UNIVERSITY OF EXETER | BUSINESS SCHOOL

Department of Economics Discussion Papers

ISSN 1473-3307

The role of shortlisting in shifting gender beliefs on performance: experimental evidence

Miguel A. Fonseca and Ashley McCrea

Paper number 23/15

RePEc page: https://ideas.repec.org/s/exe/wpaper.html

The role of shortlisting in shifting gender beliefs on performance: experimental evidence*

Miguel A. Fonseca†, Ashley McCrea‡

November 7, 2023

Abstract

In labour markets, women are often underrepresented relative to men. This underrepresentation may be due to inaccurate beliefs about ability across genders. Inaccurate beliefs might cause a sampling problem: to have accurate beliefs about a group, one must first collect information about that group. However, inaccurate beliefs may persist due to biased belief updating. We run a stylized hiring experiment to disentangle these two effects. We ask participants to create shortlists from a male and a female pool of workers and give them feedback on the skill of those they shortlist. Based on that information, participants hire workers, and provide us with their beliefs about the distribution of skills in the male and female pots. We study how recruiters update their beliefs as a function of their past shortlisting behaviour, and how they shortlist given their beliefs. As expected, participants were more likely to sample from the pool with the highest subjective mean quality (on average men) and lowest subject variance. Participants were not Bayesian updaters but there were no gender-specific biases in updating. Sampling more from a pool and, somewhat surprisingly, greater time spent engaging in sampling behaviour yield more accurate beliefs.

Keywords: Inaccurate Statistical Discrimination, Belief Updating, Gender, Shortlisting, Chess

JEL Codes: C91, D83, J71, J78, M51

*We thank Katherine B. Coffman, Boon Han Koh, and Edwin Ip for detailed discussions. We also thank Ben Balmford, Surajeet Chakravarty, Christine Exley, J.Braxton Gately, Brit Grosskopf, Neeraja Gupta, Oliver Hauser, Stephen Nei, Dario Sansone, Mattie Toma, and conference participants at MBEES/MBEPS 2023, ESA 2023 European Meeting, and Nordic Conference 2023 for their valuable comments and suggestions. Financial support from the ESRC is gratefully acknowledged.

†University of Exeter. E-mail: *m.a.fonseca@exeter.ac.uk*.

‡University of Exeter. E-mail: *am749@exeter.ac.uk*. Postal address: University of Exeter Business School, Streatham Court, Rennes Drive, Exeter EX4 4ST, United Kingdom.

1 Introduction

In labour markets, women are often underrepresented relative to men (Reuben et al., 2014; Cheryan et al., 2017). In UK politics, only 35% of House of Commons MPs and 28% of the members of the House of Lords are female (Uberoi and Mansfield, 2022). In the private sector, 37.7% of directors of FTSE100 companies are women (Buchanan et al., 2022). In law enforcement, 32% of UK police officers are female; and in academia, only 28% of UK professorships are held by women (Uberoi and Mansfield, 2022; HESA, 2022). This underrepresentation has serious consequences for economic performance and growth, as institutions are under-recruiting talent from roughly 50% of the talent pool.

Labour market discrimination between genders has been well studied and a prominent explanation for discrimination in the literature is statistical discrimination (Phelps, 1972; Arrow, 1977).[1,2] Recently, greater attention has been given to inaccurate statistical discrimination (Bordalo et al., 2016; Bohren et al., 2019, 2020; Feld et al., 2022), in which employers do not have access to an unbiased signal about productivity, but rely on (potentially inaccurate) subjective beliefs about a group. Inaccurate statistical discrimination occurs when an employer holds inaccurate beliefs about the group averages and other potential characteristics, and then uses this inaccurate information as proxies in their decision making.

Inaccurate statistical discrimination can potentially cause a sampling issue: in order to have accurate beliefs about a group, one must collect information about that group. However, in labour markets, it is costly to collect information (Bartoš et al., 2016; Fershtman and Pavan, 2021). If an employer has negative beliefs about a group, the employer may believe there are low returns to searching within that group and neglect it. As a result, that employer may not update beliefs about that group and continue to discriminate against it in their hiring practices. This reinforces the difference in beliefs across these groups (Lepage, 2023)[3]. For this reason, shortlisting restrictions such as the "Rooney Rule" in the NFL[4] may aid in correcting any distortions in beliefs about groups (DuBois, 2015).

There are a number of potential reasons as to why an employer may not hold correct beliefs about men and women, when they perform similarly. We focus on two mechanisms. The first mechanism involves sampling and search arguments. It may be that employers may not have had sufficient opportunity to search for and hire workers, meaning that employers have not had a chance to learn

[1]The other main theory of discrimination is taste-based discrimination (Becker, 1957), which can be described as having a disutility or utility from interacting with members of certain groups, creating a preference to either hire or avoid hiring from those particular groups.

[2]See Riach and Rich (2002); Bertrand and Duflo (2017); Blau and Kahn (2017); Neumark (2018) for reviews of the literature on discrimination in labour markets

[3]For example, if an employer believes that women are on average worse in STEM jobs, the employer is less likely to search for, interview, and hire women. By not searching and hiring women, they obtain new information about women's productivity. This then does not allow for the opportunity to correct their beliefs, potentially reinforcing already biased beliefs.

[4]The Rooney Rule is a policy that was introduced in the NFL in 2003. It required that NFL teams interview at least one ethnic-minority candidate for all head coaching and senior executive roles.

and update their beliefs about the abilities of men and women. It could also be that the employer believes, for example, that men are better than women, leading to them sampling less women and more men, resulting in more accurate beliefs about men and inaccurate beliefs about women.

The second mechanism is that employers may be updating beliefs differently for different, but similarly performing, populations, which has the potential to cause belief differences to persist. This difference in belief updating for different populations may come about in a number of ways. One may be, for example, some form of motivated reasoning, where an employer may want to hold certain beliefs, and might distort their belief updating in such a way that leads to them to be able to hold these beliefs (Eil and Rao, 2011; Möbius et al., 2022). For example, they may wish to hold beliefs in line with particular gendered stereotypes, that one group is better at a particular task than another group (Coffman, 2014; Bordalo et al., 2016, 2019).

In this paper we present evidence from a laboratory experiment designed to disentangle and compare these mechanisms while controlling for initial priors. We are interested in studying information search and belief updating about performance across genders in a highly stereotyped context. While there are many domains that suffer from gender stereotypes, not all are amenable to laboratory experiments, through which we can elicit belief information. We chose performance in chess. Chess is a male-dominated activity: women account for only 11% of FIDE rated players at all levels, 1% of the top-100 players, and only 2% of Grandmasters (Smerdon, 2022). Gender stereotypes are ingrained in chess to the extent that Nigel Short, a grandmaster and the vice-president of FIDE, the governing body for chess, when asked about the gender gap at the top of competititve chess stated that "men are "hardwired" to be better at the game than women" (Ellis-Petersen, 2015) [5]. Importantly, chess ability is a function of an individual's capacity to do backward induction, and working memory. There is no evidence of gender differences in these cognitive functions (Robert and Savoie, 2006; Harness et al., 2008). As such, chess is a promising domain to study discrimination in the lab, as there is little basis for differential performance.

In this study, we run a stylised hiring task paired with a belief updating task. Our subjects were placed in the role of recruiters. At the start of the experiment, we measured recruiters' prior beliefs about the distributional performance for men and women. Then, in each round of the experiment, recruiters had to construct a shortlist of 10 individuals that could be drawn from an all-female pool and/or an all-male pool. Recruiters then found out the productivity of each shortlisted candidate and selected one. Then, we asked recruiters to update their beliefs about the distribution of performance for each gender. The payoff to the recruiter was proportional to the selected candidate's performance in a real effort task: the number of chess puzzles successfully solved in 5 minutes (based on the performance data of Chess.com's puzzle rush task). Focusing on puzzles instead of actual games removes two dimensions which may introduce gender differences in performance: competitiveness

[5] Some argue that the perceived gender difference in chess playing ability is down to the fact that less women choose to play chess competitively (Ingle, 2021). Other social factors, such as starting age and perseverance, may play a role (Blanch et al., 2015; Blanch, 2016). In addition, this is a domain in which stereotype threat seems to affect the performance of women (Maass et al., 2008; Rothgerber and Wolsiefer, 2014).

(Niederle and Vesterlund, 2007), and risk/ambiguity attitudes (Borghans et al., 2009), although see (Crosetto and Filippin, 2016)[6].

We are interested in both the way recruiters update their beliefs as a function of their past shortlisting (i.e. sampling) behaviour, as well as how they sample given their current beliefs. We find an initial difference in priors about means but not priors about variances in the context of chess puzzles. We consider the two mechanisms in turn, starting with differences in belief updating. We find no difference in how people diverge from a Bayesian benchmark distribution when updating for women compared to men. This suggests that differences in belief updating might not cause differences in beliefs in this context.

The second mechanism - the sampling and search mechanism - comprises two components. The first component is whether the relative subjective belief distributions drive behaviour. We find that both the difference in subjective means and variance impact sampling behaviour. The second component is where having more observations from a group does in fact impact the accuracy of beliefs. We find evidence that greater sampling does generate more accurate beliefs. Interestingly, greater time spent engaging in sampling behaviour also seems to give more accurate beliefs. This suggests that inaccurate differences in beliefs between two groups may indeed cause inaccurate beliefs to persist. Additionally, we find lighter forms of affirmative action may not elicit a change in sampling behaviour, and as such, provide no stimulus to correct inaccurate beliefs.

The rest of the paper proceeds as follows. Section 2 presents the experimental design. Section 3 presents the results the results of our confirmatory analysis. Section 4 presents the results of our exploratory analysis. Section 5 discusses these results and concludes.

2 Experimental Design and Procedures

Consistent with the previous stylized hiring experiments, our participants act as 'employers'. In each experimental round, employers had to (i) state their beliefs about the distribution of performance in two pools of potential workers: men and women; (ii) construct a shortlist of 10 candidates from either/both pools; (iii) select one candidate after observing the true performance of the shortlisted candidates; (iv) revise their beliefs about each pool. The payoff to recruiters was a function of the performance by their chosen candidate in a real effort task.

2.1 The real-effort task and the worker pools

The two pools of workers were created based on publicly available information on profiles on chess.com where players have completed 5-min puzzle rush, a timed chess puzzle challenge on chess.com. In it, players have 5 minutes to complete as many puzzles as possible. Puzzles increase in difficulty as the player progresses through the challenge. The challenge ends either after 5

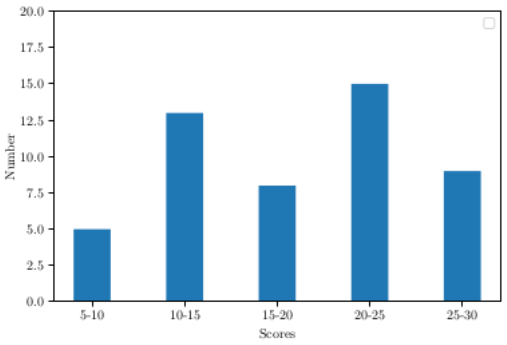
[6] A caveat that risk attitudes are context-dependent is warranted here (Weber et al., 2002). An individual's estimated risk attitude will depend on the measure used (Crosetto and Filippin, 2013).

Table 1: Summary Statistics for Performance Distribution of the Pots

	Mean	S.D.	Var	Min	Max
Men/'Pot A'	18.33	5.90	34.89	7.00	28.9
Women/'Pot B'	18.29	5.90	34.81	7.50	27.00

	5-10	10-15	15-20	20-25	25-30
Interval Frequency	5	13	8	15	9
Midpoint Mean	18.5				
Midpoint Var	41.83				

Figure 1: True Distribution for both Men and Women



minutes, or if the player answers 3 puzzles incorrectly.

Across the publicly available profiles on chess.com, we collected profiles of both genders with Elo ratings between 1600 and 1800, to ensure sufficient proficiency at the task [7]. We classified these profiled by presented gender based on a publicly available profile picture, or a publicly available name or a chess title.[8] We excluded any profile who did not have any recorded attempts at the 5-minute puzzle rush, as well as any profiles we could not classify with a gender. Of those profiles that could be classified by gender and had attempted puzzle rush, we calculated the average performance score from their recent history of play. This score acts as our performance measure within the main experiment. Across all individuals who attempted 5-minute puzzle rush at least once, we created two pots of 50 individuals. One pot had only male players, while the other had only female players. We constructed these pots such that both pots had a similar performance distribution (see Table 1 for summary statistics on both pots; see Figure 1 for bar chart) and the same number of observations in a score range (for example, between 5 and 10). Importantly, using chess puzzles rules out two behavioural factors that may rationalise differences in performance across genders in actual chess matches: risk attitudes and competitiveness (Niederle and Vesterlund, 2007; Gerdes

[7]Elo is a standard performance measure of chess rating.
[8]A player that holds the title of Female Master (FM) or Female Grandmaster (FGM) is necessarily female.

and Gränsmark, 2010; Smerdon, 2022) . Chess puzzles are fundamentally a test of cognitive ability. Yet, gender stereotypes about chess ability may persist in this setting.

2.2 An experimental round

The experiment consisted of a number of rounds, which were divided into sub-tasks. We explain each task in turn.

2.2.1 Belief Elicitation

At the start of each round, we elicited participants' full distribution of beliefs about the performance of the individuals in each pot. That is, instead of eliciting the first moment of their belief distribution (i.e., the mean, median or mode), as is common in the literature, we elicited the full distribution.

The belief elicitation mechanism worked as follows. For each pot, we presented the participant with a set of 5 sliders, each referring to a discrete set of intervals of possible performance levels: [5,10), [10, 15), [15, 20), [20, 25), and [25, 30) (see Figures Ax-Az in the Appendix for screenshots of the experimental interface). The lowest performance across the pots was 7 and the highest score across all pots in all sets was 28.9. Employers used the sliders to indicate their belief about the relative frequency of each performance interval in each pot, which corresponded to the probability of drawing an individual whose performance was in that range. The stated probabilities in all 5 sliders had to sum to 100%.

In the first round, the sliders were set to 0. In all subsequent rounds, the starting position of the sliders is the belief distribution given in the previous round.

In the first round, we incentivised beliefs through a generalised quadratic scoring rule (Harrison et al., 2017). We implemented this scoring rule as follows. For each performance interval, a payoff is associated with that interval, based on how many individuals that the participant believes got a score in that interval. The payoff for an interval is:

$$S = 1.5 + 1.5 \left[(2 \times r_k - \sum_{i=1}^{5} r_i^2 \right] \tag{1}$$

where r_k is the number of individuals assigned to an interval.

At the end of the experiment, a worker was randomly drawn from one of the two pots. The employer was then paid the payoff associated with the score pertaining to the interval to which the randomly drawn worker belonged. As employers adjusted the sliders for each interval, they could see in real time the payoff they would receive should the random worker be drawn from that interval. This tool had the advantage of not requiring an explanation about the underlying formula that determined their payoff.

In all later rounds, we did not incentivise beliefs: the only change in how the sliders were presented was that the payment values next to the sliders were removed. We choose to only elicit beliefs in

the first round of the experiment, prior to any hiring decision, because incentivised beliefs would likely interfere with the hiring task. Note that by shortlisting from a pot, an employer acquires information about that pot; as such, incentivising the belief elicitation in later rounds might have artificially shifted employers' behaviour to more exploration than they would otherwise chosen.

We do not believe this to be a problem, however, as there is evidence that introspection may perform just as well as more complicated incentivised belief elicitation mechanisms (Charness et al., 2021).

2.2.2 Worker Shortlisting and Hiring

After beliefs were elicited, participants completed the worker hire task. This started with the creation of a 'shortlist' of workers. The participants first stated how many workers they wished to sample from each of the two pots; for each pot they could state any number between 0 and 10, subject to the constraint that both numbers had to add up to 10. The software then randomly selected the specified number of workers from either pot, and added it to the shortlist.

