assessment*, 27(8):1870–1885, 2020. doi: 10.1177/1073191118793208.

D. S. Ma, J. Correll, B. Wittenbrink, Y. Bar-Anan, N. Sriram, and B. A. Nosek. When fatigue turns deadly: The association between fatigue and racial bias in the decision to shoot. *Basic and applied social psychology*, 35(6):515–524, 2013.

C. N. Macrae, A. B. Milne, and G. V. Bodenhausen. Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, 66(1):37–47, 1994.

F. Matějka and A. McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015.

J. J. McCall. Economics of information and job search. *Quarterly Journal of Economics*, 84(1): 113–126, 1970.

M. Moreno, H. Ñopo, J. Saavedra, and M. Torero. Detecting gender and racial discrimination in hiring through monitoring intermediation services: the case of selected occupations in metropolitan lima, peru. *World Development*, 40(2):315–328, 2012.

E. Mourelatos. Does mood affect sexual and gender discrimination in hiring choices? evidence from online experiments. *Journal of Behavioral and Experimental Economics*, 106:102069,

2023.

D. Neumark. Experimental research on labor market discrimination. *Journal of Economic Literature*, 56(3):799–866, 2018.

P. Parasurama and P. Ipeirotis. Human-ai collaboration in fair algorithmic hiring. Working paper, under review, Feb. 2026. URL https://parasurama.github.io/_papers/fair_hiring.pdf.

J. W. Payne, J. R. Bettman, and E. J. Johnson. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):534–552, 1988.

J. W. Payne, J. R. Bettman, and E. J. Johnson. *The Adaptive Decision Maker*. Cambridge University Press, Cambridge, 1993.

A. Peng, B. Nushi, E. Kıcıman, K. Inkpen, S. Suri, and E. Kamar. What you see is what you get? the impact of representation criteria on human bias in hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):125–134, 2019. doi: 10.1609/hcomp.v7i1.5281. URL https://doi.org/10.1609/hcomp.v7i1.5281.

E. S. Phelps. The statistical theory of racism and sexism. *American Economic Review*, 62(4): 659–661, 1972.

F. Pratto and J. A. Bargh. Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology*, 27(1):26–47, 1991.

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

C. Rollero, G. Peter, S. Tartaglia, et al. Psychometric properties of short versions of the ambivalent sexism inventory and ambivalence toward men inventory. *TPM. Testing, Psychometrics, Methodology in Applied Psychology*, 21(2):149–159, 2014.

M. Rothbart, S. Fulero, C. Jensen, J. Howard, and P. Birrell. From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology*, 14 (3):237–255, 1978.

B. Schwartz. *The Paradox of Choice: Why More Is Less*. HarperCollins, New York, 2004.

H. A. Simon. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1): 99–118, 1955.

C. A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003.

T. Sühr, S. Hilgard, and H. Lakkaraju. Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 989–999, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3461702.3462602. URL https://doi.org/10.1145/3461702.3462602.

J. K. Swim, K. J. Aikin, W. S. Hall, and B. A. Hunter. Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, 68(2):199–214, 1995. doi: 10.1037/0022-3514.68.2.199.

S. E. Taylor, S. T. Fiske, N. L. Etcoff, and A. J. Ruderman. Categorical and contextual bases of person memory and stereotyping. *Journal of personality and social psychology*, 36(7):778, 1978.

C. Townsend and B. E. Kahn. The "visual preference heuristic": The influence of visual versus verbal depiction on assortment processing, perceived variety, and choice overload. *Journal of Consumer Research*, 40(5):993–1015, 2014.

M. L. Weitzman. Optimal search for the best alternative. *Econometrica*, 47(3):641–654, 1979.
C. Wolff and N. Keith. Motives relate to cooperation in social dilemmas but have an inconsistent association with leadership evaluation. *Scientific reports*, 9(1):10118, 2019.

