Organisational Values, Self-Image and Inclusion: Evidence from a Field Experiment*

Girum Abebe[†]
Siân Brooke[‡]
Tom Gole[§]
Simon Quinn[¶]
Tom Schwantje[‖]

May 5, 2026

Abstract

We conduct a field experiment within a business plan competition to examine how institutional features influence inclusive decision-making. In the control group, over the course of independent sequential assessments, evaluators become less likely to recommend female candidates. This pattern reduces the quality of their decisions – as measured by expert assessments and machine-learning methods. Informing judges of the organisation's commitment to equal opportunity entirely offsets this decline, whereas requiring judges to justify their decisions to peers has a more muted effect. Our results show that decision fatigue can undermine both decision quality and inclusivity, but simple organisational messaging can resolve this.

*We thank Vojtěch Bartoš, Elia Benveniste, Marianne Bertrand, Ralph de Haas, Maddalena Ronchi, Chris Roth, Marc Witte and Chris Woodruff for helpful comments and suggestions, and in particular to Karmini Sharma, Rodolfo Stucchi and Anasuya Narasimhan their thoughtful discussions of the paper. We are grateful to audiences at the CSAE Conference and research workshops at Bocconi and Oxford University. This project was funded by the Private Enterprise Development in Low-Income Countries programme (PEDL). We also want to thank Gezahegn Gbremedhin for outstanding research assistance. The project would not have been possible without the constant support of Rose Page and the Centre for the Study of African Economies (University of Oxford), nor without the support of EconInsights in Addis Ababa. The experiment was pre-registered with the AEA RCT registry (AEARCTR-0009893) and IRB approval was obtained from the University of Oxford (ECONCIA21-22-39).

[†]**International Finance Corporation**: gtefera@ifc.org
[‡]**University of Amsterdam**: s.j.m.brooke@uva.nl
[§]**Queensland Treasury Corporation**
[¶]**Imperial College London**: simon.quinn@imperial.ac.uk
[‖]**Bocconi University**: tom.schwantje@unibocconi.it

Recent debates over DEI – diversity, equity and inclusion – have often been framed as arguments about meritocratic evaluation. In Silicon Valley, for example, some firms and investors have advocated replacing DEI initiatives with principles of 'MEI': 'merit, excellence, and intelligence'. Advocates of this alternative view argue that organisations should evaluate candidates solely on quality rather than demographic criteria: that candidate diversity is often a *substitute* for candidate quality. In contrast, proponents of DEI contend that their efforts correct biases against high-quality candidates from under-represented groups: that candidate diversity and candidate quality can be *complements*.[1] Within organisations, decisions such as hiring, lending, and grading are typically delegated to individual decision-makers who possess context-specific information but whose biases the organisation cannot fully control (Aghion and Tirole, 1997; Dessein, 2002). Together, these observations raise a central question (Goldin and Rouse, 2000): can institutions make delegated processes more inclusive without reducing decision quality? This challenge may be particularly acute when decision-makers are busy or fatigued: conditions under which implicit biases are more likely to surface (Bertrand, Chugh, and Mullainathan, 2005; Chugh, 2004).

To study how fatigue and institutional design shape evaluative decisions, we embed a field experiment in a business plan competition in Ethiopia. In this competition, senior human resources managers from large Ethiopian firms assess young professionals' business proposals for a prize of 50,000 Ethiopian Birr (approximately 800 USD). This experimental design allows us to answer three questions. First, *how does decision fatigue affect inclusive institutional decision-making*? To answer this, we ask judges to repeatedly assess pairs of candidates in an environment in which other plausible mechanisms are unlikely by design. Second, *to what extent can this be offset by institutional design*? To answer this question, we focus on the interaction of two common institutional features: communication of organisational values about equal opportunity, and self-image concerns generated by the prospect of having to justify decisions to peers. Specifically, we randomly assign half of our judges to an 'Organisational Values' treatment – in which they are exposed to a message emphasizing the importance to the organisation of promoting equal access to capital for female entrepreneurs. We also randomly assign half of our judges to a 'Self-Image' treatment, in which they are told in advance that they will be required to discuss and justify their decisions with peer judges – but only after making all their decisions. We randomize these two interventions independently, allowing us to estimate both their individual effects and their interaction.

This leads to a critical third question: *do these institutional designs improve decision quality?* We think of decision quality as the quality of the proposals recommended for prizes, irrespective of candidates' gender. If discrimination is taste-based, and particularly if fatigue makes such preferences more likely to shape decisions, then reducing this bias should increase both inclusion and decision quality by making high-quality proposals from female candidates more likely to be recommended. In contrast, the treatments could simply make judges vote more for women regardless of proposal quality, thus increasing inclusivity at the expense of decision quality. The answer matters for organisations: it determines whether institutions face a trade-off between inclusion and decision quality, or whether simple changes to evaluative processes can improve both.

The assessments are structured as a series of binary comparisons – each between a male and a female candidate.

[1] For example, Scale AI CEO Alexandr Wang outlined his view of MEI in a 2024 company blog post: Wang (2024). By contrast, Apple CEO Tim Cook defended Apple's inclusion efforts at the company's 2025 annual shareholder meeting – arguing that Apple's strength comes from hiring the best people and fostering a culture in which people with diverse backgrounds and perspectives can do their best work (Salon, 2025).