The employer then saw the profiles of the drawn candidates. Each profile contained that worker's actual performance and the pot from which they were drawn. The employer then had to choose one worker from the shortlist.

For this task, the employer earned a payoff equal to the score of the chosen worker's multiplied by £0.05 (rounded down).

2.2.3 End of round

At the end of each round, participants were presented with a random number from 1 to 20. If the random number was 18 or less, they played another round of the experiment. If the random number was 19 or 20, they proceeded to the final stages of the study. This corresponds to a probability of 10% that the game ends each round.

We drew a random sequence of numbers prior to the start of the data collection and used that sequence on all participants. Each participant played 6 rounds of the experiment.

After all rounds were completed, we elicited participants' individual risk aversion using the bomb risk elicitation task (Crosetto and Filippin, 2013).

Participants were paid for three events. Firstly, they were paid for the outcome of the belief elicitation task in the first round. Secondly, they were paid for the performance of the chosen worker in the final round, at a rate of £0.05 multiplied by the selected worker's score, rounded down. Thirdly, they were paid for their decisions in the BRET at a rate of £0.03 for each box opened. Participants also received £3 for completing the experiment.

The average payoff was £2.90 with a show up fee of £3, for an experiment which lasted on average 25 minutes. Our experimental software was programmed in oTree (Chen et al., 2016). We recruited 600 participants were recruited through Prolific.co. They were 37.52 years old on average; 50.8% were women; 44.6% were in full employment.

The experiment (methods, sampling, procedures and statistical analysis) was pre-registered on OSF [`https://osf.io/8n5sq/`]. All statistical analysis reported in the paper is pre-registered unless signposted as exploratory.

2.3 Hypotheses

Our first hypothesis concerns prior beliefs. As we argued earlier, chess is a male-dominated activity, in which there are ingrained gender stereotypes about performance. It is possible that women are perceived to have higher variance in ability, or that the beliefs about women are noisier and result in higher variance (Aigner and Cain, 1977). As a result, we hypothesise that these stereotypes will carry over to our experimental setting.

Hypothesis 1a *(Mean of Prior Beliefs): Employers' beliefs about the performance of male workers will have a higher mean than beliefs about performance of female workers.*

Hypothesis 1b *(Variance of Prior Beliefs): Employers' beliefs about the performance of male workers will have a lower variance than beliefs about performance of female workers.*

Our second hypothesis pertains to how employers update their beliefs conditional on new information. Gender bias in performance could manifest itself in how individuals update their beliefs given new information about male and female performance. It may be that stereotypes drive gender difference and as such participants update their beliefs in a direction that would match these gender stereotypes (Bordalo et al., 2019). Additionally, it may be that participants wish to hold these gender stereotyped beliefs, further encouraging belief updating in this direction through motivated reasoning (Eil and Rao, 2011; Möbius et al., 2022; Thaler, 2020).

Hypothesis 2 *(Belief Updating): Employers will update their beliefs differently for women than for men.*

An important element of the argument we discussed in the introduction is that beliefs may persist through belief-driven sampling behaviour, such that if an individual only ever samples their a priori better perceived group, they will never hire from their worst perceived group.

The theoretical foundation for this is thus. Under Bayesian inference, when an individual draws a very large sample, their prior beliefs are crowded out by the observed data. Under the assumption that a large sample is more likely to be reflective of the true population, an individual's posterior belief will be more in line with this representative sample. In our case, a larger sample occurs with a greater number of observations being drawn across the rounds played. As such, with a greater number of observations, an individual's posterior belief will converge to the true distribution, which is analogous to saying that the participant will have a greater level of accuracy over the distribution.

Hypothesis 3 *(Sampling and Belief Accuracy): The more sampling from a given pot, the more accurate beliefs will be.*

Given the previous discussion about the relation between the accuracy of beliefs and sampling, it may be that one reason there is a difference in beliefs about similar groups is that individuals have not sample sufficiently from one or both groups. One policy that may help is the use of some form of soft affirmative action, as this then requires that an employer at the very least samples from a group. One such form of affirmative action is the previously mentioned Rooney Rule. Given that the Rooney Rule requires that a participant samples from a pot, previously under-sampled groups will be sampled more, and greater sampling will tend towards more accurate beliefs. As such, the implementation of the Rooney Rule may lead to more accurate beliefs about the two populations.

Hypothesis 4 *(Sampling Restrictions and Belief Accuracy): Shortlisting restrictions on the number of workers an employer can shortlist from a given gender will improve the accuracy of beliefs.*

We now turn to the act of shortlisting candidates. We expect that the likelihood of sampling from a particular candidate pool will be positively correlated with the mean belief about performance, and negatively correlated with the variance in beliefs about performance of that candidate pool (risk aversion).

Hypothesis 5 *(Sampling Behaviour): Employers are more likely to sample from a candidate pool the higher the mean beliefs about performance and the lower the variance in their beliefs.*

2.4 Experimental Design

In order to test the hypotheses set out above, we constructed six different treatments, which we describe next. Numbers in each treatment and balance tests are available in appendix I.

BASELINE (**BASE**): This treatment is the baseline as described above. The participant was able sample from the pots in any combination they choose. That is, they can choose, for example, 10 men and 0 women; 9 men and 1 women; and so on.

Probabilistic Exogeneity (PE): In this treatment, players were not guaranteed to draw the shortlist that they request. At the point of creating a shortlist, players still gave their preference over the shortlist composition. There was then a 50% chance they receive a shortlist constructed in their requested way, and a 50% chance they get a shortlist randomly selected from the other possible shortlists, where the other shortlists have an equal probability of being chosen.

Shortlisting Restrictions (SR): In this treatment, players were required to draw at least one profile from each pot at the stage of creating a shortlist.

Only Beliefs-Framing (OB-F): In this treatment, employers only completed the belief elicitation task and did not decide how to shortlist workers. In contrast to the treatments with the shortlisting task, we incentivised all belief elicitation. At the start of each round, participants gave their beliefs

with the same generalised quadratic scoring rule mechanism described for the first round of the baseline. Employers knew that one pot contains only men and the other contains only women. We then drew a random number of workers from each pot, totalling up to 10; that is, the number of men and women sampled in any given round to an employer was random. We then presented these profiles to employers, each of which contained the worker's score and the pot they came from. Employers did not hire a worker based on the shortlist. The number of rounds completed was random; participants were paid for their beliefs in the final round, as opposed to the first round as in treatments that included a shortlisting decision.

Only Beliefs – No Framing (OB-NF): This treatment was identical to OB-F, other that it was neutrally framed. We told employers that one pot is called "Pot-A" and the other pot is called "Pot-B." When presented with the profiles, they were told whether they came from "Pot-A" or "Pot-B"; there was no mention of gender in the instructions and no gender information was revealed to employers.

Only Beliefs – Shortlisting Samples (OB-SS): This treatment was be run after the initial data collection of the other five treatments. This treatment functioned similarly to Only Beliefs – Framing. The only difference was that the samples presented to participants follow similar distributions to what was seen in the baseline treatment (BASE). Using the data from the initial prior belief elicitation in the shortlisting treatments, we split participants into 3 groups – those who have had a higher mean for men; those who had a higher mean for women; and those with the same mean for both distributions. Within each group, the number of men and women shortlisted in each round is averaged, such that we have an average of the sampling behaviour within each group for each round. This creates 3 sets of paths of sampling for the different groups of initial priors. After participants in this treatment give their initial priors, they are matched with one of these sampling paths based on the mean of their given priors. Across the 6 rounds, these participants were presented with a randomly drawn sample in accordance with this sampling path.

The three OB treatments allow us to test for the effect of different sampling approaches on belief updating while avoiding the simultaneity problem inherent to the shortlisting treatments. The framing manipulation in the OB treatments allows us to identify any gender bias, as we will explain in detail when describing the econometric approach. The shortlisting treatments naturally address the relationship between beliefs and shortlisting behaviour.

One concern is that participants who choose to sample more men or more women are a) likely to hold better beliefs about the individuals in that pot and b) may have some desire to hold beliefs about those pots. The probabilistic exogeneity treatment (PE) allows us to do two things. Firstly, it allows us to see what participants wish to sample given their belief set, with some greater probability they will receive that sample. Secondly, it allows us to ensure that participants who do wish to sample

in a certain way do update their beliefs more correctly if they receive a sample that might not be in accordance with their prior beliefs.

Within our experiment, we collect both incentivised beliefs and unincentivized beliefs. We collect incentivised beliefs in the "beliefs only treatments" and unincentivized beliefs in the "shortlisting treatments." We choose to collect unincentivised beliefs in the shortlisting treatments as the presence of an incentivised belief elicitation task may skew the behaviour in favour of exploration when the intention of the shortlisting treatments is to look explicitly at how one's initial beliefs impact their shortlisting behaviour. As we collect these different types of beliefs, we wished to check that there was no significant differences in belief updating between incentivised and unincentivized beliefs. We created a beliefs only treatment that would present similar samples based on initial beliefs to those that were collected in the baseline (BASE). This ensured that the samples seen in OB-SS and BASE were comparable, allowing us to directly look at the effect of incentivised beliefs on behaviour.

One of our hypotheses relates to the idea that some form of Rooney rule or restriction on shortlisting may improve the accuracy of beliefs about pots by encouraging a minimum amount of shortlisting of the pots. The shortlisting restrictions treatment (SR) requires that the participants must sample at least one from each pot. This removes the possibility that a participant never learns about one of the groups and as such, provides a route through which accuracy may be improved.

3 Results

We now present our main results. We begin by investigating whether employers' priors are different for each gender, and whether gender impacts their belief updating in this context. From here, we study how beliefs impact shortlisting behaviour in this setting. Finally, we address questions surrounding whether increased sampling from a group impacts the accuracy of beliefs.

3.1 Initial Conditions

3.1.1 Are priors about men and women in chess different?

We first examine whether there are gender differences in priors, both in means and in variances. To this effect we use all round 1 data from all treatments except OB-NF as this treatment was neutrally framed.

Figure 2 presents the average distribution of beliefs about the performances of men and women in round 1. The horizontal axis denotes the intervals in which scores may lie. The vertical axis denotes the participants' beliefs about the number of men and women who scored within each interval. The figures suggest there is a difference in distributions, primarily driven by the extremes: the belief distribution about men places has more weight than that of women's performance in the top 2 categories, while the reverse is true for the bottom 2 categories.

Given this visual difference in distributions, we formally test whether there is a difference in

Figure 2: Average Prior Beliefs about Men and Women's Performance

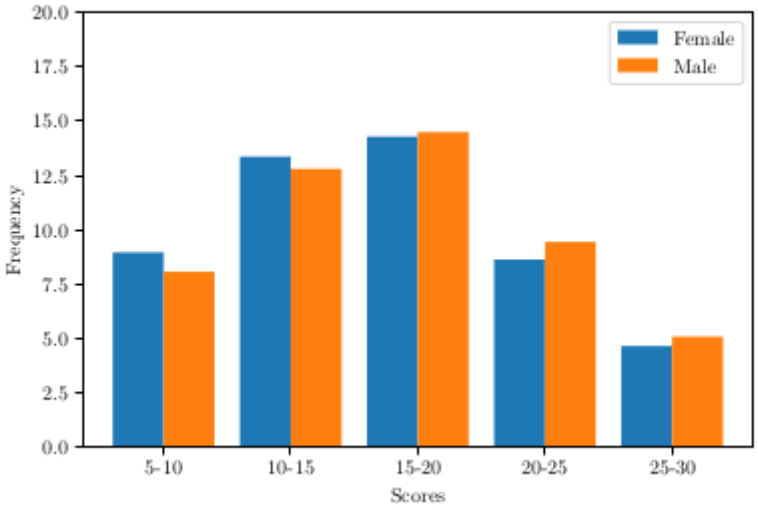
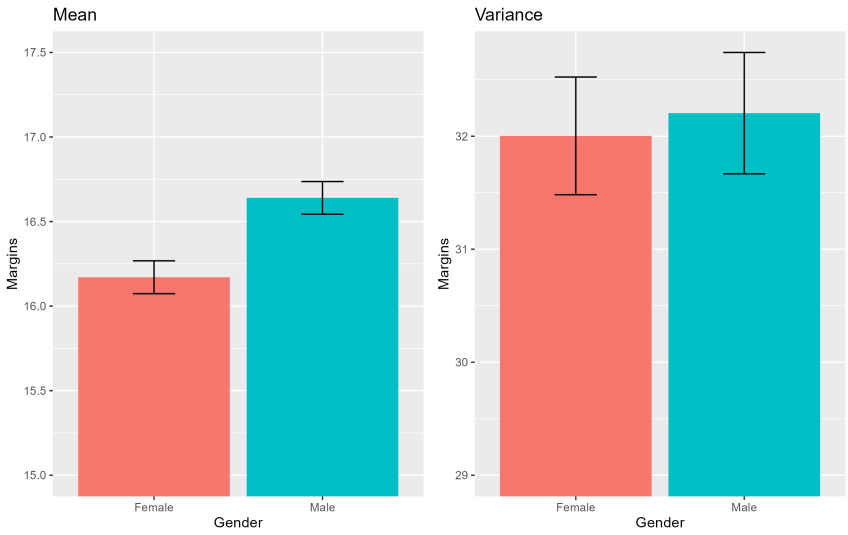


Figure 3: Margins of the Mean and Variance of Performance by Gender



two moments of the belief distributions – their mean and variance. The margins of the results of this analysis are displayed graphically in Figure 3. The full details of this analysis are available in Appendix 6.3.1 in Table 6.

Figure 3 presents the margins of the beliefs about the means and variances of the distributions by gender. The left side figure displays the margins by gender about the means of the distributions. The right hand side displays the margins by gender about the variances of the distributions. Within each figure, the left bar represents the beliefs about the female pot and the right bar represents the beliefs about the male pot. Looking at these figures, one can clearly see that the average beliefs about the means are such that participants perceived the men to have performed better, whereas there is no clear difference in beliefs about the variances.

In the full analysis in Table 6 in Appendix 6.3.1, when regressing the mean of the elicited distribution on a dummy for whether the distribution represents the female workers, the coefficient on this dummy in regression (2) is negative and significant. This coefficient of -0.469 implies that women are believed to perform worse by roughly 3% on average. When we run a similar regression, but instead regressing the variance of the distributions on this dummy, we get a negative but statistically insignificant coefficient, suggesting that there is no difference in the beliefs about the variance of the distributions by gender.

It is important to note that while the difference in puzzles solved is not that large at 3%, this could be due to the belief elicitation mechanism we used. The generalised quadratic scoring rule preserves the true relative ranking of probabilities but tends to generate probability distributions that are flatter than the true distributions, likely because of risk aversion (Harrison et al. (2017)). As such, we should take this difference as a lower bound on the true difference in perceived ability. This argument extends to the lack of difference in perceived variance.