A Decision Problem Generation

To construct the decision problems used in our experiment, we drew on publicly available data from the Exley and Kessler (2022) study. In total, we generated 28 decision problems to accommodate the small and large choice set conditions across treatment arms. Twelve of these problems were used in Block 1, which features within-subject variation (six small and six large problems). The remaining sixteen were used in Block 2, which introduces between-subject variation (eight small and eight large problems).

We began by extracting participant profiles from the Exley and Kessler (2022) dataset. For each participant, we constructed a partial score by randomly selecting 10 out of the 20 quiz items and counting the number of correct responses. From each profile, we retained only the participant's ID, gender, total score, and partial score. We then dropped any profiles that lacked a matching counterpart of the opposite gender with the same total score. This ensured that each profile used in our decision problems would have a gender-swapped counterpart, allowing us to generate a paired dataset for incentive compatibility and experimental control (described below).

We then randomly sampled profiles with replacement from the eligible pool to build the 28 decision problems. This sampling process was repeated until we obtained a full set of decision problems that met a set of pre-specified constraints.

First, to simplify analysis and interpretation, we imposed the constraint that no decision problem contains a tie for the highest partial score. This ensures that the performance-maximizing choice is unambiguous in each problem. Second, we applied a series of balancing constraints to facilitate a straightforward measure of discrimination. Specifically, we required that 50% of the decision problems feature a female candidate with the highest partial score—overall and within every relevant subset. These subsets include: Block 1, Block 2, all small choice sets, all large choice sets, and each of the two treatment arms (Small and Large). These constraints allow us to interpret deviations from a 50% hiring rate for female candidates as evidence of discrimination, without confounding from imbalance in candidate quality distributions.

At the end of this process, we produced two final data files. The first contains the full set of 28 decision problems, constructed according to the constraints above. The second is a gender-swapped mirror of the first, in which the gender of each candidate is flipped (male becomes female, and vice versa), while all other information is held constant. This paired structure enables us to isolate the effects of gender on hiring decisions, independent of score distributions or sampling noise.

B Survey Items

B.1 Scale items

In this Appendix, we describe in further detail the *Hostile*, *Denial*, and *Accept* attitude scales, as well as the need for cognition (*Cognition*) scale used in our investigation of the moderators of the effect of large decisions. Each scale is constructed as the average of three survey items where managers were asked to "please indicate how much you agree with the following statements." Responses were on a five-point Likert response scale (Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree).

B.1.1 Hostile

The Hostile Sexism sub-scale of the Ambivalent Sexism Inventory (ASI) measures explicitly antagonistic, derogatory, and adversarial attitudes toward women. Items reflect beliefs that women seek to dominate men, exaggerate or complain unfairly, and are dishonest about their grievances. Hostile sexism is characterized by contempt toward women as a group. A person high in hostile sexism may be more skeptical of a woman's performance claims, attributing success to luck, see hiring women as unfair to men, or view a woman's qualifications as undeserved. We hypothesized that hostile sexism may lead to a decrease in hiring women across the board, more strongly so in large decisions.

The Ambivalent Sexism Inventory was developed by Glick and Fiske (1996) as a 22-item instrument measuring two dimensions: Hostile Sexism (HS) and Benevolent Sexism (BS). The full HS sub-scale contains 11 items. The instrument has been widely validated and is among the most frequently used measures of sexist attitudes in psychology. Glick and Whitehead (2010) developed a shorter 6-item version of the HS sub-scale that correlates highly with the full scale and demonstrates equivalent reliability. A subsequent analysis by Rollero et al. (2014) examined the psychometric properties of the 6-item ASI short form. For the hostile sexism items, all six items loaded substantially on the latent HS factor, with loadings ranging from .57 to .78. In a study across 62 countries, Bosson et al. (2021) further reduced the HS sub-scale to three items. These are the items we use:

1. "Women seek to gain power by getting control over men."
2. "Women exaggerate problems they have at work."
3. "When women lose to men in a fair competition, they typically complain about being discriminated against."