In the control group, evaluators vote for female candidates in 48.6% of the assessments in the first six of twelve assessments, suggesting limited bias against female candidates. In the last six rounds, this falls by 10.2 percentage points, to 38.4%. We show that this decline is driven by worse assessments of those female candidates that, according to two independent HR experts, should have won – indicating a fall in both inclusivity and in decision quality.[2] Following the literature in organisational psychology and behavioural economics, we refer generically to such behaviour as '*decision fatigue*': as evaluators become tired over repeated assessments, they rely more on heuristics and less on careful evaluation (Augenblick and Rabin, 2019; Bodenhausen, 1990; Danziger, Levav, and Avnaim-Pesso, 2011).[3]

In the first six rounds, none of our three treatments has a significant effect. However, in the last six rounds, the Organisational Values treatment succeeds in eliminating the deterioration of both decision quality and the success rate of female candidates. This treatment effect is concentrated among those female candidates who were preferred by the experts. The Self-Image treatment by itself has a directionally similar, though more muted, effect. To reduce our reliance on subjective expert scores and better understand which female candidates benefit, we use two machine learning methods – endogenous stratification (Abadie, Chingos, and West, 2018) and generic machine learning (Chernozhukov, Demirer, Duflo, and Fernández-Val, 2025). Together, these methods show that the Organisational Values treatment is particularly valuable for female candidates who would have performed well had they been assessed in the first six rounds (before the decision fatigue took effect). These candidates are significantly more likely – for example – to have a clear business concept and an understanding of the market, are more likely to discuss business operations, are more likely to be perceived as articulate and to use specific examples; they are also judged as less likely to have spent effort on their physical appearance.

First, we contribute to a large literature documenting gender disparities in entrepreneurial finance and capital allocation (Ewens, 2023; Hebert, 2025), including evidence from both credit markets and equity finance (Alesina, Lotti, and Mistrulli, 2013; Bartoš, Castro, Czura, and Opitz, 2024; Brock and De Haas, 2023). We show that bias against female candidates emerges dynamically over repeated assessments, consistent with decision fatigue.[4] This finding helps reconcile recent evidence from field experiments in Ethiopia showing limited gender bias in access to finance (Ayalew, Manian, and Sheth, 2023; Buehren and Papineni, 2025) with descriptive work highlighting that women face substantially worse access to finance in this setting (World Bank, 2022). More broadly, our results provide evidence consistent with the theoretical distinction between implicit and explicit discrimination (Bertrand et al., 2005; Chugh, 2004) – biases that are absent under full cognitive control surface as decision fatigue sets in (Wigboldus, Sherman, Franzese, and Knippenberg, 2004). A set of closely related papers highlight different, related mechanisms. Miserocchi (2024) shows that constraints on evaluators' ability to retrieve individual-level information amplify gender discrimination. Bartoš, Bauer, Chytilová, and Matějka (2016) show that discrimination can operate through attention (employers devote less effort to screening minority applicants). Benson, Li, and Shue (2026) find that gender gaps arise from subjective assessments of 'potential' rather than performance.

[2] Three related papers also study how the sequence of candidates distorts evaluations: Chen, Moskowitz, and Shue (2016); Kessler, Low, and Shan (2026); Radbruch and Schiprowski (2023).

[3] See also Weinshall-Margel and Shapard (2011) and Glöckner (2016) for alternative explanations for the result in Danziger et al. (2011).

[4] Kessler, Low, and Sullivan (2019) provide related evidence of this mechanism.

Second, we contribute to the literature on the design of organisational processes. A recent literature shows that communicating organisational mission can substitute for financial incentives in motivating effort (Ashraf, Bandiera, and Jack, 2014; Cettolin, Cole, and Dalton, 2024; Khan, 2025). We extend this insight: reminding judges of the organisation's commitment to equal opportunity sustains the effortful processing needed to assess candidates on merit, without constraining their ability to rely on individual expertise. This is light-touch relative to related interventions such as informing evaluators of their implicit biases directly (Alesina, Carlana, La Ferrara, and Pinotti, 2024) or changing evaluation processes (Bohnet, Van Geen, and Bazerman, 2016; Miller, Lall, Goldstein, and Montalvao, 2023), but is also effective. Closer to our treatment, Bursztyn, Fiorin, Gottlieb, and Kanz (2019) show that reminding borrowers of their moral obligation to repay increases repayment – suggesting that simple normative messages can sustain prosocial behaviour across settings. We then show that a second mechanism – self-image concerns arising from anticipated peer accountability – has a less pronounced effect. This complements evidence in Chen, Chen, and Yang (2025) and Brock (2025) by showing that self-image concerns have some positive effects on both reducing biases and improving decision quality. Together, our results suggest that low-cost institutional interventions can substantially improve both inclusivity and decision quality by counteracting decision fatigue.

The remainder of the paper proceeds as follows. Section 1 describes the experimental design, setting and sample in detail, Section 2 presents both the regression and machine learning results, and Section 3 concludes.

1 Experimental design

In this section, we describe the setting for our experiment – a business plan competition in Addis Ababa. We go on to explain the four treatment conditions, and then describe the sample. Appendix D provides detail on the pre-specified experimental design and randomisation.

1.1 Setting: A business plan competition

To implement our experiment, we ran a business plan competition among young professionals in Addis Ababa.[5] Competitors were invited to a studio to record a pitch for their business proposal, and these videos form the input for our experiment.[6] The competition took place in a league-style format: each competitor was placed in one out of ten pools of ten candidates, with the winner of each pool receiving a cash prize of 50,000 Birr. Each pool of ten candidates comprised five female candidates and five male candidates. We asked the business plan judges to cast a series of twelve binary votes – where each vote compared one female and one male candidate.[7] For each competition, the candidate with the most weighted votes – pooling judge and HR expert assessments – won the prize.

[5] Specifically, young professionals were eligible for this competition based on their assessment by established firm managers in a separate field experiment: see Abebe, Fafchamps, Koelle, Quinn, and Schwantje (2026).

[6] Specifically, competitors were asked to prepare a three-minute pitch briefly introducing themselves and covering the following five points: (i) their business idea, (ii) their intended target market, (iii) their potential competition, (iv) their operations and (v) their cost of business.

[7] The remaining pairs – that is, the male-male pairs and the female-female pairs – were assessed by two human resources experts. In the interests of fairness across competitors, we weighted the votes such that the votes from the human resources experts had the same total weight as those of the judges.