Result 1a: *The mean of the distribution of prior beliefs about women are lower than that of men.*

Result 1b: *The variance of the distributions of prior beliefs about men and women are not different.*

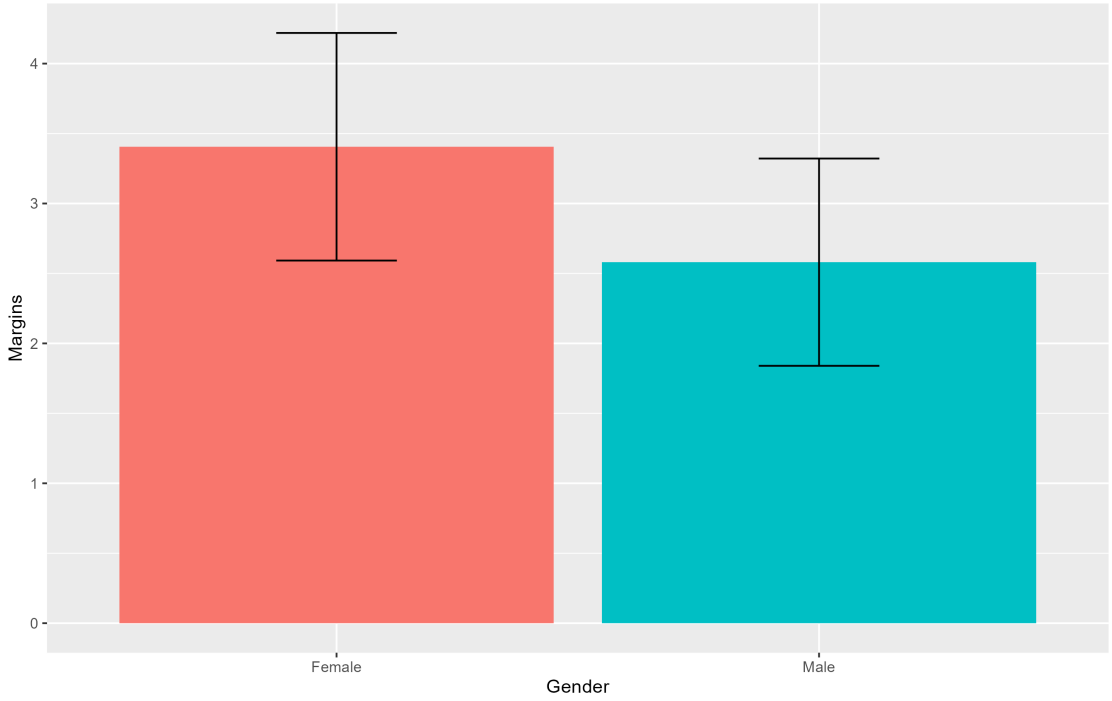
3.2 Mechanism 1: Differences in Belief Updating

3.2.1 Is there a gender-based difference in belief updating?

Having established that there is a difference in initial beliefs about the distributions, we study two factors that may affect how beliefs are updated.

We first verify whether or not our participants diverge from the reference Bayesian posterior mean when giving their own posterior beliefs (and by how much). In our setting, we were unable to elicit the full distribution of second-order priors as eliciting a probability for each possible combination for the numbers of workers who got each possible score is unfeasible to do in an experimental setting.

Figure 4: Deviation from the Bayesian Benchmark by Gender



In order to create a reference Bayesian posterior and to make this analysis possible, we assume that participants hold a set of prior beliefs of the possible compositions of the pots in accordance with a Dirichlet-Multinomial distribution. As the samples that participants draw have a discrete support, as expected under a multivariate hypergeometric distribution, we are able to update beliefs through the manipulation of the hyperparameters of the Dirichlet-Multinomial distribution[9]. This allows us to generate our Bayesian benchmark in this situation where we cannot see the entire distribution of priors[10]. After the theoretical distribution of priors have been calculated, we take the mean prior of the Dirichlet-Multinomial distribution to be the benchmark posterior distribution against which the participants' posteriors are compared. This benchmark allows us to calculate a new theoretical mean and compare it to the empirically observed mean of the participant's distribution. We do this

[9]The Dirichlet-Multinomial distribution is a family of probability distributions on finite, integer support. As such, it is well suited to our belief elicitation task, in which we elicit a probability distribution over five events from employers. Importantly, if the prior distribution of beliefs is Dirichlet-Multinomial and the likelihood function follows multivariate hypergeometric distribution, the posterior distribution of beliefs is also Dirichlet-Multinomial, which greatly facilitates our computations. This is due to the conjugacy of these prior and posterior distributions.

[10]An important thing to note when calculating with the Dirichlet-Multinomial is that it is not able to handle a calculation where an interval is observed to be 0. In order to address this, we add a very small inconsequential observation to the interval of 0.0001. This allows us to compute a posterior when there are observations within this interval. Then, in order to tackle the potential that this means the posterior does not equal the prior when there are no observations for that population as a whole, the posterior is then reset to be the same as the prior, which would have been predicted in a situation with no observations.

by constructing a ratio for how the empirically observed mean diverges from the theoretical mean with the fraction $\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}}$.[11]

In the full pre-registered analysis in Table 7 in Appendix 6.3.2, we regress this divergence ratio on whether people are updating their beliefs about the male pot or female pot with controls for treatments. For clarity, we present the margins of the results of a non-preregistered version of this analysis using the treatments "Beliefs Only - Framing" and "Beliefs Only - Shortlisting Samples" to show this behaviour in our framed treatments. This is displayed graphically in figure 4.

Figure 4 presents the margins of how this divergence ratio differs when updating for each distribution. The left bar represents the divergence when updating about women and the right bar represents the divergence when updating about men. Looking at these figures, one can clearly see that people generally over update compared to the Bayesian benchmark, but this does not seem to be different by gender. In the full analysis in 6 in appendix 6.3.2 for both the pre-registered version and previously presented non-pre-registered version , the coefficient on the Female Employer dummy is positive, which would suggest people overreact more in response to information about women, but this is not statistically significant.[12] As such we cannot reject the null that our participants updated their beliefs in the same direction as predicted by the Bayesian benchmark.

Result 2: *There is no difference in how posterior beliefs deviate from the Bayesian benchmark posterior when updating beliefs about men and women.*

Robustness: As an important robustness check, we check how the framing of the pots impacted belief updating in the beliefs only treatments. This check was provided by the inclusion of the "Belief Only - Not Framed" treatment. In the appendix 6.3.2 in table 7, we provide a regression model with the divergence ratio regressed on whether the beliefs being updated are about women/pot-B and interactions with the "Belief Only - Not Framed" and "Beliefs Only - Shortlisting Samples"

[11] In order to deal with the issues surrounding a 0 theoretical mean change in the ratio $\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}}$, for example in cases where they do not observe anyone from that group, we add 0.02 to the theoretical mean if for any observation the theoretical mean change is 0 and the difference between the empirical posterior and prior is positive; we subtract 0.02 from the theoretical mean if for any observation the theoretical mean change is 0 and the difference between the empirical posterior and prior is negative; and we code the ratio as 1 if the participant does not change beliefs and the theoretical mean change is 0. By doing so, we remove the possibility that this ratio is undefined, as well as giving participants the benefit of the doubt for that observation that they would have updated in the correct direction if the theoretical mean change had been anything other than 0. However, this does allow us to encode a "punishment" metric if they shift beliefs when they should not have. Additionally, to control for the highly influential effect of small theoretical means on this ratio, we choose to set the theoretical mean change to be 0.02 for any observation where the theoretical mean change is between 0 and 0.02; and we choose to set the theoretical mean change to be -0.02 for any observation where the theoretical mean change is between -0.02 and 0.

[12] The noisy estimates could be due to a number of reasons. Some participants chose not to update their beliefs when the model predicted small changes in the distribution of beliefs. In the beliefs-only treatments, a theoretical change in mean smaller than 0.02 occurred 313 times, of which a participant did not update beliefs 34 times, and a theoretical change in mean smaller than 0.1 occurred 583 times, of which a participant did not update beliefs 123 times. Some participants updated their beliefs when that was not warranted i.e. participants updated beliefs about a group when there were no observations about that group. This happened 43 times for beliefs about women and 54 times for beliefs about men. Instances of incorrect belief updating (i.e., in the wrong direction) were rare. In the beliefs only treatments, this happened 584 times (20.53%).

treatments. Using the estimates in model (2) in Table 7, we run a Wald test checking whether the addition of the coefficients of "1(Female Worker)" and "1(Female Worker) X OB-NF" is different from 0 (F(1,305)=0.13, p=0.7212, FDR Sharpened q-value = 0.727). Under both the p-value and the FDR sharpened q-value, we fail to reject the null of a difference, suggesting that belief updating is not different for the pots when framed neutrally.

Additionally, we run two other checks on this result in appendix 6.5. The first is whether this result is robust to only including people shifting their beliefs in the correct direction.[13] The second is to look at divergence from the Bayesian benchmark as the mean squared error between the theoretical posterior distribution and the empirically observed posterior distribution. We the same result maintains with these analysis.

Exploratory:[14] To provide a further check on whether people are updating beliefs differently by gender in a more raw fashion[15], we run an exploratory analysis in which we classify belief changes by whether they update in the correct direction given their observations, don't change beliefs or update in the wrong direction.[16] We find that people are not differentially likely to update in the correct direction or incorrect direction by the gender of the distribution about which they are updating their beliefs. Additionally, we explore whether people update differently in response to seeing a "high-skilled worker" or "low-skilled worker" and whether this differs by gender.[17] We do not find strong evidence of this, with some very weak evidence that participants over-update less when they see a low-skilled women.

3.3 Mechanism 2: Sampling Behaviour

3.3.1 Does more sampling from a distribution impact accuracy of beliefs?

We consider next how the accuracy of beliefs evolves with respect to sampling from each distribution. We initially consider the evolution of accuracy of beliefs in the beliefs-only treatments, since sampling is strictly exogenous. We then study the impact of the shortlisting restriction on the accuracy of beliefs in the shortlisting treatments. The measure of belief accuracy we use is the absolute distance of the mean of participants' belief distributions from the mean of the true distributions of beliefs[18]. A smaller absolute distance implies more accurate beliefs.

Figure 5 presents the evolution of the accuracy of beliefs about the performances of men (Pot A) and women (Pot B) in the framed and non-framed beliefs only treatments. The horizontal axis

[13] That is, people shifting the mean of their belief upwards if the theoretical posterior has a higher mean than their prior belief, and vice versa.

[14] Exploratory analysis can be found in appendix 6.8

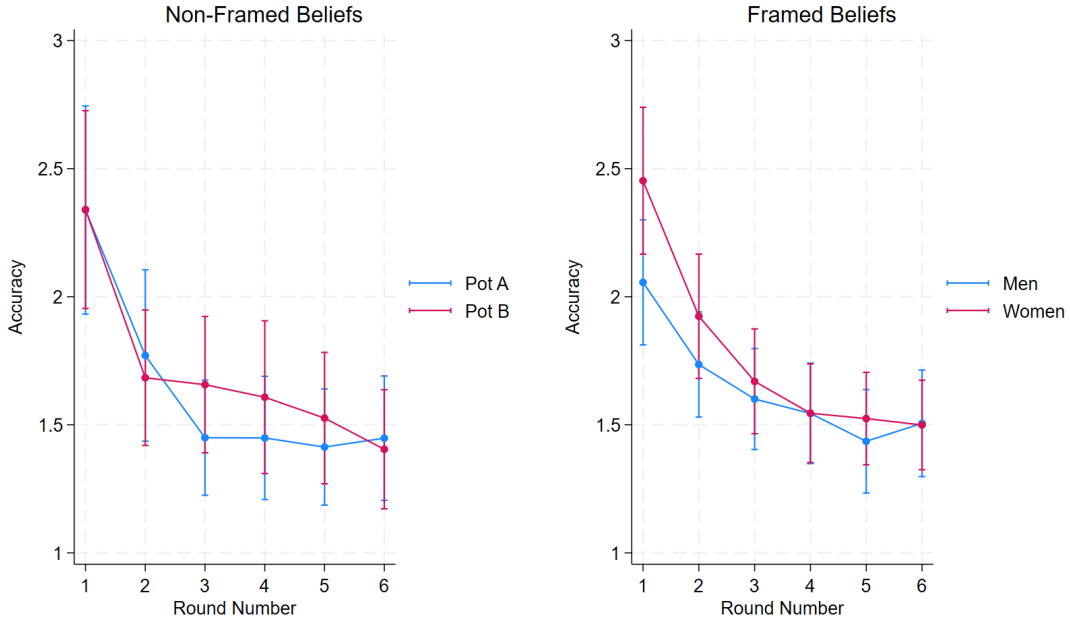
[15] As opposed to exclusively using the processed belief updating with the Dirichlet-Multinomial.

[16] If the mean of the observations the participants receive is higher than their subjective mean, they should move their beliefs about the mean upwards.

[17] Here, we define a high-skilled worker as a worker whose score was between 25-30 and a low-skilled worker as a worker whose score was between 5-10.

[18] Note that as we collect beliefs as an interval, we use the interval-midpoints for our mean calculation.

Figure 5: Evolution of Accuracy of Beliefs by Round



denotes the round number. The vertical axis denotes our measure of accuracy of beliefs about the distribution. The downward trend suggests that our participants gain more accurate beliefs about both distributions in all treatments as participants progress through the rounds. We formally test the impact of having sampled from a population on accuracy of beliefs with a Pooled OLS model. The full results for this are displayed in Appendix 6.3.3 in table 8. We display a shortened non-pre-registered version of this in table 2 using the data from "Only Beliefs - Framing" and "Only Beliefs - Shortlisting Samples," again to show what happens in the framed treatments. In model (1), we present the regression of the accuracy of beliefs on the number of observations seen by participants, with controls. In model (2), we present the accuracy regressed on the number of observations seen, a second half dummy for rounds 4-6 and an interaction between the two, with controls.

The coefficient on cumulative observations is significant in both versions of the presented regression models. The negative coefficient for cumulative observations suggests that participants are shifting their beliefs in the correct direction and as such, they are gaining in accuracy of beliefs. This supports the hypothesis that greater sampling from a distribution improves the accuracy of beliefs. Interestingly, participants appear to be more accurate in the second half compared to the first, even when we include the number of observations. The coefficient on "Second Half" is significant in model (2). This is interesting as there may in fact be a time effect in learning in addition to an effect from the number of observations a participant has had for a group.

Table 2: Pooled OLS Regression: Determinants of Accuracy in Beliefs Only-Framed and Beliefs Only - Shortlisting Samples

	(1)	(2)
Cumulative Observations	-0.026***	-0.057***
	(0.005)	(0.012)
1(Second Half)=1		-0.526**
		(0.206)
1(Second Half)=1 X Cumulative Observations		0.050***
		(0.012)
Constant	1.250***	1.421***
	(0.407)	(0.402)
Controls	Yes	Yes
Observations	2400	2400
Clusters	200	200
R2	0.063	0.070

Cluster robust standard errors in parentheses.

Where appropriate, FDR sharpened q-values are reported in {}. * p<0.1 ** p<0.05, *** p<0.01.

Exploratory:[19] This previous statement of it being easier to get more accurate beliefs is in line with something we observe in the exploratory analysis. Again, classifying belief changes by whether they update in the correct direction given their observations, don't change beliefs or update in the wrong direction, we find that people are more likely to update in the correct direction if the theoretical shift in mean as given by our Bayesian benchmark is large. When participants beliefs are highly inaccurate, it is more likely that the size of the theoretical shift is larger, and as participants beliefs get more accurate, the size of the theoretical shift is smaller. This would suggest that the rate at which a participant gets more accurate beliefs will be diminishing as this size of theoretical shift decreases.

Additionally, there is some suggestion that later in the experiment, more observations reduces the rate at which a participant gains accuracy. Given that in Figure 5 that the average accuracy about either pots in either treatments end up being fairly similar, there may be some level of accuracy that can be easily reached i.e. participants struggle to get a greater accuracy level than what they get too.

Result 3: *Greater sampling from a distribution improves the accuracy of beliefs.*

3.3.2 How do the moments of subjective beliefs impact sampling?

We move on to examining shortlisting decisions themselves, and how they depend on beliefs.

Specifically, we examine whether participants' shortlisting decisions were influenced by their beliefs about the ability of men and women, both in terms of the mean and variance of their belief distribution. We regress the number of women shortlisted on the difference in beliefs about the means of the male and female pot and on the difference in beliefs about the variance of the male and female pot, with a Pooled Panel Poisson model. The full pre-registered results of this analysis are available in Appendix 6.3.4. We display a shortened non-pre-registered version here using the data from the "Baseline" and "Shortlisting Restrictions" treatments in table 3, without the "Probabilistic Exogeneity" treatment as this is primarily a robustness treatment.