B.1.2 Denial

Modern sexism captures the belief that gender discrimination is no longer a problem in modern society, potentially picking up a more subtle form of bias. Items reflect beliefs that deny the existence or persistence of workplace discrimination against women. This aligns with a "post-feminist" narrative and may reflect either genuine belief in gender equality or motivated reasoning to avoid confronting ongoing inequality. It may also predict dismissal of discrimination concerns and less scrutiny applied to one's own potentially biased decisions. We conjectured that this scale may represent an implicit bias that is more likely to emerge under cognitive strain, and thus predict a treatment effect.

The Modern Sexism Scale was developed by Swim et al. (1995) as an 8-item measure of contemporary, subtle forms of sexism and includes a 5-item Denial of Continuing Discrimination subscale. Wolff and Keith (2019) reduced this sub-scale to three items that are either general or work-related and which coincide with the three highest-loading items in the original Swim et al. (1995) study. We adopt the same method and use the following three items:

1. "Discrimination against women is no longer a problem in the United States." [Reverse-scored]
2. "Women often miss out on good jobs due to sexual discrimination."
3. "Society has reached the point where women and men have equal opportunities for achievement." [Reverse-scored]

B.1.3 Acceptance of Hiring Discrimination

In addition to explicit sexist attitudes, we were interested in the extent to which managers found discriminatory hiring behavior to be "acceptable." The items we use are new, developed to operationalize the two major economic theories of discrimination: taste-based discrimination Becker (1957) and statistical discrimination (Arrow, 1973; Phelps, 1972).

We explicitly ask whether subjects find gender discrimination acceptable in hiring contexts, covering taste-based discrimination, or hiring manager's personal preference to work with a particular gender; statistical discrimination, or discrimination motivated by (real or perceived) differences between genders in job performance; and general managerial discretion, or the right of a manager to favor one gender, generally. Distinct from whether one holds discriminatory beliefs, this measures whether it is acceptable to use gender as a criterion. Subjects who score high on acceptance of hiring discrimination may feel less internal constraint against using gender as a heuristic and may be less likely to second-guess or correct a gender-based choice. We conjecture that this may occur with greater frequency or intensity in larger choice sets, generating a treatment effect. The items are as follows:

1. "Even if a hiring manager prefers one gender, that preference should not influence hiring decisions." [Reverse-scored]
2. "It is acceptable for an employer to consider average gender differences in job performance when making hiring decisions."
3. "It is acceptable for an employer to favor hiring candidates of a particular gender if one prefers working with that gender."

B.1.4 Need for Cognition

The Need for Cognition Scale (NFC/NCS) reflects the extent to which a person is motivated to engage in effortful cognitive processing (Cacioppo and Petty, 1982). Persons high in NFC enjoy thinking, find cognitive tasks intrinsically satisfying, and spontaneously engage in complex information processing even when simpler heuristics are available. Persons low in NFC prefer to avoid effortful thinking, find complex information-processing tasks aversive, and readily adopt simple decision rules and heuristics. In this sense, NFC measures the "preference cost" of systematic processing, as distinct from the "ability cost" measured by the incentivized task in the Math Part of

the experiment. We conjecture that subjects high in NFC are less susceptible to the "choice overload" effect because they enjoy thinking and are more likely to engage systematically with complex information even when decision sets are large.

Cacioppo and Petty (1982) developed a 34-item NFC construct that was later shortened to 18 items by Cacioppo et al. (1984). Lins de Holanda Coelho et al. (2020) developed and validated a six-item form that correlates very highly with the 18-item scale ($r > .90$). We adopt the three highest-loading items from Lins de Holanda Coelho et al. (2020) with factor loadings (in the U.S. sample) of 0.77, 0.81, and 0.82. They are as follows:

1. "I like to have the responsibility of handling a situation that requires a lot of thinking."
2. "Thinking is not my idea of fun." [Reverse-scored]
3. "I would rather do something that requires little thought than something that is sure to challenge my thinking abilities." [Reverse-scored]