1.2 Assessing the candidates: Four treatment conditions

Each judge was randomly assigned to one of four different conditions under which to assess the videos. Our experiment took place over a period of three weeks in the summer of 2022 – during which we ran two sessions per day, randomly varying which treatment arm was implemented in each session. Randomisation was done at the judge level: after agreeing to participate, each judge was assigned to one treatment condition (and invited to a specific session accordingly). We grouped the assessments so that each pair of candidates was viewed by 12 judges: three judges in each of the four treatment conditions. We randomised the order in which assessments took place (so that, for example, what is the first pair for one judge might be the sixth pair for a different judge). To avoid any learning or spillovers, judges were asked not to discuss the assessment of the competitors with other judges until they had finished all assessments. We now describe each treatment condition in turn.

Condition 1: Control. Under the 'control condition', each judge assessed the videos individually, and without any messaging as to organisational purpose. Specifically, judges were invited to a venue in central Addis Ababa in groups of 12 to 15. All judges entered a single room in a classroom-style arrangement. Once all judges had arrived, the judges were shown two videos. The first video explained the competition and the judges' task (including details on prizes, and on how the competition winners would be determined). In the second video, a prominent Ethiopian businessman and former Olympian, Haile Gebrselassie, spoke briefly about the importance of the business plan competition in providing capital for Ethiopian entrepreneurs. Specifically, he said the following:

> As you know, access to capital is limited for entrepreneurs in Ethiopia. This competition will provide an opportunity for entrepreneurs to access capital to start or grow their business. Your vote is important in deciding which individual will win the 50,000 Ethiopian birr prize; please consider your choices carefully.

After watching these videos, the judges started their assessments. These were done on tablet computers, with the assistance of facilitators. Specifically, each judge was played a series of video pairs; for each pair, the judge cast a vote for the candidate he or she recommended for the grant. Judges under the control condition knew that their peers were also assessing candidates – but judges were not told anything about which candidates their peers were assessing, and there was no subsequent discussion about any vote cast. Judges were only asked to, privately, provide some feedback to two of the candidates they assessed at the end of their assessments.

Condition 2: Organisational values condition. The 'Organisational Values' condition was designed to emphasise to judges that the organisers of the business plan competition particularly value the inclusion of female candidates. This condition differed from the control condition in just one respect. Specifically, the second video included an additional prompt on the importance of equal opportunity for the organisation – noting, in particular, the specific constraints that female entrepreneurs face in accessing capital in Ethiopia. We deliberately designed this as a light-touch treatment, acknowledging that judges need to consider a range of factors in assessing business plans. Specifically, the text is as follows (with the text in italics distinguishing the Organisational Values condition from the

control condition):[8]

> As you know, access to capital is limited for entrepreneurs in Ethiopia. This competition will provide an opportunity for entrepreneurs to access capital to start or grow their business. *Considering equal opportunity: I realise you need to take into account a large number of factors when making your decision but would like you to keep in mind that when starting a business, female entrepreneurs face additional constraints due to lenders' biases. A recent World Bank report finds that male entrepreneurs are more likely to take out loans than female entrepreneurs. In terms of loan sizes, male entrepreneurs borrow about 50 percent more than female entrepreneurs. In this competition, we are committed to gender equality and want to promote male and female entrepreneurs equally.* Your vote is important in deciding which individual will win the 50,000 Ethiopian birr prize; please consider your choices carefully.

Condition 3: Self-Image. The 'Self-Image' condition was designed to allow the possibility of self image concerns. We did this in several complementary respects. First, before watching the two explanatory videos[9] (and before starting their assessments), judges were asked to introduce themselves to the other judges – providing their name, company and position. Second, prior to each of their binary votes, each judge was shown the names and photographs of two other judges who would also be voting on the same pair of candidates. Third, at the conclusion of all voting, we randomly chose one of the pairwise votes; each judge then sat together with the other judges who assessed that pair (that is, the other two judges whose names and photographs had been shown before that particular vote). In their groups of three, an enumerator shared who each judge voted for, before judges justified their votes to their peers, and then jointly agreed on feedback for the two candidates. Judges under the 'Self-Image' condition were told in advance, and before casting each vote, that this meeting would take place.

Condition 4: Combined condition. The final condition combined the features of the 'Organisational Values' condition and the 'Self-Image' condition: this was designed to test whether communication of organisational values has a different impact when assessors are exposed to Self-Image concerns. Specifically, this condition was identical to the 'Self-Image' condition, but judges watched the same second video as those in the 'Organisational Values' condition.

Theoretical model. To guide intuition in understanding our treatments, we present a game-theoretic model in Appendix A. In this model, evaluators trade off three forces: (i) alignment with their own view of what matters, (ii) organisational pressure to emphasise particular dimensions, and (iii) a cost of disagreeing with other evaluators. The model highlights a simple mechanism. As attention degrades – for example due to fatigue or cognitive load – evaluators place greater weight on their own preferred evaluation rules and reduce alignment with organisational objectives. Organisational messages can partially offset this by re-orienting evaluators toward the intended criteria. These forces generate three predictions that map directly to our experimental design: (i) increases in cognitive costs

[8] The 'recent World Bank report' referred to in this text is the Ethiopia Gender Diagnostic Report (World Bank, 2022).
[9] The first video was slightly adapted to reflect the change in the protocol.

increase reliance on individual preferences and reduce alignment with organisational criteria, (ii) organisational pressure shifts evaluations without necessarily affecting dispersion, and (iii) greater concern about disagreement compresses variation in evaluations. The model also implies that when alignment is already high, evaluations may already represent the organisational benchmark, so that additional increases in organisational pressure have limited marginal impact.

1.3 Experimental participants: Ethiopian managers

We invited senior Ethiopian managers from established Ethiopian firms to serve as judges.[10] These judges are, by the nature of their roles, deeply familiar with the Ethiopian business environment and the challenges facing firms; they are also a highly relevant sample for testing attitudes towards issues of hiring and inclusion in the Ethiopian labour market. Table A.6 reports summary statistics describing the judges. Most judges are either the most senior manager or owner of the firm (33%) or human resources manager (32%). They are on average 41 years old, and have on average 20 years of professional experience, on average six of which are in their current position. The managers are highly educated, with 78% having a bachelor's degree and 77% having formal management education. Three quarters of the judges are male, one quarter are female. In the final column of Table A.6 we report the p-value for a Wald-test that the judges' characteristics are balanced across treatments, and find no evidence for imbalance for any of the characteristics.