From the above regression, we see that both the perceived difference in mean between the male and female distributions and the perceived difference in variance between the male and female distributions impact behaviour. From the model in (3) and its margins in (4), we see that an increase of the difference in perceived mean, calculated with the male mean subtracted from the female mean, of 1 solved puzzle increases the number of women shortlisted by 0.222. This is significant at the 1% level. Additionally, we see that an increase in the difference in perceived variance, calculated with the male variance subtracted from the female variance, decreases the number of women shortlisted by 0.019. This is also significant at the 1% level.

Result 4: *Participants are more likely to shortlist from groups with a higher mean belief and lower*

[19] Exploratory analysis can be found in appendix 6.8

Table 3: Pooled Poisson Regression: Determinants of the Number of Woman Shortlisted in Baseline and Shortlisting Restrictions

	(1)	(2) Margins
Mean Difference (W-M)	0.048***	0.210***
	(0.012)	(0.038)
1(Shortlisting Restrictions (SR))=1	0.025	0.143
	(0.032)	(0.159)
SR X Mean Difference	-0.012	
	(0.015)	
Variance Difference (W-M)	-0.008***	-0.020**
	(0.003)	(0.009)
SR X Var Difference	0.007**	
	(0.004)	
Constant	1.637***	
	(0.072)	
Controls	Yes	Yes
Treatment Interactions	Yes	Yes
Observations	1170	1170
Clusters	195	195
Pseudo-R2	0.016	.

Cluster robust standard errors in parentheses.
* p<0.1 ** p<0.05, *** p<0.01.

Table 4: Pooled OLS Regression: Do Shortlisting Restrictions Improve Accuracy in the Baseline and Shortlisting Restrictions Treatments?

	(1)	(2)
1(Shortlisting Restrictions (SR))=1	0.184	0.219
	(0.169)	(0.185)
1(Second Half)=1		-0.148
		(0.103)
SR X Second Half		-0.070
		(0.152)
Constant	1.855***	1.930***
	(0.343)	(0.344)
Controls	Yes	Yes
Observations	2340	2340
Clusters	195	195
R2	0.013	0.017

Cluster robust standard errors in parentheses.
* p<0.1 ** p<0.05, *** p<0.01.

variance in beliefs.

3.4 Do quotas on shortlisting lead to more accurate beliefs?

We next study the accuracy of beliefs in the shortlisting treatments and whether a restriction on the shortlisting leads to an increase in the accuracy of beliefs. In order to study this, we regress the accuracy of the mean belief on a dummy for the "Shortlisting Restrictions" treatment and an interaction with a dummy for rounds 4 to 6, with a Pooled OLS model. We present the full pre-registered analysis of this in table 10 in appendix 6.3.5. Here, we present a shortened version of this analysis using the "Baseline" and "Shortlisting Restrictions" treatments in table 4, without the "Probabilistic Exogeneity" treatment as this is primarily a robustness treatment.

The coefficient for the impact of neither the shortlisting restrictions nor its interaction with the second half dummy is significant. As such, we cannot reject the nulls that this form of restriction on shortlisting behaviour increases the accuracy of beliefs or that there is a different impact of the shortlisting restriction over time.

Result 5: *Shortlisting restrictions do not lead to more accurate beliefs and there is no differential impact of the shortlisting restriction across time.*

One potential reason as to why we do not observe an effect here is that the "Shortlisting Restrictions" treatment did not have a particularly strong impact on shortlisting behaviour, as observed in table 3. The extent of any potential effect is that it may have dampened the impact of the difference in variance between the true groups on shortlisting behaviour. In hindsight, it is likely that our shortlisting restriction was not strong enough to elicit a response. The mean number of women shortlisted in a round is 4.984 with a standard deviation of 1.901. If we take two standard deviations either side of the mean, the bound for the majority on the number of women shortlisted is approximately [1.18, 8.78]. Given this, the number of individuals for whom this policy could have had an effect were relatively small. On reflection, our intervention of requiring 1 from each group was unlikely to impact behaviour.

4 Discussion

4.1 Initial Priors

The initial question we ask is whether there are differences in the initial priors of the distributions in our performance domain. The belief elicitation we use allows us to capture the entire probability distribution of mean prior beliefs (i.e. how many people do you think got a score of X, how many got a score of Y). As we capture the full distribution, we are able to look at different moments of the distribution. We find a difference in mean prior means but not mean prior variances. This finding about abilities in a chess domain further contributes to potential stereotypes in the difference in abilities between men and women, such as with maths and sports (Carlana, 2019; Bordalo et al., 2019; Gupta, 2023).

There are a number of reasons why this initial difference in stereotypes exist. It could be that a participant has had a lot of potential exposure to a high-performing men through the media and less so to women. This would be line with a stereotype model of representativeness (Bordalo et al., 2016). In this case, the probability of seeing a highly rated male player is far higher than seeing a highly performing female player. Even though this may in part be a function of sample size, it is unlikely that people take this into account and as such may exhibit this type of representativeness bias (Kahneman and Tversky, 1972).

It could also be that public statements made by "experts" such as Nigel Short guides public opinion, much in line with the literature on how people take advice from experts. Additionally, while participants may not have had as much exposure to chess, participants may have different beliefs about the determinants of chess puzzle solving ability. Chess ability will partially be a function of spatial working memory and differences about this could impact beliefs about puzzle solving.

4.2 Mechanism 1: Differences in Belief Updating

When compared to the Bayesian benchmark, we see that there is no difference in updating in terms of deviations from the mean of distribution of the theoretical mean posterior. In gender-stereotyped domains, this is a surprising result. When the literature has considered belief updating in stereotyped domains, they often find that participants are more likely to update in directions that allow them to hold these stereotyped beliefs (Bordalo et al., 2019; Gupta, 2023). These are usually considered to be indicative of motivated reasoning (e.g. Eil and Rao, 2011; Möbius et al., 2022; Thaler, 2020).

However, in this context, it could be that participants do not have strong desires to hold particular beliefs. When compared to a domain such as maths ability, chess ability might not be as closely tied to perceived norms. This would suggest that participants do not gain as much utility from believing men are better in this chess domain (Akerlof and Kranton, 2000). This could be one mechanism as to why we do not see a difference in belief updating between two groups.

4.3 Mechanism 2: Sampling Behaviour

This second mechanism of the sampling behaviour driving inaccurate beliefs breaks up into two components - 1) how sampling and search impacts beliefs and 2) how beliefs impact sampling and search

We consider the first component by asking the question "do people gain more accurate beliefs with greater contact with a group?" We find evidence in support of this, through participants gaining more accurate beliefs the more they sample from a distribution. Interestingly, we also find evidence that greater time spent engaging in the activity. When thinking about inaccurate gender beliefs in a STEM domain as in (Feld et al., 2022), it could be that employers require both greater contact with both men and women to have more accurate beliefs, but also simply more time for these beliefs to adjust.

The other half of this mechanism is whether participants searched for candidates in accordance with their beliefs. Because we have the entire distribution of the mean prior beliefs, we observe whether perceptions of the mean and variance impacts sampling behaviour. In this case, both mean and variance are important as we can think about searching through a population for participants as a lottery (Aigner and Cain, 1977). In our context, both the subjective mean and subjective variance impacts this sampling behaviour. The result for the mean is inline with previous stylized hiring experiments that consider the mean of the distribution (Bohren et al., 2020; Coffman et al., 2021). Additionally, we contribute experimental evidence of how the subjective variance impacts hiring behaviour.

4.4 Affirmative Action

This sampling-based mechanism suggests that some form of affirmative action may encourage interaction with a group. Traditional affirmative action policies look at having a hard quota on hiring

(Coate and Loury, 1993; Gupta, 2023). However, the mechanic through which we manipulate is a search mechanic, as such we choose to evaluate a softer form of affirmative action in which we implement a quota at the shortlisting stage (Fershtman and Pavan, 2021; Komiyama and Noda, 2020). Results on the effectiveness of shortlisting have been mixed, with some suggestions it does not change the outcome for the intended target group (DuBois, 2015; Fanning Madden and Ruther, 2011; Solow et al., 2011). Additionally, there is scope it could in fact discourage deeper learning about minority applicants in the candidate pool (Fershtman and Pavan, 2021).

We identify another route through which affirmative action may not work, and that is the shortlisting quota itself not being strong enough to elicit a change in behaviour. We required our participants to sample at least one candidate from each group. However, the majority of hiring behaviour we saw was in the range of approximately [1.18, 8,78]. A reason that could be driving this hiring behaviour is that the difference in beliefs is not yet large enough to have caused one group to not be sampled from (Lepage, 2023).

5 Conclusion

The literature around statistical discrimination has broadened recently to have a greater focus on the idea of inaccurate statistical discrimination (Bohren et al., 2019, 2020; Feld et al., 2022). This by itself leads to a problem of some groups being underrepresented in the labour market simply because employers and firms have not accurately assessed the abilities of the group as a whole. However, in addition to this, the problem is further perpetuated by the dynamic nature of interactions (Lepage, 2023; Komiyama and Noda, 2020; Gupta, 2023). If you perceive one group to be better than another group, you'll never hire from the perceived worse group. This means you lose the opportunity to learn about that group. Due to these lack of learning opportunities, one can easily end up in a situation where you cannot fix these inaccuracies.

We run a stylized hiring experiment with a gender frame that tries to address the overarching question of why we continue to have different beliefs about similarly performing groups. Stylized hiring tasks require some performance task against which hiring is undertaken. In our case, we choose to use chess puzzles based on the performance data of Chess.com's puzzle rush task. Throughout the experiment, participants are asked to construct a shortlist of 10 potential candidates from two groups, sampling however they wish, and then hire a candidate from this list. Additionally, we continually collect participants beliefs about the full distribution of abilities of the potential workers in these two groups. While we choose to frame this through the lens of the labour market, it would be quite easy to extend this framing to a domain that is reminiscent of training young chess players. A chess coach has a limited amount of time and energy to go and find young chess players, and as such, if they have some prior that young men go on to be better at chess than young women, they may spend more time searching and training young men.

In this type of study, checking that there is an initial difference in priors is an important first

step. With this in mind, We do find an initial gender-based difference in beliefs about the mean but not in beliefs about the variance. This finding about abilities in a chess domain further contributes to potential stereotypes in the difference in abilities between men and women, such as with maths and sports (Carlana, 2019; Bordalo et al., 2019; Gupta, 2023).

A key theme of this paper is understanding why inaccurate beliefs about similar groups may persist over time. To this end, we consider two mechanisms - one based in differences in beliefs updating and one based in sampling behaviour.

Considering the first mechanism, when compared to the Bayesian benchmark, we see that there is no difference in updating in terms of deviations from the mean of distribution of the theoretical mean posterior. In the context of the literature, this is a surprising result as participants generally update to hold stereotyped beliefs (Bordalo et al., 2019; Gupta, 2023). Future research may wish to study the impact of the strength of the norm of a gender stereotype on gender differences in belief updating.

Moving onto the second mechanism, we first consider how sampling impacts the beliefs about a distribution. We have evidence that both greater sampling from a group as well as greater time spent in sampling behaviour may generate more accurate beliefs. This contributes an important result to the literature on inaccurate belief discrimination. An important facet of inaccurate belief discrimination is consideration for how to correct inaccurate beliefs. The literature on the dynamics of belief discrimination has generally considered how these initial beliefs might reinforce (Lepage, 2023; Komiyama and Noda, 2020). We contribute a route through which an employer might gain more accurate beliefs about populations.

To provide the second side of this mechanism, we consider how beliefs impact shortlisting behaviour. We find that both the subjective mean and subjective variance impact the shortlisting behaviour in the stylized hiring game. The result for the mean is inline with previous stylized hiring experiments that consider the mean of the distribution (Bohren et al., 2020; Coffman et al., 2021). Additionally, we contribute new experimental evidence of how the subjective variance impacts hiring behaviour.

In line with encouraging greater contact with a group, we find that lighter forms of affirmative action may not elicit a change in behaviour. This supplements in the literature on the Rooney Rule and soft affirmative action by providing some suggestive evidence to the mechanism as to why the Rooney Rule does not seem to have bridged the gap in the NFL (Fershtman and Pavan, 2021).

A key policy implication of our study is that encouraging long-term contact with a group in order to remove persistent inaccurate belief differences between groups. However, when implementing affirmative action, a policy maker needs to give consideration to the strength of the affirmative action and ensure that it is strong enough to elicit the desired change.

References

Aigner, D. J. and Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *Ilr Review*, 30(2):175–187.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Arrow, K. J. (1977). *The theory of discrimination.*, page 3—33. Princeton NJ: Princeton University Press.

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

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Bertrand, M. and Duflo, E. (2017). Field experiments on discrimination. *Handbook of economic field experiments*, 1:309–393.

Blanch, A. (2016). Expert performance of men and women: A cross-cultural study in the chess domain. *Personality and Individual differences*, 101:90–97.

Blanch, A., Aluja, A., and Cornadó, M.-P. (2015). Sex differences in chess performance: Analyzing participation rates, age, and practice in chess tournaments. *Personality and Individual Differences*, 86:117–121.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2020). Inaccurate statistical discrimination: An identification problem.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019). The dynamics of discrimination: theory and evidence. *American Economic Review*, 109(10):3395–3436.

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

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–773.

Borghans, L., Heckman, J. J., Golsteyn, B. H., and Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, 7(2-3):649–658.

Buchanan, I., Pratt, A., and Francis-Devine, B. (2022). Women and the uk economy. Technical report, House of Commons Library.

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

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Cheryan, S., Ziegler, S. A., Montoya, A. K., and Jiang, L. (2017). Why are some stem fields more gender balanced than others? *Psychological Bulletin*, 143(1):1.

Coate, S. and Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, pages 1220–1240.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660.

Coffman, K. B., Exley, C. L., and Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science*, 67(6):3551–3569.

Crosetto, P. and Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, 47:31–65.

Crosetto, P. and Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, 19:613–641.

DuBois, C. (2015). The impact of "soft" affirmative action policies on minority hiring in executive leadership: The case of the nfl's rooney rule. *American Law and Economics Review*, 18(1):208–233.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138.

Ellis-Petersen, H. (2015). Nigel short says men "hardwired" to be better chess players than women. *The Guardian*.

Fanning Madden, J. and Ruther, M. (2011). Has the nfl's rooney rule efforts "leveled the field"for african american head coach candidates? *Journal of Sports Economics*, 12(2):127–142.

Feld, J., Ip, E., Leibbrandt, A., and Vecci, J. (2022). Identifying and overcoming gender barriers in tech: A field experiment on inaccurate statistical discrimination.

Fershtman, D. and Pavan, A. (2021). "soft" affirmative action and minority recruitment. *American Economic Review: Insights*, 3(1):1–18.

Gerdes, C. and Gränsmark, P. (2010). Strategic behavior across gender: A comparison of female and male expert chess players. *Labour Economics*, 17(5):766–775.

Gupta, N. (2023). Can temporary affirmative action improve representation? *Job Market Paper*.

Harness, A., Jacot, L., Scherf, S., White, A., and Warnick, J. E. (2008). Sex differences in working memory. *Psychological Reports*, 103(1):214–218.

Harrison, G. W., Martínez-Correa, J., Swarthout, J. T., and Ulm, E. R. (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization*, 134:430–448.

HESA (2022). Higher education staff statistics: Uk, 2020/21. Technical report, HESA.

Ingle, S. (2021). It is not biology': Women's chess hindered by low numbers and sexism. *The Guardian*.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454.

Komiyama, J. and Noda, S. (2020). On statistical discrimination as a failure of social learning: A multi-armed bandit approach. *arXiv preprint arXiv:2010.01079*.

Lepage, L. P. (2023). Experience-based discrimination. *Working Paper*.

Maass, A., D'ettole, C., and Cadinu, M. (2008). Checkmate? the role of gender stereotypes in the ultimate intellectual sport. *European Journal of Social Psychology*, 38(2):231–245.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*, 68(11):7793–7817.

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

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

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

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

Riach, P. A. and Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, 112(483):F480–F518.