C Additional Tables and Figures

Table C.1: Treatment Balance

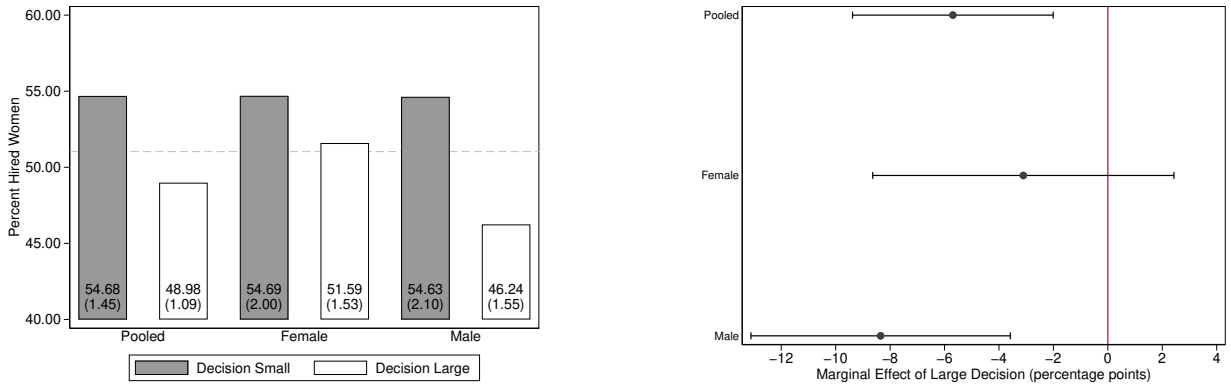
	More Small Mean (SD)	More Large Mean (SD)	p-value
Female	0.466	0.502	0.441
	(0.499)	(0.500)	
Age	36.3	37.1	0.505
	(11.3)	(13.0)	
Heterosexual	0.843	0.823	0.569
	(0.364)	(0.382)	
White	0.640	0.579	0.190
	(0.480)	(0.494)	
Employed FT	0.466	0.469	0.953
	(0.499)	(0.499)	
Low Income	0.508	0.545	0.437
	(0.500)	(0.498)	
No Bachelor's	0.390	0.392	0.957
	(0.488)	(0.488)	
Observations	236	209	

Notes: Means of manager characteristics by treatment arm, with standard deviations in parentheses (shares for binary characteristics), for the 445 managers in the analysis sample. *More Small* managers saw 14 small and 6 large decisions; *More Large* managers saw 6 small and 14 large. Each p-value is from an OLS regression of the characteristic on a *More Small* indicator, with robust standard errors (one observation per manager). The *Female* comparison excludes six managers who indicated they are "transgender, non-binary, or another gender" rather than "Male" or "Female," and the *Age* comparison excludes one manager with an invalid response.

Table C.2: Estimated logistic regression coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Large Decision	−0.128***	−0.195***	−0.199***	−0.063	−0.066	−0.195***	−0.198***
	(0.034)	(0.040)	(0.040)	(0.055)	(0.055)	(0.040)	(0.040)
Female		0.007	−0.003				
		(0.046)	(0.048)				
Large Decision × Female		0.132*	0.133*				
		(0.068)	(0.068)				
Constant	0.086***	0.085***	0.053	0.092**	0.094	0.085***	0.004
	(0.023)	(0.029)	(0.084)	(0.036)	(0.120)	(0.029)	(0.102)
Sample	Pooled	Pooled	Pooled	Female	Female	Male	Male
Demographic Controls			✓		✓		✓
Observations	8,900	8,780	8,760	4,240	4,220	4,540	4,540
Subjects	445	439	438	212	211	227	227

Notes: Log-odds coefficients from logistic regressions of *Choice Female* (one if a female candidate was hired, zero otherwise) on *Large Decision*. Models 1 to 3 use all managers, models 4 and 5 use female managers, and models 6 and 7 use male managers. Models 2 and 3 add a *Female* manager indicator and its interaction with *Large Decision*, so the *Large Decision* coefficient is the effect for male managers; models 2 to 7 exclude six managers who indicated they are "transgender, non-binary, or another gender" only. Models 3, 5, and 7 add demographic controls (age and indicators for heterosexual, white, full-time employed, low income, and no bachelor's degree), and models 3 and 5 also exclude one manager with an invalid age response. Standard errors clustered by subject in parentheses. The marginal effects in Figure 4 come from models 1, 4, and 6. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.