The managers work for large firms with a median of 54 employees (mean 200). 95% of these firms are for-profit, 62% are private limited companies, and 16% are public limited companies. Most firms are located in the capital Addis Ababa, with a few based in the nearby cities of Adama and Bishoftu.

1.4 Additional assessment: Human resources experts

In addition to the experiment's primary assessments, we had four individuals independently evaluate all competition submissions. This group included two highly experienced enumerators and two senior HR managers. The enumerators focused on transcribing more 'factual' dimensions of the vignettes, such as how the competitors were dressed, how they composed themselves, and whether they addressed each required topic in their pitch. The two senior managers, henceforth referred to as our 'HR Experts', first completed a shortened version of the enumerators' questionnaire. They then assessed the overall quality of the pitch on a scale from one to twenty and evaluated specific dimensions of the pitch, such as whether the competitors had a clear business concept and a strategy for growth.[11] These latter questions were drawn from Fafchamps and Quinn (2018). The experts, who include the owner of an Ethiopian HR firm and the HR head of a large for-profit enterprise, were selected by our local partners for their in-depth understanding of the labour market and ability to assess the quality and viability of business plans.

[10] These are drawn from a sample of managers that participated in a separate field experiment: see Abebe, Fafchamps, Koelle, and Quinn (2026). Appendix D.4 details the invitation for judges, which is translated into Amharic before being shared with enumerators. Enumerators are told to closely stick to these scripts and not to give additional information discussing the experiment or how we expect judges to make their decisions.

[11] We are deliberately agnostic about the criteria: we simply ask the overall score to indicate whether the expert thinks the candidate should be awarded the grant.

In our analysis, we interpret the overall score given by these two experts as a proxy for the quality of the proposals. We find that the expert evaluations of male and female candidates are very similar. Among the candidate pairs included in our analysis, female candidates receive scores that are, on average, 0.146 points lower than their male counterparts (on a scale from one to twenty), a difference that is not statistically significant ($p = 0.518$). Similarly, in 47.3% of pairs the female candidate is preferred on average (the "expert favourite"), while in 46.0% pairs the male candidate is preferred ($p = 0.841$); the remaining pairs are tied. Using an alternative, more stringent definition of "unanimous favourite" – where a candidate must be strictly preferred by both experts – the female and male candidate are the unanimous favourite in 25.7% and 28.3% of pairs respectively. Interestingly, the individual experts show more signs of potential biases, with one expert rating the male candidates higher on average, and one expert rating female candidates higher on average. Appendix Table A.7 details these statistics.

2 Results

2.1 Result 1: Decision fatigue in the control group

We begin by documenting gender discrimination in the control group: this pattern is central to motivating and interpreting our subsequent treatment effects. In Table 1, we report four probabilities in the control group: (i) the probability that the female candidate wins, (ii) the probability that a candidate favoured by the expert wins, (iii) the probability that a female candidate favoured by the expert wins, and (iv) the probability that a male candidate favoured by the expert wins. We compare each probability between assessments made in the first six rounds and those made in the second six rounds; formally, we do this by estimating the following Linear Probability Model:

$$y_{jp} = \beta_1 \cdot \texttt{Late\_Round}_{jp} + \mu_p + \varepsilon_{jp}, \tag{1}$$

where j indexes judges, p indexes pairs, and $\texttt{Late\_Round}_{jp}$ is a dummy for whether judge j assessed pair p in rounds 7-12. This regression exploits a key feature of our experimental design: the randomisation of the order in which different judges viewed the same candidate pairs. The inclusion of pair fixed effects in equation 1 (μ_p) allows us to identify the causal impact of assessing the *same* pair of candidates in either an early or late assessment round.

Column 1 of Table 1 shows that, in the first six rounds, female candidates win 48.6% of pairwise comparisons in the control group. This falls by 10.2 percentage points – to 38.4% – in the second half of the assessments ($p = 0.042$). Column 3 shows that this decline is driven by expert-favoured female candidates becoming 6.2 percentage points less likely to win: this probability falls from 27.8% in the first six rounds to 21.6% in the last six rounds ($p = 0.035$). Column 4 shows that there is no effect for male candidates favoured by the expert (we have a slight positive coefficient of 3.7 percentage points, with $p = 0.294$).

In sum, Table 1 shows that in the second half of their assessments, judges are substantially less likely to vote for female candidates – and, using the expert assessments as proxies for decision quality, that judges are making worse decisions in doing so. Following literature in organisational psychology and behavioural economics, we refer generically to this result as '*decision fatigue*'. This term refers broadly to a range of phenomena by which, as decision

makers become more tired, they rely more on heuristic 'rules of thumb' – and such rules often reflect underlying biases (Bodenhausen, 1990; Pignatiello, Martin, and Hickman Jr, 2020). Empirically, this phenomenon has been documented in several settings: judicial rulings deteriorate over the course of court sessions (Danziger et al., 2011), financial analysts resort to heuristic forecasts as daily decisions accumulate (Hirshleifer, Levi, Lourie, and Teoh, 2019), employers allocate less attention to resumes from negatively stereotyped groups when attention is scarce (Bartoš et al., 2016), and lab participants make more biased decisions in a shooting scenario when cognitively stressed (Ma, Correll, Wittenbrink, Bar-Anan, Sriram, and Nosek, 2013).