Robert, M. and Savoie, N. (2006). Are there gender differences in verbal and visuospatial working-memory resources? *European Journal of Cognitive Psychology*, 18(03):378–397.

Rothgerber, H. and Wolsiefer, K. (2014). A naturalistic study of stereotype threat in young female chess players. *Group processes & Intergroup relations*, 17(1):79–90.

Smerdon, D. (2022). Facts and myths about gender in chess.

Solow, B. L., Solow, J. L., and Walker, T. B. (2011). Moving on up: The rooney rule and minority hiring in the nfl. *Labour Economics*, 18(3):332–337.

Thaler, M. (2020). The fake news effect: Experimentally identifying motivated reasoning using trust in news. *arXiv preprint arXiv:2012.01663*.

Uberoi, E. and Mansfield, Z. (2022). Women in politics and public life. Technical report, House of Commons Library.

Weber, E. U., Blais, A.-R., and Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4):263–290.

6 Appendix

6.1 Appendix: Q-Value Table

Table 5: Q Values on Pre-Registered Hypotheses

Family	p-value	q-value (in family)	q-value (Whole Set)
Priors and Updating:			
Men are initially perceived to be better on average than women.	2.24e-08	.001	.001
It is possible that women are perceived to have higher variance in ability, or that the beliefs about women are noisier and result in higher variance (Aigner and Cain, 1977). We test that there is no difference.	0.645	1	.852
Belief updating is different for women compared to men.	0.814	1	.904
Belief updating is not different for the pots when framed neutrally.	0.721	1	.852
Sampling:			
A higher perceived mean for women leads to greater sampling from the women pot	0.00011	.001	.001
A higher perceived variance for women leads to less sampling from the women pot	0.0029	.003	.006
Shortlisting restrictions increases sampling of women	0.457	.180	.666
Accuracy:			
More sampling from a distribution increases accuracy of beliefs:	1.437e-08	.001	.001
The impact of greater sampling is different in the later rounds of the experiment	1.437e-07	.001	.001
Shortlisting restrictions increase the accuracy of beliefs	0.359	.315	.560
The impact of shortlisting restrictions on the accuracy of beliefs is different in the later rounds of the experiment	0.973	.918	1.00
Non-incentivised vs Incentivised Beliefs:			
Collecting unincentivized beliefs does not impact belief updating.	0.677	0.513	.852
Collecting unincentivized beliefs does not impact the accuracy of beliefs.	0.138	0.382	.226

6.2 Appendix: Does the incentivisation impact the evolution of beliefs?

In order to understand whether incentivising the belief elicitation changed behaviour in how participants gave beliefs, we ran the Only Beliefs – Shortlisting Samples (OB-SS) treatment to be able to compare the belief-only treatments with the Baseline treatment. In order to study this, we estimate the following specifications with a Pooled OLS model, using data from BASE and BO-SS:

$$\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}} = \alpha + \beta_1 SS_i + \gamma X_i + \varepsilon_{i,t} \tag{2}$$

$$A_{i,g,t} = \alpha + \beta_1 SS_i + \gamma X_i + \varepsilon_{i,t} \tag{3}$$

Where $\Delta E_{i,g,t}$ is the difference between the empirical posterior mean and stated prior; $\Delta T_{i,g,t}$ is the difference between the theoretical posterior mean and stated prior; $A_{i,g,t}$ is the absolute distance between the mean of the participants' belief distribution for group g in round t and the mean of the true distribution of group g; SS_i is a dummy, taking a value of 1 if the given beliefs are elicited the Only Beliefs – Shortlisting Samples treatment; X_i represents the vector of individual characteristics used as controls; *alpha* is a constant; and $\varepsilon_{i,t}$ is the error term. The first specification is reported in model (1) and in table **??** and the second specification is reported in model (2) and in table **??**.

Given the lack of significance on the SS_i dummy, we aren't able to reject the null that there is a difference in beliefs behaviour between the baseline treatment and the Only Beliefs – Shortlisting Samples treatment.

6.3 Appendix: Full Analysis

6.3.1 Are priors about men and women in chess different?

We estimated the following empirical specifications, using a pooled OLS regression, with standard errors clustered at the individual level.

$$\mu_{i,g} = \alpha + \beta_1 W + \gamma X_i + \varepsilon_i \tag{4}$$

$$\sigma_{i,g} = \alpha + \beta_1 W + \gamma X_i + \varepsilon_i \tag{5}$$

Where $\mu_{i,g}$ is participant i's perceived prior mean of distribution g; $\sigma^2_{i,g}$ is participant i's perceived prior variance of distribution g. Both $\mu_{i,g}$ and $\sigma^2_{i,g}$ were constructed at the individual level using the stated probabilities for each category and the mid-point of each category (e.g., we used the value 7.5 for the 5-10 category); W takes a value of 1 if the stated beliefs are about the performance distribution of women; X_i represents the vector of individual employer characteristics; α is a constant; and $\varepsilon_{i,t}$ is the error term. The vector of individual employer characteristics includes whether they are female; whether they have watched the Queen's Gambit ; whether they know the rules of chess; whether they play chess regularly; whether they themselves have played the chess performance task; whether

they have ever undertaken a hiring role at work; the age of participants in years.

The results of these estimations are displayed in Table 6 with controls in models 1 and 2.

Table 6: Pooled OLS Regression: Determinants of Mean and Variance of Prior Beliefs

	(1) Mean	(2) Variance
1(Female Worker)	-0.469***	-0.201
	(0.083)	(0.436)
	{0.001}	{1.000}
1(Female Employer)	-0.013	-0.454
	(0.194)	(1.026)
1(Queen's Gambit)	0.107	-1.048
	(0.187)	(1.031)
1(Chess Rules?)	0.023	-0.948
	(0.237)	(1.265)
1(Play Chess)	-0.125	0.065
	(0.128)	(0.692)
1(Puzzle Rush)	-0.885*	0.105
	(0.482)	(2.097)
1(Hiring Role)	0.240	1.721
	(0.217)	(1.118)
Age	-0.007	0.024
	(0.009)	(0.046)
1(Student)	0.023	0.276
	(0.241)	(1.334)
1(Full-Time Work)	-0.179	-0.114
	(0.200)	(1.042)
Constant	17.073***	31.620***
	(0.352)	(2.077)
Observations	988	988
Clusters	494	494
R2	0.027	0.010
BASE	Yes	Yes
SR	Yes	Yes
PE	Yes	Yes
BO_F	Yes	Yes
BO_SS	Yes	Yes
BO_NF	No	No

Cluster robust standard errors in parentheses.

Where appropriate, FDR sharpened q-values are reported in {}. * p<0.1 ** p<0.05, *** p<0.01.

6.3.2 Is there a gender-based difference in belief updating?

To do so we estimate the following specification, using a Pooled OLS model, using data from BO-F, BO-SS, BO-NF:

$$\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}} = \alpha + \beta_1 W + \beta_2 NF_i + \beta_3 W \times NF + \beta_4 SS_i + \beta_5 W_i \times SS_i + \gamma X_i + \varepsilon_{i,t} \quad (6)$$

Where $\Delta E_{i,g,t}$ is the difference between the empirical posterior mean and stated prior; $\Delta T_{i,g,t}$ is the difference between the theoretical posterior mean and stated prior; W is a dummy, taking a value of 1 if the given beliefs are about the distribution of women; NF_i is a dummy, taking a value of 1 if the given beliefs are elicited in the Only Beliefs - No-Framing treatment; SS_i is a dummy, taking a value of 1 if the given beliefs are elicited the Only Beliefs – Shortlisting Samples treatment; X_i represents the vector of individual characteristics used as controls; α is a constant; and $\varepsilon_{i,t}$ is the error term.

The results of these estimations are displayed in Table 7 with controls model 2. The analysis presented in the main body of the paper is in model 1 using BO-F and BO-SS. In model 3, we present the same analysis using the shortlisting treatments.

Table 7: Pooled OLS Regression: Determinants of Divergence of Belief Updating from a Bayesian Benchmark

	(1)	(2)	(3)
1(Female Worker)=1	0.425	0.425	1.585
	(1.804)	(1.801)	(1.356)
		{1.000}	
1(Shortlisting Samples (OB-SS))=1	-1.785	-1.629	
	(1.499)	(1.487)	
1(Female Worker)=1 X 1(Shortlisting Samples (OB-SS))=1	0.800	0.800	
	(2.185)	(2.181)	
1(Female Employer)	-0.922	-1.490	1.863*
	(1.103)	(1.027)	(0.958)
1(Queen's Gambit)	0.222	-0.872	-1.036
	(1.082)	(1.056)	(0.886)
1(Chess Rules?)	-2.415*	-1.564	-1.285
	(1.362)	(1.129)	(1.200)
1(Play Chess)	-0.740	-0.067	0.316
	(0.651)	(0.847)	(0.611)
1(Puzzle Rush)	1.716	1.421	1.883
	(1.633)	(1.536)	(3.280)
1(Hiring Role)	1.273	1.530	1.881*
	(1.346)	(1.034)	(0.998)
Age	-0.010	-0.003	-0.063**
	(0.087)	(0.056)	(0.031)
1(Student)	1.538	0.658	0.868
	(2.399)	(1.817)	(1.890)
1(Full-Time Work)	-0.869	-1.518	-0.650
	(1.321)	(1.018)	(0.853)
1(No Framing (OB-NF))=1		-0.651	
		(1.834)	
1(Female Worker)=1 X 1(No Framing (OB-NF))=1		0.170	
		(2.454)	
1(Probabilistic Exogeneity (PE))=1			3.890**
			(1.604)
1(Female Worker)=1 X 1(Probabilistic Exogeneity (PE))=1			-4.379*
			(2.390)
1(Shortlisting Restrictions (SR))=1			-0.517
			(0.897)
1(Female Worker)=1 X 1(Shortlisting Restrictions (SR))=1			-0.451
			(1.644)
Constant	6.429*	5.400**	2.219
	(3.549)	(2.442)	(1.754)
Observations	2000	3060	2940
Clusters	200	306	294
R2	0.006	0.004	0.009
BASE	No	No	Yes
SR	No	No	Yes
PE	No	No	Yes
BO_F	Yes	Yes	No
BO_SS	Yes	Yes	No
BO_NF	No	Yes	No

Cluster robust standard errors in parentheses.
Where appropriate, FDR sharpened q-values are reported in {}. * p<0.1 ** p<0.05, *** p<0.01.

6.3.3 Does more sampling from a distribution impact accuracy of beliefs?

We estimated the following specifications with a Pooled OLS model:

$$A_{i,g,t} = \alpha + \beta_1 S_{i,g,t} + \gamma X_i + \varepsilon_{i,t} \tag{7}$$

$$A_{i,g,t} = \alpha + \beta_1 S_{i,g,t} + \beta_2 t + \beta_3 S_{i,g,t} t + \gamma X_i + \varepsilon_{i,t} \tag{8}$$

Where $A_{i,g,t}$ is the absolute distance between the mean of the participants' belief distribution for group g in round t and the mean of the true distribution of group g; $S_{i,g,t}$ is the number of 'workers' from distribution g that individual i has sampled up to round t; t is a dummy, taking a value of 1 if the beliefs were elicited in rounds 4 to 6; X_i represents the vector of individual characteristics used as controls; α is a constant; and $\varepsilon_{i,t}$ is the error term.

The results of these estimations are displayed in Table 8 with controls in models 3 and 4. The analysis presented in the main body of the paper is in model 1 and 2 using BO-F and BO-SS. In models 5 and 6, we present the same analysis using the shortlisting treatments.

Table 8: Pooled OLS Regression: Determinants of Accuracy in Beliefs Only Treatments

	(1)	(2)	(3)	(4)	(5)	(6)
Cumulative Observations	-0.026***	-0.057***	-0.024***	-0.052***	-0.009**	-0.022***
	(0.005)	(0.012)	(0.004)	(0.009)	(0.004)	(0.008)
				{0.001}		
1(Female Employer)	0.113	0.104	0.015	0.003	0.020	0.017
	(0.147)	(0.147)	(0.120)	(0.119)	(0.136)	(0.135)
1(Queen's Gambit)	0.022	0.017	0.051	0.049	-0.175	-0.184
	(0.155)	(0.155)	(0.125)	(0.125)	(0.124)	(0.124)
1(Chess Rules?)	-0.233	-0.225	-0.044	-0.042	-0.162	-0.163
	(0.213)	(0.215)	(0.159)	(0.158)	(0.157)	(0.156)
1(Play Chess)	-0.020	-0.020	-0.079	-0.077	0.057	0.056
	(0.116)	(0.115)	(0.103)	(0.102)	(0.084)	(0.084)
1(Puzzle Rush)	0.485*	0.475*	0.420*	0.412*	-0.034	-0.049
	(0.273)	(0.273)	(0.237)	(0.235)	(0.343)	(0.342)
1(Hiring Role)	-0.285	-0.271	0.003	0.020	0.061	0.062
	(0.195)	(0.197)	(0.146)	(0.146)	(0.149)	(0.148)
Age	0.025**	0.024**	0.016**	0.015**	0.001	0.000
	(0.011)	(0.011)	(0.007)	(0.007)	(0.005)	(0.005)
1(Student)	0.064	0.058	0.139	0.130	-0.011	-0.023
	(0.160)	(0.161)	(0.147)	(0.148)	(0.165)	(0.167)
1(Full-Time Work)	0.319*	0.309*	0.083	0.068	0.113	0.128
	(0.184)	(0.185)	(0.136)	(0.136)	(0.137)	(0.137)
1(Second Half)=1		-0.526**		-0.630***		-0.588***
		(0.206)		(0.144)		(0.176)
1(Second Half)=1 X Cumulative Observations		0.050***		0.051***		0.035***
		(0.012)		(0.009)		(0.010)
				{0.001}		
Constant	1.250***	1.421***	1.475***	1.651***	1.911***	2.042***
	(0.407)	(0.402)	(0.274)	(0.271)	(0.269)	(0.273)
Observations	2400	2400	3672	3672	3528	3528
Clusters	200	200	306	306	294	294
R2	0.063	0.070	0.038	0.047	0.009	0.015
BASE	No	No	No	No	Yes	Yes
SR	No	No	No	No	Yes	Yes
PE	No	No	No	No	Yes	Yes
BO_F	Yes	Yes	Yes	Yes	No	No
BO_SS	Yes	Yes	Yes	Yes	No	No
BO_NF	No	No	Yes	Yes	No	No

Cluster robust standard errors in parentheses.

Where appropriate, FDR sharpened q-values are reported in {}. * p<0.1 ** p<0.05, *** p<0.01.

Result 3: *Greater sampling from a distribution does impact the accuracy of beliefs.*

6.3.4 How do the moments of subjective beliefs impact sampling?

We estimate the following specification, with a Pooled Panel Poisson model:

$$\begin{aligned} S_{i,t} = \alpha + \beta_1 \mu_{i,w-m,t} + \beta_2 \sigma^2_{i,w-m,t} + \beta_3 PE_i + \beta_4 SR_i + \beta_5 PE_i \mu_{i,w-m,t} + \beta_6 SR_i \mu_{i,w-m,t} \\ + \beta_7 PE_i \sigma^2_{i,w-m,t} + \beta_8 SR_i \sigma^2_{i,w-m,t} + \gamma X_i + \varepsilon_{i,t} \end{aligned} \tag{9}$$

Where $S_{i,t}$ is the number of women shortlisted in round t; $\mu_{i,w-m,t}$ is individual i's perceived difference in mean of Pot-Women and Pot-Men at time t; $\sigma^2_{i,w-m,t}$ is individual i's perceived difference in variance of Pot-Women and Pot-Men at time t; PE_i is a dummy, taking a value of 1 if the given beliefs are elicited in the Probabilistic Exogeneity treatment; SR_i is a dummy, taking a value of 1 if the given beliefs are elicited in the Shortlisting Restrictions treatment; X_i represents the vector of individual characteristics used as controls; α is a constant; and $\varepsilon_{i,t}$ is the error term.