(a) Female candidate raw hiring rates

(b) Regression-adjusted marginal effects

Figure C.1: Large Decision effect conditional on sub-optimal choice

Notes: Subfigure (a): Percent of decisions in which a female candidate was hired, for small decisions (4 candidates; grey bars) and large decisions (16 candidates; white bars), pooling all managers (Pooled) and separately for female and male managers. Labels give each mean with its standard error in parentheses. Subfigure (b) reports the corresponding average marginal effects from logistic regressions. In each, the sample excludes 5,600 (62.92%) observations where the manager chose the candidate with the highest Partial Score.

Table C.3: Robustness of the Large Decision effect

	Full Sample	Agree	Data File	Treatment	Block	Gender Mentioned	Round 1
Large Decision	−0.147***	−0.130***	−0.122***	−0.125**	−0.160***	−0.179***	−0.391**
	(0.032)	(0.034)	(0.047)	(0.052)	(0.041)	(0.053)	(0.191)
Robustness Variable		−0.023	−0.032	−0.027	−0.039	−0.041	
		(0.030)	(0.045)	(0.051)	(0.047)	(0.046)	
Large Decision × Robustness Variable		0.041	−0.010	−0.042	0.091	0.051	
		(0.038)	(0.067)	(0.071)	(0.068)	(0.067)	
Constant	0.102***	0.088***	0.101***	0.105**	0.090***	0.108***	0.234*
	(0.022)	(0.023)	(0.031)	(0.044)	(0.029)	(0.037)	(0.138)
Observations	10,180	8,900	8,900	8,900	8,900	8,900	445
Subjects	509	445	445	445	445	445	445

Notes: Log-odds coefficients from logistic regressions of *Choice Female*, with standard errors clustered by subject in parentheses. *Full Sample* regresses *Choice Female* on *Large Decision* and includes the 64 managers excluded under the pre-registered criteria. *Round 1* uses only each manager's first decision. The other columns add a robustness variable and its interaction with *Large Decision*, so the *Large Decision* coefficient is the effect when the robustness variable equals zero. The robustness variables are: *Agree*, the standardized index of agreement that the manager took the experiment seriously; *Data File*, an indicator for the Gender-Swapped data file; *Treatment*, an indicator for the More Small arm; *Block*, an indicator for Block 2; and *Gender Mentioned*, an indicator that the manager mentioned gender or bias when asked what the study was about. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table C.4: Conditional and Mixed Logit: Effect of Candidate Gender on Selection

	(1)	(2)	(3)	(4)	(5)
Panel A: Conditional Logit					
Female Candidate	0.086***	0.099***	0.099***	0.099***	0.096***
	(0.023)	(0.030)	(0.030)	(0.030)	(0.033)
Female × Large	−0.128***	−0.149***	−0.153***	−0.152***	−0.164***
	(0.033)	(0.042)	(0.042)	(0.042)	(0.050)
Partial Score		0.757***	0.761***	0.749***	0.998***
		(0.032)	(0.032)	(0.032)	(0.045)
Position			−0.043***	−0.043***	0.034***
			(0.005)	(0.005)	(0.005)
Subtraction N				−0.184***	−0.226***
				(0.040)	(0.051)
Panel B: Mixed Logit					
Means					
Female Candidate	0.174***	0.142***	0.147***	0.147***	0.161***
	(0.055)	(0.040)	(0.042)	(0.042)	(0.053)
Female × Large	−0.269***	−0.209***	−0.217***	−0.217***	−0.230***
	(0.094)	(0.064)	(0.065)	(0.065)	(0.087)
Partial Score		1.312***	1.362***	1.346***	2.016***
		(0.067)	(0.073)	(0.072)	(0.104)
Position			−0.060***	−0.060***	0.057***
			(0.007)	(0.007)	(0.009)
Subtractions				−0.222***	−0.347***
				(0.056)	(0.078)
Standard Deviations					
SD(Female)	2.701***	0.024*	0.003	0.005	0.345
	(0.744)	(0.014)	(0.022)	(0.022)	(0.284)
SD(Female × Large)	1.389**	0.987***	0.946***	0.934***	0.707**
	(0.640)	(0.204)	(0.219)	(0.221)	(0.293)
SD(Partial Score)		0.991***	1.036***	1.033***	1.361***
		(0.041)	(0.044)	(0.044)	(0.063)
SD(Position)			0.101***	0.101***	0.120***
			(0.015)	(0.015)	(0.019)
SD(Subtractions)				0.007	0.012
				(0.022)	(0.047)
Sample	Full	Full	Full	Full	Viewed
Observations	87,704	87,704	87,704	87,704	60,981
Subjects	445	445	445	445	435