Indeed, in the same context as the present experiment, Abebe et al. (2026) run an experiment evaluating management traits among young professionals. That experiment does not concern gender bias or organisational values. Nonetheless, the evaluation there required Ethiopian HR managers – similar to the subjects in our experiment – to rank three candidates within each of five triplets of young professionals, assessing their capability as a manager or entrepreneur. In Appendix Table A.18, we show that the result of column 1 of Table 1 replicates in that distinct sample: by the fifth round of assessments in that data, a female candidate is 15.2 percentage points less likely to be ranked first in her triplet, relative to a baseline of 36.5% in round one – a decline of 3.4 percentage points per round in a linear specification.[12]

2.2 Result 2: The Organisational Values treatment offsets this decision fatigue

Motivated by Result 1, we now evaluate the impact of our treatments by expanding equation 1 to interact with our three treatments.[13] Table 2 reports the results, using the same four outcomes as Table 1. It shows that the Organisational Values treatment – both on its own and interacted with the Self-Image treatment – offsets completely the decision fatigue described earlier. Specifically, the Organisational Values treatment increases the probability of a woman winning by 11.8 percentage points in the second half, and by 10.7 percentage points the probability that an expert-favoured woman wins. The equivalent coefficients for the interacted treatment are 6.7 percentage points and 5.1 percentage points (the latter also significant).[14]

In Appendix J we investigate heterogeneity in treatment effects by judges' gender. Although we lack power to detect differences by judge gender due to the small number of female judges, the point estimates suggest that the fatigue-driven decline is concentrated among male judges: they become 10 percentage points less likely to vote for expert-favoured female candidates in the last six rounds in the control group, compared to 5 percentage points for female judges.

[12] Three of the authors of this paper were authors of that earlier paper: Girum Abebe, Simon Quinn and Tom Schwantje.

[13] Specifically, let $X = \{\text{Organisational values, Self-Image, Combined treatment}\}$ be the set of treatment arms (omitting control), let t_j be the treatment assigned to judge j, and and denote by $\mathcal{I}$ the indicator function. Then we estimate:

$$y_{jp} = \sum_{x \in X} \beta_x \cdot \mathcal{I}\left(t_j = x\right) + \sum_{x \in X} \gamma_x \cdot \mathcal{I}\left(t_j = x\right) \cdot \texttt{Late\_Round}_{jp} + \delta \cdot \texttt{Late\_Round}_{jp} + \mu_p + \varepsilon_{jp}.$$

[14] Online Appendix Table A.8 shows this result is robust to interacting treatment with a continuous round variable. Online Appendix Figure A.2 plots treatment effects by round. The coefficients on the dummy `Late_Round` in Table 2 differ slightly to the equivalent coefficients in Table 1; this is unsurprising, given the way that pair fixed effects enter the estimation.

2.3 Result 3: The Organisational Values treatment improves decision quality

To this point, we have used the assessments of external human resource experts as a proxy for decision quality. As we noted in our discussion of Table 1 and 2, these assessments were strongly suggestive that the Organisational Values treatment improved decision quality: in Table 1, we found that in the control group agreement with the experts about women fell due to decision fatigue, and in Table 2, we found that this effect was offset by the Organisational Values treatment.

To further understand how the Organisational Values treatment achieved this, we introduce a large vector of candidate characteristics. Broadly, these covariates fall into three categories. First, we have the human assessments (provided by our human resource experts and by trained enumerators) of candidate content and presentation style. Second, we transcribe and translate the audio using a proprietary algorithm provided by Hasab AI. Third, we use the videos themselves to generate 'emotion AI' measures, using a proprietary algorithm provided by the start-up Affectiva.[15] We summarise these covariates in detail in Appendix Table A.1.

To incorporate these candidate characteristics into our analysis, we use two recent complementary machine learning methods. First, we study heterogeneity in treatment effects based on the expected outcome in the control group; to do this, we use the 'endogenous stratification' method of Abadie et al. (2018). Second, we assess which characteristics of candidates predict which female candidates will benefit from the Organisational Values treatment in the last six rounds. To do so, we implement the Generic Machine Learning algorithm of Chernozhukov et al. (2025).

Endogenous Stratification. We implement the method of Abadie et al. (2018) as follows. First, we partition the sample using unique pair identifiers, to ensure that no observations from a given pair are in both the auxiliary or main sample; half of the pairs in the first six rounds are randomly assigned to an auxiliary sample, and the other half of the pairs in the last six rounds are allocated to the main sample. In the auxiliary sample, we use a LASSO-Logit model to estimate the probability – in each pair – that the female candidate will win. We use these model estimates to predict the probability of women winning in the main sample had they been assessed in the first six rounds. We then use a median split on these predicted probabilities – so that we can estimate treatment effects separately for the set of pairs where the woman is less likely to win (the 'low predicted probability' sample) and for the set of pairs where the woman is more likely to win (the 'high predicted probability' sample). Following Abadie et al. (2018), we construct confidence intervals by repeated sample-splitting.

We report results in Table 3. We find clear heterogeneity in the treatment effect of Organisational Values in the last six rounds. Women predicted to perform well in the control group enjoy a substantial increase in their likelihood of winning under treatment (an increase from 49.8 percent to 65.8 percent) – while those predicted to perform poorly exhibit smaller gains (an increase from 34.9 percent to 42.4 percent). We find qualitatively similar patterns of heterogeneity under the interacted treatment, with women predicted to do well in the control group winning 56.0 percent of their assessments, compared to 36.1 percent for those predicted to perform poorly. The Self-image treatment instead does not significantly benefit either subsample, with those predicted to perform well and poorly winning respectively 3.2 and 2.0 percentage points more often. In the first six rounds, none of the treatments

[15] These include measures of how positive (or negative) the emotions displayed by each candidate are, and how engaged the candidate is during the pitch.

significantly benefit any of the subsamples, and we do not find evidence for heterogeneity. We interpret these results as supporting our earlier conclusions: that the Organisational Values treatment stops deterioration in the quality of decisions due to decision fatigue.

Generic Machine Learning. The endogenous stratification exercise shows that, in the later rounds, the Organisational Values treatment is more effective for candidate pairs in which the woman has a high predicted probability of winning. What characterises these candidate pairs? To answer this question, we implement the generic machine learning framework of Chernozhukov et al. (2025) to study heterogeneity in treatment effects. Appendix H.3 describes the procedure in detail.