The results of this regression are displayed in Table 9 in model 3 with controls and its margins in 4. The analysis presented in the main body of the paper is in models 1 and 2 using BASE and SR.

Table 9: Pooled Poisson Regression: Determinants of the Number of Woman Shortlisted in Shortlisting Treatments

	(1)	(2) Margins	(3)	(4) Margins
Mean Difference (W-M)	0.048***	0.210***	0.048***	0.218***
	(0.012)	(0.038)	(0.012)	(0.031)
			{0.001}	
1(Shortlisting Restrictions (SR))=1	0.025	0.143	0.024	0.142
	(0.032)	(0.159)	(0.032)	(0.162)
			{0.145}	
SR X Mean Difference	-0.012		-0.013	
	(0.015)		(0.015)	
Variance Difference (W-M)	-0.008***	-0.020**	-0.007***	-0.019***
	(0.003)	(0.009)	(0.002)	(0.007)
			{0.005}	
SR X Var Difference	0.007**		0.007*	
	(0.004)		(0.004)	
1(Female Employer)	-0.018	-0.090	-0.014	-0.070
	(0.035)	(0.175)	(0.030)	(0.149)
1(Watched Queen's Gambit)	-0.040	-0.199	-0.049	-0.242
	(0.035)	(0.173)	(0.032)	(0.157)
1(Chess Rules?)	-0.031	-0.154	-0.013	-0.062
	(0.040)	(0.201)	(0.033)	(0.165)
1(Play Chess Regularly)	0.012	0.057	0.001	0.006
	(0.018)	(0.088)	(0.015)	(0.075)
1(Played Puzzle Rush)	-0.136	-0.679	-0.136**	-0.671**
	(0.102)	(0.510)	(0.069)	(0.340)
1(Hiring Role)	-0.048	-0.241	-0.048*	-0.237*
	(0.035)	(0.175)	(0.027)	(0.133)
Age	-0.000	-0.002	-0.002**	-0.011**
	(0.001)	(0.006)	(0.001)	(0.006)
1(Student)	0.010	0.052	-0.028	-0.140
	(0.047)	(0.235)	(0.044)	(0.218)
1(Full-Time Work)	0.062**	0.310**	0.062**	0.308**
	(0.032)	(0.156)	(0.026)	(0.128)
1(Probabilistic Exogeneity (PE))=1			-0.007	-0.030
			(0.033)	(0.163)
PE X Mean Difference			0.001	
			(0.017)	
PE X Var Difference			0.004	
			(0.003)	
Constant	1.637***		1.725***	
	(0.072)		(0.066)	
Observations	1170	1170	1764	1764
Clusters	195	195	294	294
Pseudo-R2	0.016	.	0.017	.
BASE	Yes	Yes	Yes	Yes
SR	Yes	Yes	Yes	Yes
PE	No	No	Yes	Yes
BO_F	No	No	No	No
BO_SS	No	No	No	No
BO_NF	No	No	No	No

Cluster robust standard errors in parentheses.

Where appropriate, FDR sharpened q-values are reported in {}. * p<0.1 ** p<0.05, *** p<0.01.

6.3.5 Do quotas on shortlisting lead to more accurate beliefs?

We estimate the following specification with a Pooled OLS model:

$$A_{i,g,t} = \alpha + \beta_1 SR_i + \gamma X_i + \varepsilon_{i,t} \tag{10}$$

$$A_{i,g,t} = \alpha + \beta_1 SR_i + \beta_2 t + \beta_3 SR_i \gamma X_i + \varepsilon_{i,t} \tag{11}$$

Where $A_{i,g,t}$ is the absolute distance between the mean of the participants' belief distribution for group g in round t and the mean of the true distribution of group g; SR_i is a dummy, taking a value of 1 if the given beliefs are elicited in the No Probabilistic Exogeneity – Shortlisting Restrictions treatment; t is a dummy, taking a value of 1 if the beliefs were elicited in rounds 4 to 6; X_i represents the vector of individual characteristics used as controls.

The results of this regression are displayed in Table 10 in models 3 and 4 with controls. The analysis presented in the main body of the paper is in models 1 and 2 using BASE and SR.

Table 10: Pooled OLS Regression: Do Shortlisting Restrictions Improve Accuracy in Shortlisting Treatments?

	(1)	(2)	(3)	(4)
1(Shortlisting Restrictions (SR))=1	0.184	0.219	0.146	0.144
	(0.169)	(0.185)	(0.143)	(0.157)
				{0.315}
1(Female Employer)	-0.093	-0.093	0.020	0.020
	(0.181)	(0.181)	(0.135)	(0.135)
1(Queen's Gambit)	-0.280*	-0.280*	-0.173	-0.173
	(0.156)	(0.156)	(0.123)	(0.123)
1(Chess Rules?)	-0.023	-0.023	-0.165	-0.165
	(0.216)	(0.217)	(0.157)	(0.157)
1(Play Chess)	0.027	0.027	0.053	0.053
	(0.116)	(0.116)	(0.086)	(0.086)
1(Puzzle Rush)	-0.168	-0.168	-0.047	-0.047
	(0.325)	(0.325)	(0.339)	(0.339)
1(Hiring Role)	-0.047	-0.047	0.074	0.074
	(0.221)	(0.221)	(0.151)	(0.151)
Age	0.002	0.002	0.000	0.000
	(0.007)	(0.007)	(0.005)	(0.005)
1(Student)	-0.126	-0.126	0.001	0.001
	(0.211)	(0.211)	(0.165)	(0.165)
1(Full-Time Work)	0.001	0.001	0.111	0.111
	(0.200)	(0.200)	(0.139)	(0.139)
1(Second Half)=1		-0.148		-0.223***
		(0.103)		(0.071)
1(Shortlisting Restrictions (SR))=1 X 1(Second Half)=1		-0.070		0.004
		(0.152)		(0.133)
				{0.918}
Constant	1.855***	1.930***	1.774***	1.886***
	(0.343)	(0.344)	(0.262)	(0.266)
Observations	2340	2340	3528	3528
Clusters	195	195	294	294
R2	0.013	0.017	0.008	0.013
BASE	Yes	Yes	Yes	Yes
SR	Yes	Yes	Yes	Yes
PE	No	No	Yes	Yes
BO_F	No	No	No	No
BO_SS	No	No	No	No
BO_NF	No	No	No	No

Cluster robust standard errors in parentheses.

Where appropriate, FDR sharpened q-values are reported in {}. * p<0.1 ** p<0.05, *** p<0.01.

6.4 Appendix: Deviations from Pre-Registration

We now report the deviations in analysis from the initial pre-registration. For ease of reading, we re-state the research questions specified in the pre-registration here. These are:

- RQ1: Are priors about men and women different?
- RQ2: Do people update their beliefs differently about men and women?
- RQ3: How does greater sampling from a distribution impact the accuracy of beliefs?
- RQ4: Do shortlisting restrictions at the stage of search lead to more accurate beliefs?
- RQ5: How do different moments, specifically the mean and variance, of the perceived distributions over a population impact sampling behaviour?
- RQ6: Do non-incentivized beliefs systematically differ from incentivized beliefs?

The first thing to note is that throughout all original specifications, we had not included a constant. This was a mistake and has not been included. For RQ2, we originally reported the specification:

$$\frac{1+\Delta E_{i,g,t}}{1+\Delta T_{i,g,t}} = \beta_1 H_{i,g} + \beta_2 L_{i,g} + \beta_3 W + \beta_4 H_{i,g} \times W + \beta_5 L_{i,g} \times W + \gamma X_i + \varepsilon_{i,t} \tag{12}$$

This has been changed to:

$$\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}} = \alpha + \beta_1 H_{i,g} + \beta_2 L_{i,g} + \beta_3 W + \beta_4 H_{i,g} \times W + \beta_5 L_{i,g} \times W + \gamma X_i + \varepsilon_{i,t} \tag{13}$$

With the following changes to the theoretical mean to avoid a denominator with a value of 0:

- we add 0.02 to the theoretical mean if for any observation the theoretical mean change is 0 and the difference between the empirical posterior and prior is positive
- we subtract 0.02 from the theoretical mean if for any observation the theoretical mean change is 0 and the difference between the empirical posterior and prior is negative

To control for the highly influential effect of small theoretical means on this ratio:

- we choose to set the theoretical mean change to be 0.02 for any observation where the theoretical mean change is between 0 and 0.02
- we choose to set the theoretical mean change to be -0.02 for any observation where the theoretical mean change is between -0.02 and 0

The reason for these deviations is due interpretability. By using $\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}}$, it is far clearer to see that the Bayesian benchmark is 1 and the interpretation of deviations from this benchmark is easier to see when a participant over-updates or under-updates compared to the benchmark.

To deal with the issues from a 0 theoretical mean change:

- we add 0.02 to the theoretical mean if for any observation the theoretical mean change is 0 and the difference between the empirical posterior and prior is positive
- we subtract 0.02 from the theoretical mean if for any observation the theoretical mean change is 0 and the difference between the empirical posterior and prior is negative
- we code the ratio as 1 if the participant does not change beliefs and the theoretical mean change is 0

The reason for this addition is to the initial analysis is to adjust for the possibility that this ratio is undefined, as well as giving participants the benefit of the doubt for that observation that they would have updated in the correct direction if the theoretical mean change had been anything other than 0.

Similarly, for RQ6, we originally reported the specification:

$$\frac{1+\Delta E_{i,g,t}}{1+\Delta T_{i,g,t}} = \beta_1 SS_i + \gamma X_i + \varepsilon_{i,t} \tag{14}$$

This has been changed to:

$$\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}} = \alpha + \beta_1 SS_i + \gamma X_i + \varepsilon_{i,t} \tag{15}$$

This has been changed for the same reasons as stated above.

For RQ3, RQ4 and RQ6, we had originally planned to use the Kullback-Leibler divergence to demonstrate accuracy by looking at the divergence of the reported beliefs from the true distribution of the pots. We discovered that this was too noisy a measure of accuracy after having run the study. As such, we have moved to use the absolute value in difference of the mean of the distribution of the reported beliefs from the mean of the true distribution of the pots. A presentation of the analysis with the Kullback-Leibler divergence is available in appendix XX.

Within RQ2, for the hypothesis "*Belief updating is not different for the pots when framed neutrally*," we had original specified the null for the hypothesis test as:

$$H_0 : \beta_2 = \beta_3 = 0 \tag{16}$$

The correct test is in fact:

$$H_0 : \beta_2 + \beta_3 = 0 \tag{17}$$

This correction is to correctly analyse the hypothesis given. The original stated hypothesis would have tested that the no-framed treatment and the interaction between the no-framed treatment and pot B did nothing. This is correct as one would expect moving from the Framed-Female Pot to Non-Framed-Pot-B would see no changed when compared to non-framed pot-A. As such, one would expect the addition of the effects of these 2 pots to be 0.

For RQ3, we originally reported:

$$A_{i,g,t} = \beta_1 S_{i,g,t} + \beta_2 S_{i,g,t} t + \gamma X_i + \varepsilon_{i,t} \tag{18}$$

This has been changed to:

$$A_{i,g,t} = \alpha + \beta_1 S_{i,g,t} + \beta_2 t + \beta_3 S_{i,g,t} t + \gamma X_i + \varepsilon_{i,t} \tag{19}$$

The original specification would have captured both the impact of being in the later round within the β_2 coefficient as well as the impact of sampling in the later round of the experiment. In order to get a clean look at the impact of sampling the later rounds of the experiment, we introduce the variable for the second half of the experiment by itself to separate this effect out.

Similarly, for RQ4, we originally reported:

$$A_{i,g,t} = \beta_1 SR_{i,g,t} + \beta_2 SR_{i,g,t} t + \gamma X_i + \varepsilon_{i,t} \tag{20}$$

This has been changed to:

$$A_{i,g,t} = \alpha + \beta_1 SR_{i,g,t} + \beta_2 t + \beta_3 SR_{i,g,t} t + \gamma X_i + \varepsilon_{i,t} \tag{21}$$

Similar reasons apply as for the previous change. We want to be able to isolate the impact of the shortlisting treatment and having been in the shortlisting treatments in later rounds for accuracy.

For RQ5, we originally reported the specification:

$$\begin{aligned} S_{i,t} = {} & \beta_1 \mu_{i,w-m,t} + \beta_2 \sigma^2_{i,w-m,t} + \beta_3 PE_i + \beta_4 SR_i + \beta_5 PE_i \cdot SR_i + \beta_6 PE_i \mu_{i,w-m,t} + \beta_7 SR_i \mu^2_{i,w-m,t} \\ & + \beta_8 PE_i \cdot SR_i \cdot \mu_{i,w-m,t} + \beta_9 PE_i \sigma^2_{i,w-m,t} + \beta_{10} SR_i \sigma^2_{i,w-m,t} + \beta_{11} PE_i \cdot SR_i \cdot \sigma^2_{i,w-m,t} + \gamma X_i + \varepsilon_{i,t} \end{aligned} \tag{22}$$

This was changed to:

$$\begin{aligned} S_{i,t} = {} & \alpha + \beta_1 \mu_{i,w-m,t} + \beta_2 \sigma^2_{i,w-m,t} + \beta_3 PE_i + \beta_4 SR_i + \beta_5 PE_i \mu_{i,w-m,t} + \beta_6 SR_i \mu_{i,w-m,t} \\ & + \beta_7 PE_i \sigma^2_{i,w-m,t} + \beta_8 SR_i \sigma^2_{i,w-m,t} + \gamma X_i + \varepsilon_{i,t} \end{aligned} \tag{23}$$

The original specification was a mistake from a previous design of the experiment in which there were interactions across the shortlisting treatments, such that some participants would have simultaneously experienced the probabilistic exogeneity treatment and the shortlisting restriction treatment. This was no longer the case with the most up-to-date design and as such the interactions between the treatments are no longer required. For robustness checks, we originally specified that we would run a Hausman test for RQ2, RQ3, RQ4 and RQ5. However, we are unable to run the Hausman test for RQ2 and RQ5 as the variable of interest does not change across rounds, and as such we cannot

estimate a fixed effects model for these.

6.5 Appendix: Additional Checks on Belief Updating

In this appendix section, we provide two key checks on the belief updating analysis we see in the main paper and in the pre-registered analysis.

The first check we run is on whether the non-difference in divergence from the Bayesian benchmark remains when we only consider participants who shifted beliefs in the correct direction. By correct direction, we mean that they shifted their beliefs such that the mean of their newly elicited belief is greater than the mean of their previously elicited beliefs if the mean of the Bayesian benchmark distribution is higher than their current mean. Similarly, they should shift their elicited means downward if the mean of the Bayesian benchmark is lower than their original mean. We estimate the following specification, using a Pooled OLS model, using data from BO-F, BO-SS, BO-NF:

$$\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}} = \alpha + \beta_1 W + \beta_2 NF_i + \beta_3 W \times NF + \beta_4 SS_i + \beta_5 W_i \times SS_i + \gamma X_i + \varepsilon_{i,t} \tag{24}$$

Where $\Delta E_{i,g,t}$ is the difference between the empirical posterior mean and stated prior; $\Delta T_{i,g,t}$ is the difference between the theoretical posterior mean and stated prior; W is a dummy, taking a value of 1 if the given beliefs are about the distribution of women; NF_i is a dummy, taking a value of 1 if the given beliefs are elicited in the Only Beliefs - No-Framing treatment; SS_i is a dummy, taking a value of 1 if the given beliefs are elicited the Only Beliefs – Shortlisting Samples treatment; X_i represents the vector of individual characteristics used as controls; α is a constant; and $\varepsilon_{i,t}$ is the error term.

Table 11 presents the results of this specification. Model 1 shows the pre-registered analysis with all the data. Model 2 restricts the observations to those who update in the correct direction.