Notes: Log-odds coefficients from conditional logistic (Panel A) and mixed logistic (Panel B) regressions of whether a candidate profile was chosen. The unit of observation is a candidate profile, and the choice set is the decision. Choice-set fixed effects absorb the main effect of large decisions. Columns add *Partial Score*, *Position* (the profile's position on the screen), and *Subtractions* (the number of subtraction operations in the Partial Score expression); column 5 keeps only profiles whose Partial Scores were viewed. In Panel B, the coefficients are normally distributed, and the table reports their means and standard deviations, estimated by simulated maximum likelihood. Standard errors clustered by subject in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table C.5: Direct Effect of Moderators

	Search		Beliefs		Attitudes		
	Math Ability	Cognition	Direct	Stereotype	Hostile	Denial	Accept
Moderator	0.016	0.016	0.107***	0.021	−0.065***	−0.046**	−0.019
	(0.018)	(0.020)	(0.038)	(0.029)	(0.021)	(0.019)	(0.021)
Constant	0.023	0.023	0.035*	0.035	0.022	0.024	0.023
	(0.020)	(0.019)	(0.020)	(0.027)	(0.019)	(0.019)	(0.019)

Notes: Log-odds coefficients from logistic regressions of *Choice Female* on a single manager-level moderator; each column uses a different moderator. Moderators are defined and scaled as in Table 2: Search and Attitudes moderators are standardized, and Beliefs moderators run from -2 to $+2$. Standard errors clustered by subject in parentheses. $N = 8{,}900$ decisions from 445 managers in every model. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