In summary, we restrict attention to the last six rounds – where the reduced-form treatment effects are concentrated. To test the impact of the Organisational Values treatment, we pool the Organisational Values and the Combined treatments and compare to Control (omitting Self-image for this exercise for simplicity). In each repetition, we split the sample at the pair level into auxiliary and main halves. On the auxiliary half, we estimate separate elastic-net logistic models for treated and control observations using the same rich set of pair-level and judge-level observables as in the endogenous-stratification exercise. We then use these models to construct – for each observation in the main half – a proxy for the conditional average treatment effect equal to the difference between the predicted outcome under treatment and under control.

We split the sample into two groups based on a median split of the predicted treatment effect. We show that in the group with a high predicted treatment effect, women indeed benefit more: they win 10.4 percentage points more due to the treatment, compared to 4.3 percentage points for the group with a low predicted treatment effect. We can then compare these two groups in terms of their characteristics, we report the results in Table 4. Strikingly, in pairs with a high predicted treatment effect, experts see the women as having a better business concept, better sense of the market, and a better business sense, and to more frequently mention their business operations. They are also more articulate, and give more specific examples. However, they are less likely to be seen as arrogant by the enumerators, and their appearance is regarded as less appropriate for the competition. Turning to some measures of 'gendered' behaviour, we find that in pairs with a high predicted treatment effect, women are more likely to mention their family or to discuss teamwork, but use *less* communal language – words referring to a larger collective like 'we'.

3 Discussion

Our results show that inclusive institutional decision-making can deteriorate over the course of an evaluation session, even when there is little evidence of bias at the outset. In the control group, judges are substantially less likely to select female candidates in later assessments, and this decline is concentrated among female candidates whom external HR experts judge to be stronger. We interpret this pattern as evidence that decision fatigue reduces both inclusivity and decision quality: as evaluators become cognitively depleted, they are more likely to disadvantage strong female candidates. By contrast, a simple prompt emphasising the organisation's commitment to equal opportunity fully offsets this deterioration. The self-image treatment has directionally similar but more muted effects. Taken together, these findings suggest that low-cost features of organisational design can improve inclusivity and meritocracy

simultaneously, not by constraining discretion, but by helping evaluators sustain attention to the organisation's objectives

To assess whether the treatments affect whether judges identify promising candidates, we collect follow-up data on candidates' labour market outcomes approximately 18 months after the competition. In line with the existing literature on business plan competitions (Fafchamps and Woodruff, 2017; McKenzie and Sansone, 2019), judges' assessments have limited predictive power for candidates' subsequent employment, income, and business investment – and this holds regardless of whether the assessments occurred in early or late rounds. Since only ten candidates won prizes, we estimate these correlations across the full sample, controlling for winning. A caveat is that judges may select candidates whom they expect would benefit most from winning – rather than those with the strongest unconditional prospects – in which case limited predictive power for long-run outcomes may not imply poor selection. We provide additional details in Appendix K.

Our results point to a broader implication for organisations. Much of the public discussion around inclusive decision-making is framed as a trade-off between fairness and performance. In our setting, that framing is misleading. The same institutional feature that improved outcomes for female candidates also improved alignment with expert assessments. This suggests that some forms of observed exclusion may reflect not a deliberate preference against under-represented groups, but a failure of institutional design to support careful evaluation under realistic conditions of limited attention. Crucially, the classic Beckerian argument that competition erodes discrimination (Becker, 1957; Falco, Gatti, Islam, and Menzel, 2025) applies to taste-based bias, but the fatigue-driven bias we document imposes costs that firms may not recognise – and that competition alone will not correct – without deliberate institutional design. For organisations, our results are encouraging: simple, scalable interventions – such as reminders of organisational values, perhaps combined with better management of evaluator workload – may improve decision-making without removing discretion or imposing rigid rules.

Several questions remain for future research. One is external validity across other evaluative environments, such as hiring, lending, promotion, and admissions, where evaluators face long sequences of subjective judgments. A second is whether stronger accountability mechanisms can make peer justification more effective. More broadly, social incentives in organisations may improve or impair performance depending on how they are designed; understanding which forms of peer accountability sustain evaluative effort is an open question (see Ashraf and Bandiera (2018) for an extensive discussion). A third is whether these effects extend beyond gender to other dimensions of disadvantage. More generally, our findings suggest that understanding discrimination requires attention not only to who decision-makers are, but also to the organisational environments in which they decide. Small changes in those environments can materially affect both whom institutions select and how well they select them.

Table 1: Gender Bias in the Control Group

	(1) Woman Wins	(2) Expert-favoured Candidate Wins	(3) Expert-favoured Woman Wins	(4) Expert-favoured Man Wins
Late Round	-0.102**	-0.025	-0.062**	0.037
	(0.049)	(0.044)	(0.029)	(0.035)
Mean (early rounds)	0.486	0.576	0.278	0.299
N	567	567	567	567
Pair FE	Yes	Yes	Yes	Yes

Notes: This table reports OLS estimates of late-round effects on judge decisions using only the control group, regressing each outcome on a Late Round indicator (rounds 7–12). The dependent variables are: (1) woman wins, (2) expert-favoured candidate wins, (3) expert-favoured woman wins ($= 1$ if the female candidate is both expert-favoured and wins); and (4) expert-favoured man wins ($= 1$ if the male candidate is both expert-favoured and wins). The omitted category is Early Rounds (rounds 1–6). All specifications include pair fixed effects. Standard errors clustered at the judge level are in parentheses. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table 2: The treatment effect in early versus late rounds