After running this analysis, our results we found in our pre-registered analysis do not change.

Table 11: Is the non-difference in belief updating robust to people going in the correct direction?

	(1)	(2)
1(Female Worker)=1	0.425	0.322
	(1.801)	(1.593)
1(No Framing (OB-NF))=1	-0.651	-0.794
	(1.834)	(1.672)
1(Female Worker)=1 X 1(No Framing (OB-NF))=1	0.170	2.801
	(2.454)	(2.093)
1(Shortlisting Samples (OB-SS))=1	-1.629	-0.592
	(1.487)	(1.355)
1(Female Worker)=1 X 1(Shortlisting Samples (OB-SS))=1	0.800	0.267
	(2.181)	(1.750)
1(Female Employer)	-1.490	-0.065
	(1.027)	(0.813)
1(Queen's Gambit)	-0.872	-0.440
	(1.056)	(0.570)
1(Chess Rules?)	-1.564	1.032
	(1.129)	(0.731)
1(Play Chess)	-0.067	-0.320
	(0.847)	(0.420)
1(Puzzle Rush)	1.421	1.078
	(1.536)	(1.272)
1(Hiring Role)	1.530	1.830*
	(1.034)	(1.079)
Age	-0.003	-0.131**
	(0.056)	(0.057)
1(Student)	0.658	-1.484*
	(1.817)	(0.880)
1(Full-Time Work)	-1.518	-0.411
	(1.018)	(0.795)
Constant	5.400**	7.220***
	(2.442)	(1.723)
Observations	3060	1399
Clusters	306	296
R2	0.004	0.017
BASE	No	No
SR	No	No
PE	No	No
BO_F	Yes	Yes
BO_SS	Yes	Yes
BO_NF	Yes	Yes

Cluster robust standard errors in parentheses.

* p<0.1 ** p<0.05, *** p<0.01.

(1) displays the original pre-registered analysis. (2) restricts the analysis to people who updated in the correct direction.

A second check we ran was to use a different metric for divergence in belief updating. This new metric looks at the distribution of the Bayesian benchmark and we compare this to the distribution of the participant's posterior beliefs, using the mean squared error.

We estimate the following specification, using a Pooled OLS model, using data from BO-F, BO-SS, BO-NF:

$$MSE = \alpha + \beta_1 W + \beta_2 NF_i + \beta_3 W \times NF + \beta_4 SS_i + \beta_5 W_i \times SS_i + \gamma X_i + \varepsilon_{i,t} \tag{25}$$

Where MSE is the mean squared error between the distribution of the Bayesian benchmark posterior and the participant's posterior distribution; W is a dummy, taking a value of 1 if the given beliefs are about the distribution of women; NF_i is a dummy, taking a value of 1 if the given beliefs are elicited in the Only Beliefs - No-Framing treatment; SS_i is a dummy, taking a value of 1 if the given beliefs are elicited the Only Beliefs – Shortlisting Samples treatment; X_i represents the vector of individual characteristics used as controls; α is a constant; and $\varepsilon_{i,t}$ is the error term.

The results of this regression are reported in (1) in table 12. We find no significant difference in divergence from the Bayesian benchmark, this time with divergence given by mean squared error.

Table 12: Do people show less error (MSE) in their update compared to the Bayesian benchmark by gender?

	(1)
1(Female Worker)=1	-10.494
	(8.014)
1(No Framing (OB-NF))=1	-1.522
	(17.692)
1(Female Worker)=1 X 1(No Framing (OB-NF))=1	30.478
	(20.634)
1(Shortlisting Samples (OB-SS))=1	-5.064
	(19.196)
1(Female Worker)=1 X 1(Shortlisting Samples (OB-SS))=1	16.735
	(13.491)
1(Female Employer)	34.464**
	(17.263)
1(Queen's Gambit)	22.568
	(21.515)
1(Chess Rules?)	-10.784
	(19.341)
1(Play Chess)	-1.646
	(16.941)
1(Puzzle Rush)	-4.748
	(32.530)
1(Hiring Role)	13.883
	(14.306)
Age	0.739
	(0.728)
1(Student)	17.139
	(20.711)
1(Full-Time Work)	14.410
	(16.175)
Constant	51.328
	(44.537)
Observations	3060
Clusters	306
R2	0.013
BASE	No
SR	No
PE	No
BO_F	Yes
BO_SS	Yes
BO_NF	Yes

Cluster robust standard errors in parentheses.
Where appropriate, FDR sharpened q-values are reported in {}. * p<0.1 ** p<0.05, *** p<0.01.

6.6 Appendix: Kullback-Leibler Divergence Analysis

In this section, we include the original analysis for the evolution of accuracy using the relative entropy between the true distribution and the elicited beliefs. We calculated relative entropy using the Kullback-Leibler Divergence:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log\left(\frac{P(x)}{Q(x)}\right) \tag{26}$$

In order to maintain comparability across individuals, the true distributions are inputted as $Q(x)$, and the elicited beliefs are inputted as $P(x)$. A lower relative entropy score indicates a small divergence between the two distributions. In our case, this means that a lower score implies greater accuracy.

Figure A3.1 shows the evolution of accuracy over the number of rounds using the KL divergence as a measure. We show two graphs. On the left we have the evolution of accuracy of beliefs in the framed scenario. On the right, we have the evolution of accuracy of beliefs in the non-framed scenario. On the vertical axis we have the accuracy of beliefs. On the horizontal axis we have the round number. The key thing that stands in this graph is the overall downward trend in all beliefs. However, the confidence intervals around these beliefs is far noisier than when using the deviation of mean from the beliefs as in figure **??** of the main paper. It is clear, however, that participants are getting more accurate beliefs. Additionally, we repeat the analysis of number observations and time on the accuracy of beliefs. For clarity of coefficients, we multiple the KL divergence by 100 to address the issues with small decimals. We estimated the following two specifications with a Pooled OLS model:

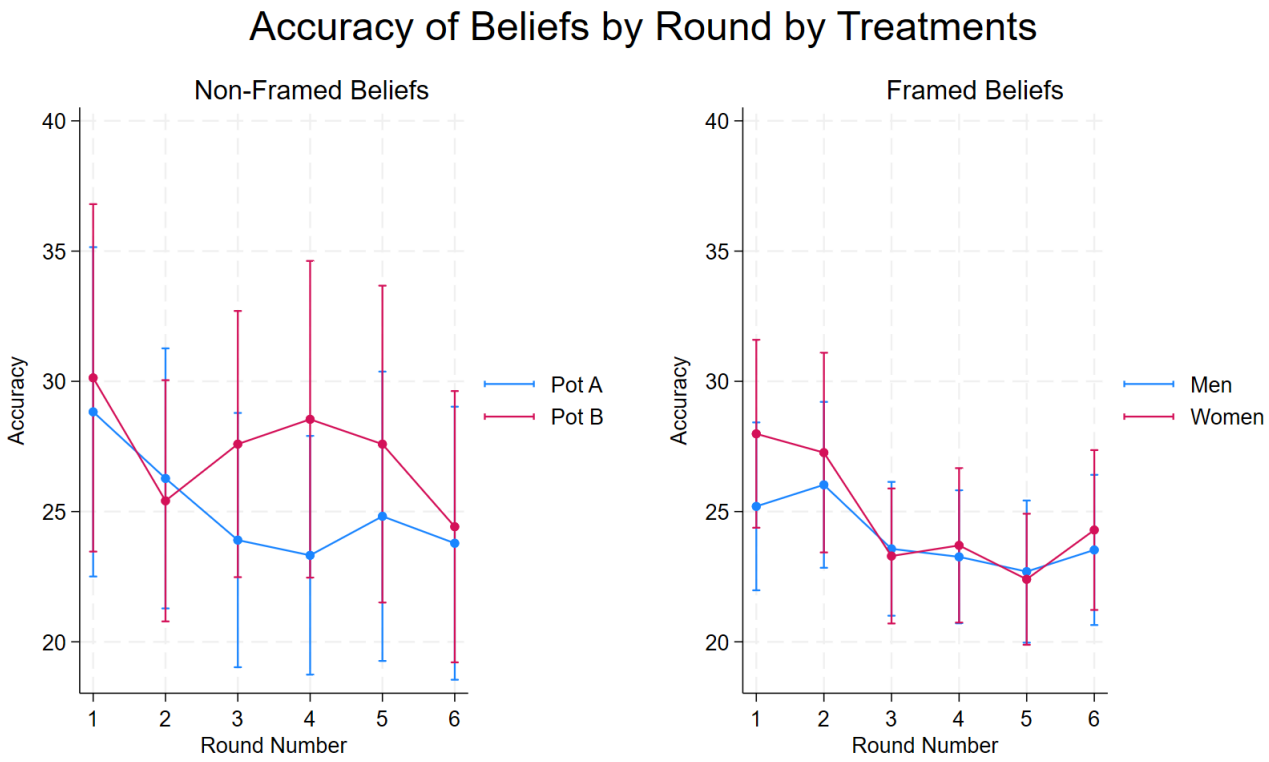
$$A_{i,g,t} = \beta_1 S_{i,g,t} + \gamma X_i + \varepsilon_{i,t} \tag{27}$$

$$A_{i,g,t} = \beta_1 S_{i,g,t} + \beta_2 t + \beta_3 S_{i,g,t} \gamma X_i + \varepsilon_{i,t} \tag{28}$$

Where $A_{i,g,t}$ is the relative entropy between i's beliefs about the distribution of group g in round t and the true distribution of group g, as given by the Kullback-Leibler divergence; $S_{i,g,t}$ is the number of 'workers' from distribution g that individual i has sampled up to round t; t is a dummy, taking a value of 1 if the beliefs were elicited in rounds 4 to 6; X_i represents the vector of individual characteristics used as controls.

The results of these regressions are displayed in Table **??** in (4) and (6). We report these estimations without controls in (1) and (3), respectively. Furthermore, we report the interaction between cumulative sampling and the second half dummy without the second half dummy by itself in (2), with controls in (5).

Figure 6: Average Prior Beliefs about Men and Women's Performance



Comparing this table to the table **??** in the main text, the coefficients on the Cumulative Observations are no longer significant. However, given the negative coefficients, this is suggestive that participants are getting more accurate beliefs as participants have more observations from the groups. Combining this table with the noisiness of the margins in figure 6, it is likely we were under-powered to detect the effect.

Table 13: Pooled OLS Regression: Determinants of Accuracy in Beliefs Only Treatments (with K-L divergence)

	(1)	(2)	(3)	(4)
Cumulative Observations	-0.085	-0.093	-0.221	-0.217
	(0.058)	(0.057)	(0.140)	(0.136)
1(Female Employer)		3.527		3.438
		(2.179)		(2.171)
1(Queen's Gambit)		2.403		2.414
		(2.450)		(2.442)
1(Chess Rules?)		3.480		3.494
		(2.341)		(2.338)
1(Play Chess)		-1.052		-1.041
		(1.776)		(1.774)
1(Puzzle Rush)		-2.514		-2.579
		(4.628)		(4.628)
1(Hiring Role)		1.904		2.044
		(2.128)		(2.130)
Age		0.045		0.042
		(0.088)		(0.088)
1(Student)		3.449		3.352
		(2.728)		(2.748)
1(Full-Time Work)		2.957		2.838
		(2.286)		(2.263)
1(Second Half)=1			-5.307**	-4.709*
			(2.570)	(2.460)
1(Second Half)=1 X Cumulative Observations			0.332**	0.300**
			(0.149)	(0.141)
Constant	26.108***	18.849***	27.134***	19.917***
	(1.193)	(4.735)	(1.379)	(4.748)
Observations	3672	3672	3672	3672
Clusters	306	306	306	306
R2	0.001	0.025	0.003	0.027

Cluster robust standard errors in parentheses.

* p<0.1 ** p<0.05, *** p<0.01.

6.7 Appendix: Hausman Tests: Random vs Fixed Effects

The purpose of this appendix section is to run the pre-registered Hausman tests where possible. It is important to note that the analyses in this section are done without controls or robust standard errors. These are done without controls as the fixed effects model removes time invariant regressors, which includes our controls. These are done without robust standard errors, as the Hausman test does not allow for robust standard errors in its analysis.

6.7.1 How does greater sampling from a distribution impact the accuracy of beliefs?

In order to compare like with like, our analysis will look at the impact of greater sampling from a distribution, a dummy for a second half round and the interaction between the two. The results of this analysis are reported in table 14 with the Random Effects model reported in model (1) and the Fixed Effects model reported in model (2)

Table 14: Random vs Fixed Effects: Determinants of Accuracy in Beliefs Only Treatments

	(1) Random	(2) Fixed
Cumulative Observations	-0.053***	-0.054***
	(0.006)	(0.006)
1(Second Half)=1	-0.473***	-0.423***
	(0.096)	(0.098)
1(Second Half)=1 X Cumulative Observations	0.044***	0.042***
	(0.007)	(0.007)
Constant	2.152***	2.155***
	(0.056)	(0.037)
Observations	3672	3672
Pseudo-R2	.	0.066

Standard errors in parentheses.

* p<0.1 ** p<0.05, *** p<0.01.

The first thing to note is that the results are fairly similar. That is, the coefficients do not vary substantially and everything remains significant at the p=0.01 level across both model types. The results of the Hausman test suggest that the coefficients are different (chi=10.84, p=0.01), which

under normal circumstances would mean we would prefer the fixed effects model.

However, for the analysis in the main text of the evolution of accuracy in response to sampling from a distribution, we choose to continue to present the random effects. The key reason for this is to maintain consistency with the other questions that also use random effects models.

6.7.2 How do different moments, specifically the mean and variance, of the perceived distributions over a population impact sampling behaviour?

In order to be able to compare like with like, we restrict the analysis to look at the impact of solely mean and variance on shortlisting behaviour. This is without the inclusion of treatments, treatment interactions and controls. These could not be included as, for each individual, the treatments and the controls are fixed effects, so any inclusion of these would mean we cannot compare like with like. The results of this analysis are reported in table 15 using the Pooled Poisson model for random effects in model (1) and the Conditional Fixed Effects Poisson model for fixed effects in model (2).

Table 15: Random vs Fixed Effects: Determinants of the Number of Woman Shortlisted in Shortlisting Treatments

	(1) Random	(2) Fixed
Mean Difference (W-M)	0.045***	0.023***
	(0.005)	(0.008)
Variance Difference (W-M)	-0.004***	-0.003*
	(0.001)	(0.002)
Constant	1.610***	
	(0.011)	
Observations	1764	1734
Pseudo-R2	0.013	.

Standard errors in parentheses.

* p<0.1 ** p<0.05, *** p<0.01.

The first thing to note is that all of the coefficients point in the same direction. That is, as the relative mean for women increases compared to men, more women are shortlisted and as the relative variance for women increases compared to men, less women are shortlisted. In the comparison between model (1) and (2), the significance of the impact of the mean continues to be significant

at p=0.01 whereas the significance of the variance weakens slightly to p=0.052. The results of the Hausman test suggest that the coefficients are sufficiently different (chi=11.46, p=0.003), which under normal circumstances would mean we would prefer the fixed effects model.

However, for the analysis in the main text of the determinants of shortlisting behaviour, we choose to continue to present the random effects. The key reason for this is we wished to still be able to investigate the impact of the treatments themselves on shortlisting behaviour. With a fixed effect model, this would not have been possible. Additionally, by using a random effects model, in this case the Pooled Poisson model, we were able to maintain consistency with the other questions that also use a random effects model.

6.8 Appendix: Exploratory Results

In this section, we conduct some exploratory analysis. First, we examine further belief updating behaviour and then we consider shortlisting behaviour.

6.8.1 Exploratory Belief Updating Analysis

One thing that we were interested in examining were the determinants of whether a participant updated in the correct direction. It may be too strong an assumption that participants update as a Bayesian would. However, even if we relax the requirement for correct updating to be in line with the Bayesian posterior, we can still gain insight into the correctness of some updating behaviour. One way to do this is by considering whether a participant updates in the correct direction.