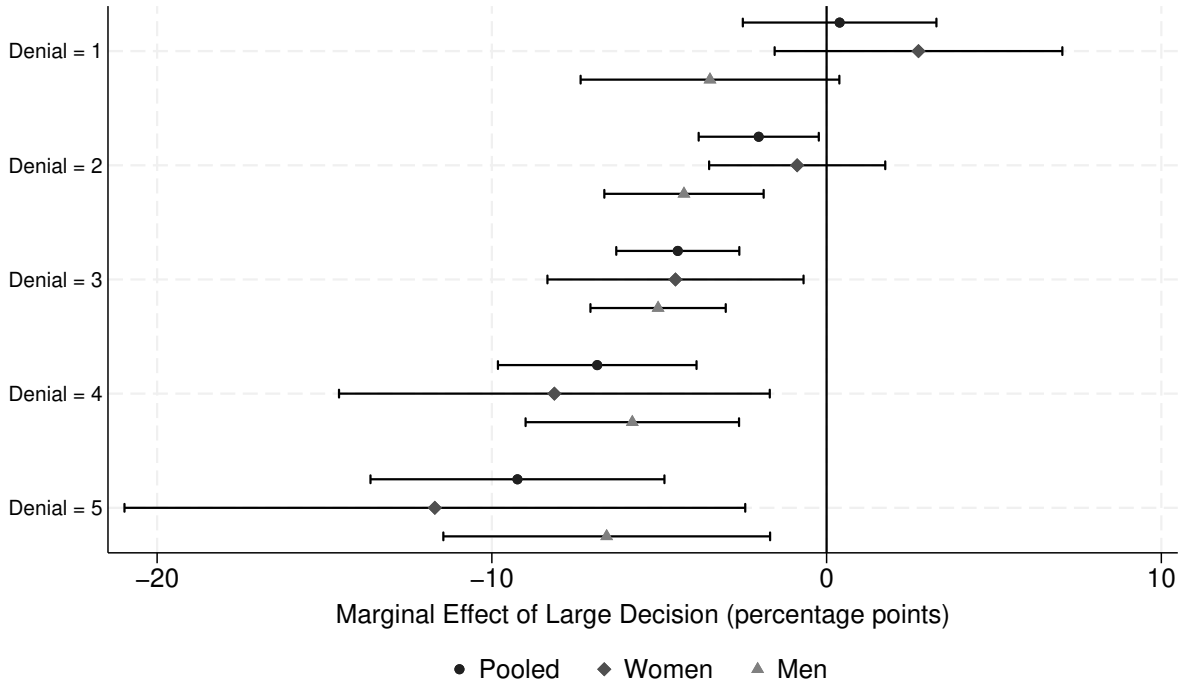


Figure C.2: Denial of sex-based discrimination moderates the Large Decision effect

Notes: Average marginal effect of a large decision (16 candidates) relative to a small decision (4 candidates) on the probability of hiring a female candidate, in percentage points, with 95% confidence intervals, evaluated at each integer value of the denial of discrimination index (*Denial*). *Denial* is the average of three agreement items from the Denial of Continuing Discrimination subscale of the Modern Sexism Scale (e.g., Swim et al., 1995), ranging from one (least in denial) to five (most in denial); see Appendix B. Estimates come from logistic regressions of *Choice Female* on *Large Decision*, unstandardized *Denial*, and their interaction, with standard errors clustered by subject, estimated separately for all managers (Pooled), female managers, and male managers. Samples are as in Figure 4.