	(1) Woman Wins	(2) Expert-favoured Candidate Wins	(3) Expert-favoured Woman Wins	(4) Expert-favoured Man Wins
Late Round	-0.105**	-0.059	-0.084***	0.025
	(0.045)	(0.038)	(0.029)	(0.031)
Org. Values × Late Round	0.137**	0.110**	0.124***	-0.014
	(0.057)	(0.050)	(0.039)	(0.038)
Self-Image × Late Round	0.055	0.065	0.063*	0.001
	(0.054)	(0.049)	(0.036)	(0.037)
Org. Values & Self-Image ×	0.101*	0.073	0.086**	-0.014
Late Round	(0.057)	(0.052)	(0.038)	(0.040)
Org. Values	-0.019	-0.021	-0.017	-0.004
	(0.043)	(0.036)	(0.027)	(0.028)
Self-Image	-0.005	-0.020	-0.014	-0.006
	(0.040)	(0.039)	(0.026)	(0.029)
Org. Values & Self-Image	-0.033	-0.037	-0.035	-0.003
	(0.042)	(0.037)	(0.026)	(0.030)
Pair FE	Yes	Yes	Yes	Yes
Control mean (early rounds)	0.486	0.576	0.278	0.299
p: Org. Values (late)	0.005	0.031	<0.001	0.515
p: Self-Image (late)	0.206	0.256	0.078	0.860
p: Combined (late)	0.114	0.402	0.086	0.575
N	2,649	2,649	2,649	2,649

Notes: This table reports OLS estimates of treatment effects on individuals' decision outcomes, interacted with evaluation timing. The dependent variable in column (1) is an indicator equal to one if the female candidate is selected; column (2) indicates whether the expert-favoured candidate wins; columns (3) and (4) split this by whether the expert-favoured the woman or man wins, respectively. Treatment indicators—Org. Values, Self-image, and their combination—are each interacted with *Late round*, an indicator for rounds 7–12 (late rounds). The omitted category is Control × early rounds. p-values at the bottom test whether the total treatment effect in late rounds (main + interaction) equals zero. "Control mean (early rounds)" reports the mean of the dependent variable for the control group in rounds 1–6. All specifications include candidate-pair fixed effects. Standard errors clustered at the judge level are reported in parentheses. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table 3: Heterogeneity in treatment effects on female candidates' performance

	Control	Org. Values	Self-Image	Both
Panel A: All rounds				
	Conditional Average Treatment Effect			
Low predicted perf.		3.8%	1.0%	-0.9%
		$[-1.2, 8.7]$	$[-4.2, 6.2]$	$[-5.7, 4.1]$
High predicted perf.		10.8%	3.6%	5.2%
		$[5.9, 15.3]$	$[-1.4, 8.7]$	$[0.2, 9.9]$
	Group Mean by Treatment			
Low predicted perf.	37.5%	41.4%	38.5%	36.6%
High predicted perf.	51.6%	62.3%	55.1%	56.8%
Panel B: Early rounds (complementary half)				
	Conditional Average Treatment Effect			
	Control	Org. Values	Self-Image	Both
Low predicted perf.		-2.8%	-0.4%	-4.5%
		$[-12.4, 6.9]$	$[-11.0, 10.0]$	$[-14.7, 6.1]$
High predicted perf.		0.1%	3.9%	3.2%
		$[-10.1, 10.9]$	$[-6.5, 15.4]$	$[-8.5, 14.6]$
	Group Mean by Treatment			
Low predicted perf.	42.2%	39.5%	41.8%	37.7%
High predicted perf.	55.4%	55.5%	59.3%	58.6%
Panel C: Late rounds				
	Conditional Average Treatment Effect			
	Control	Org. Values	Self-Image	Both
Low predicted perf.		7.5%	2.0%	1.2%
		$[2.9, 12.4]$	$[-2.9, 7.5]$	$[-3.4, 5.9]$
High predicted perf.		16.0%	3.2%	6.2%
		$[11.0, 20.4]$	$[-2.2, 8.0]$	$[2.0, 10.6]$
	Group Mean by Treatment			
Low predicted perf.	34.9%	42.4%	36.9%	36.1%
High predicted perf.	49.8%	65.8%	52.9%	56.0%

Notes: This table reports results from heterogeneity analysis using the methodology of Abadie et al. (2018). Panel A shows all rounds, Panel B shows early rounds (1-6), Panel C shows late rounds (7-12). Within each panel, the top rows show conditional average treatment effects with 90% confidence intervals for candidates with low vs. high predicted probability of winning in the control group. The bottom rows show outcome levels by treatment arm.

Table 4: CLAN: Characteristics of most and least affected groups

	G1	G2	Diff
Concept (expert)	-0.12	0.20	0.33**
Market (expert)	-0.14	0.13	0.27**
Business sense (expert)	-0.10	0.17	0.27**
Communal language	0.11	-0.18	-0.30**
Appearance (expert)	0.31	-0.15	-0.46**
Mentions operations	-0.22	0.11	0.33**
General appearance (enum.)	0.15	-0.09	-0.25**
Arrogance (enum.)	0.18	-0.19	-0.37**
Teamwork (enum.)	-0.02	0.20	0.22**
Mentions family (enum.)	-0.14	0.08	0.22**
Certainty (enum.)	-0.04	0.17	0.22**
Articulate	-0.12	0.32	0.44**
Specific examples	-0.17	0.21	0.38**

Notes: This table reports mean covariate values within the least-affected (G1) and most-affected (G2) GATES groups from the Chernozhukov et al. (2025) classification analysis (CLAN). The table only reports the covariates that are significantly different between the two groups at the 90% level; Appendix Table A.23 reports all covariates. All covariates are standardised differences between the female and male candidate in the pair, with positive values indicating that the female candidate scores higher. The treated group pools treatment arms 2 (organisational values) and 4 (both treatments); the control group is treatment arm 1. The estimation sample is restricted to late rounds (rounds 7–12). Point estimates are medians across 250 random sample splits. ** denotes that the 90% confidence interval for the G2−G1 difference excludes zero.

References

Abadie, A., M. M. Chingos, and M. R. West (2018). Endogenous stratification in randomized experiments. *Review of Economics and Statistics 100*(4), 567–580.

Abebe, G., M. Fafchamps, M. Koelle, and S. Quinn (2026). Matching, management and employment outcomes: A field experiment with firm internships. *CEPR Discussion Paper 21194*.

Abebe, G., M. Fafchamps, M. Koelle, S. Quinn, and T. Schwantje (2026). Management Style Under the Spotlight: Evidence from Studio Recordings. *CEPR Discussion Paper 21043*.

Aghion, P. and J. Tirole (1997). Formal and real authority in organizations. *Journal of Political Economy 105*(1), 1–29.