Firstly, we need to define what it means to update in the correct direction. If the mean of the observed sample is greater than that of the mean of the elicited prior distribution, the mean of the posterior should be greater than that of the prior. Similarly, if the mean of the observed sample is less than that of the mean of the elicited prior distribution, the mean of the posterior should be smaller than that of the prior. As such, we define a participant as having updated in the correct direction if either the mean of their posterior is greater than the mean of their prior when the sample mean was greater than that of the prior; or if the mean of their posterior is smaller than the mean of their prior when the sample mean was smaller than that of the prior. On the hand, we define a participant as having updated incorrectly if their posterior is different to their prior, but they have moved in the opposite direction to what would have been suggested by the mean of the third sample. A third case is that a participant did not update their beliefs when presented with an observed sample mean that did not match their prior. We define this as "no shift." This third case does point to an edge case of correct updating, where the participant does not shift their mean, but the observed sample mean equals the mean of their prior. In this case, we classify this as a correct shift. A breakdown of how participants updated by round is in table **??**. The data in the breakdown comes from the all beliefs only treatments.

Observation 1: *Participants do a fairly good job of updating in the correct direction, successfully*

Table 16: Breakdown of how participants updated by round (in Beliefs Only Treatments)

	(1)					
	2	3	4	5	6	Total
Incorrect Shift	108	125	123	126	102	584
	(19.08)	(21.93)	(21.54)	(22.07)	(18.02)	(20.53)
No Shift	121	147	160	201	232	861
	(21.38)	(25.79)	(28.02)	(35.20)	(40.99)	(30.27)
Correct Shift	337	298	288	244	232	1399
	(59.54)	(52.28)	(50.44)	(42.73)	(40.99)	(49.19)
Total	566	570	571	571	566	2844
	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)

The frequency of each type of shift in a round. Round percentages are displayed in ().

updating 49.19% of the time. A further interesting observation is that the percentage of no shifts in beliefs increases as the rounds progress, moving from 21.38% of cases in round 1 to 40.99% of cases in round 6. This seems to be driven by a movement from correct shifts to no shifts.

This leads onto the interesting question of what the determinants of correct updating are when defined as above. One potential update was the size of the theoretical shift i.e. the absolute size of the difference between the Bayesian posterior and the elicited prior. In order to estimate this, we estimated the following multinomial logit model, using a correct shift as the base outcome:

$$Pr(Y = y|X) = logit(\alpha + \beta_1 \Delta T_{i,g,t} = \gamma X + \varepsilon_{i,t}) \tag{29}$$

Where $\Delta T_{i,g,t}$ is the absolute difference between the theoretical posterior mean and prior; X_i represents the vector of individual characteristics used as controls; *alpha* is a constant; and $\varepsilon_{i,t}$ is the error term.. The data used for these models comes from the beliefs only treatments.
The marginal effects of this regression are given in Table **??**.

Observation 2: *Participants are more likely to update correctly if the theoretical shift is large.*

Support: From the above regression results, we see some slightly significant evidence that participants are more likely to update in the correct direction as the size of the theoretical mean increases. The coefficient suggests that an increasing the absolute size of the theoretical shift by 1 increases the probability that a participant updates in the correct direction by 22%. There is also some evidence that women are less likely to update in the correct direction in favour of not shifting their beliefs.

It is interesting to note that in some contexts, seeing no shifts may be rationalizable. The metric

Table 17: Marginal Effects of for the Probabilities of Updating Behaviour

	P(Incorrect Shift)	P(No Shift)	P(Correct Shift)
1(Female Worker)	0.0146	-0.000341	-0.0142
	(0.0174)	(0.0164)	(0.0211)
Absolute Theoretical Change	-0.0747	-0.0519	0.127
	(0.0738)	(0.0318)	(0.0651)
1(Female Employer)	0.00638	0.0735	-0.0799
	(0.0226)	(0.0362)	(0.0321)
1(Queen's Gambit)	-0.00394	-0.00168	0.00562
	(0.0219)	(0.0363)	(0.0335)
1(Chess Rules?)	-0.0311	0.0787	-0.0475
	(0.0310)	(0.0415)	(0.0367)
1(Play Chess)	0.00880	-0.0280	0.0191
	(0.0183)	(0.0277)	(0.0261)
1(Puzzle Rush)	0.0216	-0.0233	0.00166
	(0.0525)	(0.0654)	(0.0647)
1(Hiring Role)	-0.0216	0.0517	-0.0301
	(0.0260)	(0.0392)	(0.0353)
Age	-0.000667	-0.00166	0.00233
	(0.00103)	(0.00156)	(0.00150)
1(Student)	-0.0236	-0.00216	0.0258
	(0.0271)	(0.0470)	(0.0405)
1(Full-Time Work)	0.0515	-0.00640	-0.0451
	(0.0241)	(0.0374)	(0.0338)

Standard errors in parentheses

used to define a correct shift is very sharp, in that any size of theoretical update matters, even if it is only epsilon higher than the prior belief. It could be that some of the no shifts are a result of the granularity of how the sliders move. Our sliders move through integers, and as such, participants may not be able to change the sliders with the precision required to update by this epsilon amount. However, it could also be that, due to bounded rationality and the effort that it would require, subjects may be unable to unwilling to compute this theoretical change, and therefore report no shift.

Additionally, we investigated whether participants react differently to different types of draws when updating beliefs in the framed beliefs only treatments. In particular, we looked at whether participants reacted differently to receiving samples consisting of observations from either the high or low end of the distribution. We defined receiving an observation from the high end as seeing at least 1 observation who got at least 25 but less than 30; and we defined receiving an observation from the low end as seeing at least 1 observation who got at least 5 but less than 10. Additionally, we looked at whether the gender of these observation impacted the updating behaviour.

We estimated the following specification, with a Pooled OLS model:

$$\frac{\Delta E_{i,g,t}}{\Delta T_{i,g,t}} = \alpha + \beta_1 H_{i,g} + \beta_2 L_{i,g} + \beta_3 W + \beta_4 H_{i,g} \times W + \beta_5 L_{i,g} \times W + \gamma X_i + \varepsilon_{i,t} \tag{30}$$

Where $H_{i,g}$ is a dummy, taking a value of 1 if there is an observation within the presented sample from gender g, with a score of at least 25 but less than 30; $L_{i,g}$ is a dummy, taking a value of 1 if there is an observation within the presented sample from gender, g, with a score of at least 5 but less than 10; W is a dummy, taking a value of 1 if the given beliefs are about the distribution of women; X_i represents the vector of individual characteristics used as controls; X_i represents the vector of individual employer characteristics; *alpha* is a constant; and $\varepsilon_{i,t}$ is the error term. The data used in these treatments come from Only Beliefs-Framing and Only Beliefs – Shortlisting Samples as we are also concerned with the impact of gender on this behaviour. The results of this regression are given in Table **??** in model (2) and without controls in model (1).

Observation 3: *Participants do not update differently when seeing workers from the high end or low end of the distribution.*

Support: The above regressions include dummies for both seeing a high valued worker and a low valued worker. The majority of coefficients are reported to be insignificant, with the exception of slight significance on the interaction between a low worker and a female worker, such that a participant would over-update less than if they hadn't seen this type of worker.

Table 18: Do participants update differently to high or low observations?s

	(1)	(2)
1(High Worker)=1	-0.398	-0.357
	(2.406)	(2.433)
1(Low Worker)=1	0.315	0.673
	(1.811)	(1.827)
1(Female Worker)=1	1.782	1.780
	(1.351)	(1.344)
1(High Worker)=1 X 1(Female Worker)=1	-3.732	-3.707
	(2.878)	(2.884)
1(Low Worker)=1 X 1(Female Worker)=1	-3.696*	-3.724*
	(2.156)	(2.162)
1(Female Employer)		-0.920
		(1.105)
1(Queen's Gambit)		0.047
		(1.107)
1(Chess Rules?)		-2.628*
		(1.378)
1(Play Chess)		-0.614
		(0.654)
1(Puzzle Rush)		1.853
		(1.604)
1(Hiring Role)		1.166
		(1.350)
Age		0.009
		(0.090)
1(Student)		1.556
		(2.419)
1(Full-Time Work)		-0.881
		(1.290)
Constant	2.607***	4.885
	(0.865)	(3.350)
Observations	2000	2000
Clusters	200	200
R2	0.003	0.008

Cluster robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

6.8.2 Exploratory Shortlisting Analysis

Another area we wished to explore was the perceived certainty of beliefs and its impacts on shortlisting behaviour. In the post-game questionnaire, we collected responses to a number of simple Likert-Scale questions for how certain our participants were about their beliefs of the groups. The first set of these asked "In round 1, how certain were you about your estimates about 'Pot-Men'?" for the male pot, and a similar question as asked in reference to the female pot. The second set of these asked "If there was more than one round, in later rounds, how certain were you about your estimates about 'Pot-Men'?" for the male pot, and again a similar question was asked in reference to the female pot.

Observation 4: *There was no reported difference in initial certainty about beliefs for men and women.*

Support: Using the data from all treatments except "OB-NF," we ran a paired t-test comparing participants' reported initial certainty of the male pot and the female pot, against the null of no difference. We exclude "OB-NF" from this test, as this treatment asks a similar set questions but framed as "Pot A" and "Pot B." We fail to reject the null, suggesting there is, at the population level, no difference in certainty.

Observation 5: *Participants' reported certainty increases across rounds for both men and women.*

Support: We ran paired t-tests comparing participants' reported initial certainty of beliefs against reported certainty in later rounds twice, once for the male pot and once for the female pot, again excluding "OB-NF". Both of these were compared to the null of no change. We reject the null of no difference for both the male pot (mean difference: -.757, p=0.000) and the female pot (mean difference: -.738, p=0.000).

While there was no difference in population perceptions of certainty, some participants did report initially being more certain about one pot than another pot. We go onto investigate whether differences in initially certainty drives some of the differences in shortlisting behaviour, in the shortlisting treatments. We compute a variable of the difference in initial certainty between women minus that of men, and use it as a dependent variable in the following specifications, estimated with Poisson models:

$$S_i = \alpha + \beta_1 C_{i,w-m} + \gamma X_i + \varepsilon_i \tag{31}$$

$$S_i = \alpha + \beta_1 C_{i,w-m} \beta_2 \mu_{i,w-m,t} + \beta_3 \sigma^2_{i,w-m,t} + \gamma X_i + \varepsilon_i \tag{32}$$

Where S_i is the number of women shortlisted in round 1; $C_{i,w-m}$ is individual i's perceived initial difference in certainty of Pot-Women minus Pot-Men; $\mu_{i,w-m,t}$ is individual i's perceived difference

in mean of Pot-Women minus Pot-Men in round 1; $\sigma^2_{i,w-m,t}$ is individual i's perceived difference in variance of Pot-Women minus Pot-Men in round 1; X_i represents the vector of individual characteristics used as controls; *alpha* is a constant; and $\varepsilon_{i,t}$ is the error term. The data for this regression comes from all shortlisting treatments.

The results of these regressions are displayed in Table **??**. The first regression is reported in model (1) with its margins reported in (2). The second regression is reported in model (3) with its margin reported in (4).

Observation 6: *Participants shortlist more from the pot about which they have more certain beliefs.*

Support: From the above regression, we see that differences in initial certainty about beliefs of the male distribution and initial certainty about beliefs of the female distribution impact behaviour. From the model in (2) and its margins in (1), we see that an increase in the initial difference in certainty of 1 point on the Likert-Scale, which would represent an increase of comparative certainty about the female distribution, increases the number of women shortlisted by 0.749. This is significant at the 1% level. Similar results are obtained when controlling for the perceived differences in mean and variance in the other models of the table.

Table 19: Do differences in initial certainty impact shortlisting behaviour?

	(1)	(2) Margins	(3)	(4) Margins
Certainty Difference (W - M)	0.182***	0.832***	0.165***	0.752***
	(0.047)	(0.215)	(0.045)	(0.208)
Mean Difference (W-M)			0.043***	0.195***
			(0.010)	(0.047)
Variance Difference (W-M)			-0.003	-0.012
			(0.002)	(0.008)
1(Female Employer)	-0.081**	-0.371**	-0.074**	-0.339**
	(0.039)	(0.177)	(0.038)	(0.171)
1(Watched Queen's Gambit)	-0.054	-0.248	-0.057	-0.259
	(0.039)	(0.180)	(0.038)	(0.172)
1(Chess Rules?)	-0.052	-0.238	-0.077*	-0.350*
	(0.048)	(0.219)	(0.047)	(0.212)
1(Play Chess Regularly)	0.039	0.177	0.048**	0.218**
	(0.024)	(0.111)	(0.024)	(0.107)
1(Played Puzzle Rush)	-0.059	-0.272	-0.034	-0.156
	(0.080)	(0.366)	(0.082)	(0.375)
1(Hiring Role)	-0.025	-0.114	-0.023	-0.104
	(0.038)	(0.174)	(0.035)	(0.161)
Age	-0.003*	-0.014*	-0.002	-0.011
	(0.002)	(0.007)	(0.002)	(0.007)
1(Student)	-0.073	-0.333	-0.057	-0.260
	(0.067)	(0.304)	(0.068)	(0.308)
1(Full-Time Work)	0.077**	0.351**	0.070*	0.318*
	(0.037)	(0.169)	(0.036)	(0.164)
Constant	1.656***		1.639***	
	(0.083)		(0.084)	
Observations	294		294	
Clusters	294		294	
R2	.		.	

Cluster robust standard errors in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

6.9 Appendix: Balance Test

Table 20 summarises the means of the control variables by treatment. Additionally, it reports the F statistics and p-values of a One-way ANOVA test in order to check for balance.

Given the details in Table 20, it appears the randomization failed on 3 dimensions. It appears the age and proportions of students in at least one treatment is different from the others at a 0.01% significance level and the proportion of participants who have played puzzle rush is different in at least one treatment at the 0.1% significance level. Given this, in the later analysis, it will be important to place more emphasis on the regression models with control variables.

Table 20: Means of Control Variables by Treatment

	(1) NSR-NPE	(2) SR-NPE	(3) NSR-PE	(4) OB-F	(5) OB-NF	(6) OB-SS	(7) F Stat	(8) P value
1(Female Employer)	.4653465	.5212766	.5757576	.5	.4811321	.51	0.590	0.709
1(Watched Queen's Gambit)	.3663366	.3404255	.3939394	.31	.3113208	.41	0.770	0.568
1(Chess Rules?)	.6435644	.6702128	.6262626	.62	.6226415	.71	0.550	0.741
1(Play Chess Regularly)	1.693069	1.819149	1.787879	1.74	1.660377	1.81	0.650	0.6631
1(Played Puzzle Rush)	.029703	.0319149	.020202	.08	.0283019	.09	2.05*	0.070
1(Hiring Role)	.5049505	.4680851	.4747475	.46	.6037736	.45	1.36	0.238
Age	39.10891	40.04255	38.14141	37.21	39.48113	31.16	7.030***	0.000
1(Student)	.1485149	.0744681	.1616162	.17	.0849057	.26	3.600***	0.003
1(Full-Time Work)	.4158416	.4893617	.4444444	.38	.509434	.44	0.920	0.470
Observations	101	94	99	100	106	100		

*** p<0.01, ** p<0.05, * p<0.1

6.10 Appendix: Prolific Demographic Data Note

When using the demographic data of participants provided by Prolific, it is possible that some of the data expires, in the sense that it becomes out of date. In our case, this may have been a problem for "Student status" and "Employment status," as we use these as controls, under "1(Student)" and "1(Full-Time Work)," respectively. In cases where the data had expired under "Student status," we coded the "1(Student)" variable as as 0, i.e. not a student. In cases where the data had expired under "Employment status," we coded the "1(Full-Time Work)" variable as as 0, i.e. not in full time work.