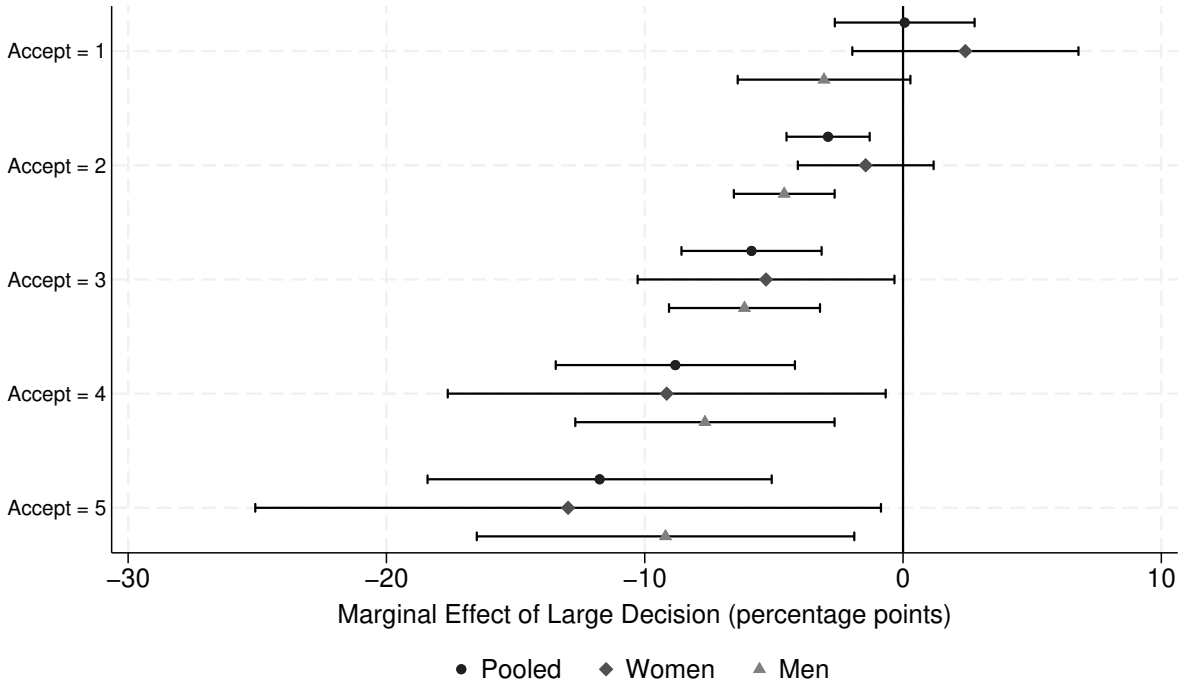


Figure C.3: Acceptance of discriminatory hiring practices moderates the Large Decision effect

Notes: Average marginal effect of a large decision (16 candidates) relative to a small decision (4 candidates) on the probability of hiring a female candidate, in percentage points, with 95% confidence intervals, evaluated at each integer value of the acceptance of hiring discrimination index (*Accept*). *Accept* is the average of three agreement items, written for this study, on whether discriminatory hiring practices are acceptable, ranging from one (least accepting) to five (most accepting); see Appendix B. Estimates come from logistic regressions of *Choice Female* on *Large Decision*, unstandardized *Accept*, and their interaction, with standard errors clustered by subject, estimated separately for all managers (Pooled), female managers, and male managers. Samples are as in Figure 4.

Table C.6: Math ability effect on search

	Candidates Viewed			Time Viewed (sec.)	
	N Viewed	Viewed None	Viewed All	Total	Per Candidate
	(1)	(2)	(3)	(4)	(5)
Large Decision	6.748***	0.506**	−2.010***	23.021***	−0.739***
	(0.228)	(0.228)	(0.114)	(1.353)	(0.195)
Math Ability	0.238***	−0.919***	0.685***	−3.534***	−1.353***
	(0.038)	(0.189)	(0.118)	(0.658)	(0.206)
Large Decision × Math Ability	1.410***	0.016	−0.181	2.060	0.479**
	(0.218)	(0.214)	(0.118)	(1.268)	(0.204)
Constant	3.578***	−3.232***	1.876***	16.630***	4.769***
	(0.043)	(0.224)	(0.113)	(0.651)	(0.197)
Model	OLS	Logit	Logit	OLS	OLS
Observations	8,900	8,900	8,900	8,895	8,325
Subjects	445	445	445	445	444

Notes: Coefficients from regressions of search behavior on *Large Decision*, *Math Ability*, and their interaction. Standard errors clustered by subject in parentheses. *Math Ability* is the number of arithmetic problems solved correctly in the 60-second Math Part, standardized, so the *Large Decision* coefficient is the effect at mean ability and the constant is the small-decision mean at mean ability. The dependent variables are defined as in Table 4. *Total* excludes five decisions with missing time data; *Per Candidate* additionally excludes 570 decisions in which no Partial Score was viewed. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table C.7: Returns to Partial Score

	Outcome: Chosen	
	(1)	(2)
Female Candidate	0.207	0.195
	(0.141)	(0.183)
Female × Large	−0.140***	−0.108
	(0.045)	(0.291)
Partial Score	0.766***	0.797***
	(0.034)	(0.045)
Female × Partial Score	−0.018	−0.015
	(0.021)	(0.028)
Large × Partial Score		−0.046
		(0.044)
Female × Large × Partial Score		−0.005
		(0.043)

Notes: Log-odds coefficients from conditional logistic regressions of whether a candidate profile was chosen (*Chosen*). The unit of observation is a candidate profile and the choice set is the decision, so choice-set fixed effects absorb the main effect of large decisions. *Partial Score* is the profile's Partial Score in points. The sample is all $N = 87{,}704$ candidate profiles shown to 445 managers, whether or not viewed, in both models. Standard errors clustered by subject in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.