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2024). Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review 114*(7), 1916–1948.

Alesina, A. F., F. Lotti, and P. E. Mistrulli (2013, 01). Do Women Pay More for Credit? Evidence From Italy. *Journal of the European Economic Association 11*, 45–66.

Ashraf, N. and O. Bandiera (2018). Social incentives in organizations. *Annual Review of Economics 10*(1), 439–463.

Ashraf, N., O. Bandiera, and B. K. Jack (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics 120*, 1–17.

Augenblick, N. and M. Rabin (2019). An experiment on time preference and misprediction in unpleasant tasks. *Review of Economic Studies 86*(3), 941–975.

Ayalew, S., S. Manian, and K. Sheth (2023). Discrimination and access to capital: Experimental evidence from ethiopia. Working Paper WPS-236, Center for Effective Global Action, University of California, Berkeley.

Bartoš, V., M. Bauer, J. Chytilová, and F. Matějka (2016). Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition. *The American Economic Review 106*(6), 1437–1475.

Bartoš, V., S. Castro, K. Czura, and T. Opitz (2024). Gendered access to finance: The roles of team formation, idea quality, and implementation constraints in business evaluations. CESifo Working Paper Series 11205, CESifo.

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

Benson, A., D. Li, and K. Shue (2026). "Potential" and the gender promotion gap. *American Economic Review 116*(2), 375–417.

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

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

Bohnet, I., A. Van Geen, and M. Bazerman (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science 62*(5), 1225–1234.

Brock, J. M. (2025). Professional motivation and the quantity–quality trade-off. *Journal of Comparative Economics 53*(3), 754–771.

Brock, J. M. and R. De Haas (2023). Discriminatory lending: Evidence from bankers in the lab. *American Economic Journal: Applied Economics 15*(2), 31–68.

Buehren, N. and S. Papineni (2025, Jun). Gender discrimination in entrepreneurial finance : Experimental evidence from ethiopia. Policy Research Working Paper Series 11152, The World Bank.

Bursztyn, L., S. Fiorin, D. Gottlieb, and M. Kanz (2019). Moral incentives in credit card debt repayment: Evidence from a field experiment. *Journal of Political Economy 127*(4), 1641–1683.

Cettolin, E., K. Cole, and P. Dalton (2024). Improving workers' performance in small firms: A randomized experiment on goal setting in ghana. *Review of Economics and Statistics*.

Chen, D. L., T. J. Moskowitz, and K. Shue (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics 131*(3), 1181–1242.

Chen, H., Y. Chen, and Q. Yang (2025). Women in the courtroom: Technology and justice. *Review of Economic Studies*, rdaf066.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2025). Fisher–schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. *Econometrica 93*(4), 1121–1164.

Chugh, D. (2004). Societal and managerial implications of implicit social cognition: Why milliseconds matter. *Social Justice Research 17*(2), 203–222.

Danziger, S., J. Levav, and L. Avnaim-Pesso (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences 108*(17), 6889–6892.

Dessein, W. (2002). Authority and communication in organizations. *The Review of Economic Studies 69*(4), 811–838.

Ewens, M. (2023). Gender and race in entrepreneurial finance. In *Handbook of the Economics of Corporate Finance*, Volume 1, pp. 239–296. Elsevier.

Fafchamps, M. and S. Quinn (2018). Networks and manufacturing firms in Africa: Results from a randomized field experiment. *The World Bank Economic Review 32*(3), 656–675.

Fafchamps, M. and C. Woodruff (2017). Identifying gazelles: Expert panels vs. surveys as a means to identify firms with rapid growth potential. *The World Bank Economic Review 31*(3), 670–686.

Falco, P., R. Gatti, A. Islam, and A. Menzel (2025, September). Does competition reduce hiring discrimination? experimental evidence from egypt. Preliminary draft.

Glöckner, A. (2016). The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision making 11*(6), 601–610.

Goldin, C. and C. Rouse (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *The American Economic Review 90*(4), 715–741.

Hebert, C. (2025). Gender stereotypes and entrepreneur financing. *The Review of Financial Studies*, hhaf049.

Hirshleifer, D., Y. Levi, B. Lourie, and S. H. Teoh (2019). Decision fatigue and heuristic analyst forecasts. *Journal of Financial Economics 133*(1), 83–98.

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

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

Khan, M. Y. (2025). Mission motivation and public sector performance: experimental evidence from pakistan. *American Economic Review 115*(7), 2343–2375.

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

McKenzie, D. and D. Sansone (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in nigeria. *Journal of Development Economics 141*, 102369.

Miller, A., S. A. Lall, M. Goldstein, and J. Montalvao (2023, 12). Asking better questions: The effect of changing investment organizations' evaluation practices on gender disparities in funding innovation. Policy Research Working Paper 10625, World Bank, Washington, DC.

Miserocchi, F. (2024). Discrimination through biased memory. *Working Paper*.

Pignatiello, G. A., R. J. Martin, and R. L. Hickman Jr (2020). Decision fatigue: A conceptual analysis. *Journal of health psychology 25*(1), 123–135.

Radbruch, J. and A. Schiprowski (2023, May). Committee Deliberation and Gender Differences in Influence. ECONtribute Discussion Papers Series 234, University of Bonn and University of Cologne, Germany.

Salon (2025, February). "Our Strength Comes from Hiring the Best People": Apple Rejects Push to End DEI Initiatives. *Salon*.

Wang, A. (2024, June). Scale is a meritocracy, and we must always remain one. Scale AI Blog.

Weinshall-Margel, K. and J. Shapard (2011). Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences 108*(42), E833–E833.

Wigboldus, D. H., J. W. Sherman, H. L. Franzese, and A. v. Knippenberg (2004). Capacity and comprehension: Spontaneous stereotyping under cognitive load. *Social Cognition 22*(3), 292–309.

World Bank (2022). Ethiopia Gender Diagnostic: Building the Evidence Base to Address Gender Inequality in Ethiopia. Technical report, World Bank, Africa Gender Innovation Lab.