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

Appendix for online publication

May 5, 2026

Contents

A Model **4**
- A.1 Probability of disagreement and geometric intuition . . . 4
- A.2 Evaluator problem and first-order condition . . . 5
- A.3 Equilibrium . . . 5
- A.4 Proof of Proposition 1 . . . 6
- A.5 Interpretation . . . 7

B Variables **8**

C Results Pre-Analysis Plan **10**
- C.1 The effect of the treatments on the probability of dissent . . . 11
- C.2 The effect of committee composition on the probability of dissent . . . 12

D Details experiment **14**
- D.1 Protocol for the control group . . . 14
- D.2 Further details on the Social Image Treatment . . . 15
- D.3 Randomization . . . 15
- D.4 Invitation judges . . . 17
- D.5 Invitation candidates . . . 17
- D.6 Summary statistics judges . . . 19
- D.7 Summary statistics expert rankings . . . 19

E Additional Results **21**
- E.1 Appendix Table and Figure: Continuous rounds . . . 21

F Robustness to the choice of sample **24**
- F.1 Main treatment effects across subsamples . . . 24
- F.2 Late-round heterogeneity across subsamples . . . 24
- F.3 Sensitivity to expert favourite definition . . . 29
- F.4 Pre-specified factorial specification across subsamples . . . 31

G External Validation: Gender Bias in Candidate Evaluation **32**

H Machine Learning Procedures **37**
- H.1 Variables and variable construction . . . 37
- H.2 Abadie-style judge-level heterogeneity procedure . . . 38
- H.3 Chernozhukov-style generic machine learning procedure . . . 40
- H.4 Machine learning results . . . 41
  - H.4.1 Abadie-style heterogeneity results . . . 41
  - H.4.2 Chernozhukov-style heterogeneity results . . . 41
  - H.4.3 Tercile-split results . . . 42

I Sequential Decision Patterns **47**
I.1 Robustness: candidate presentation order 50

J Treatment Effect Heterogeneity by Judge Gender **52**

K Long-term labour market outcomes **55**

A Model

In this appendix section, we present a stylised theoretical framework to capture the basic trade-offs that organisations face when they design assessment processes. On the one hand, the organisation wants decision-makers to use their discretion and good judgement in assessing candidates; on the other hand, the organisation cares about some dimensions in particular, and so does not want decision-makers to rely purely on their own personal and idiosyncratic preferences. This tension is a long-standing theme in organisational economics (for example, Dessein, 2002; Prendergast, 1993).

To capture this tension in our setting – and to consider potential organisational design options – we imagine different judges assessing candidates. For simplicity, imagine two judges (i and j). Each judge ℓ is characterised by a preferred evaluation rule $\omega_\ell \in [0, \pi/2]$, and chooses an evaluation rule $\theta_\ell \in [0, \pi/2]$.

Evaluators face three forces. First, they prefer to align their evaluation with their own view of what matters. Second, they face organisational pressure to emphasize particular dimensions. Third, they incur a cost from disagreeing with other evaluators (Bursztyn, Egorov, Haaland, Rao, & Roth, 2023). We capture these forces with the expected loss function:

$$L(\theta_i; \omega_i) = \underbrace{\frac{\alpha}{2}(\theta_i - \omega_i)^2}_{\text{personal preference}} + \underbrace{\beta\left(\frac{\pi}{2} - \theta_i\right)}_{\text{organizational alignment}} + \underbrace{\gamma\, \mathbb{E}_{\omega_j}[\,|\theta_i - \theta(\omega_j)|\,]}_{\text{dissent aversion}}. \tag{A.1}$$

where $\alpha > 0$ governs the cost of deviating from one's own preferred evaluation, β captures organizational pressure toward the organizational benchmark $\pi/2$, and $\gamma > 0$ captures aversion to disagreement.

A.1 Probability of disagreement and geometric intuition

To motivate the disagreement term, it is helpful to consider a geometric interpretation. Suppose candidates differ along two dimensions, (x, y), where y captures the dimension valued by the organisation and x captures an idiosyncratic or stylistic trait. For simplicity, assume (x_k, y_k) are drawn independently from a bivariate normal distribution with mean zero and equal variance across dimensions. The key property we rely on is the radial symmetry of the joint distribution.

Each judge i chooses an evaluation angle $\theta_\ell \in [0, \pi/2]$, which defines an evaluation vector $\mathbf{v}_\ell = (\cos\theta_\ell, \sin\theta_\ell)$. The score assigned to candidate k by judge ℓ is $s_\ell(k) = \cos(\theta_\ell) \cdot x_k + \sin(\theta_\ell) \cdot y_k$, and the judge votes for candidate 1 if and only if $s_\ell(1) \geq s_\ell(2)$.

Under symmetry, disagreement between two evaluators arises when their decision rules rank candidates differently. In this setting, the probability of disagreement is proportional to the angular distance between decision rules, as formalised in the following lemma.

Lemma 1. *For any evaluation angles θ_i, θ_j,*

$$\Pr(\text{disagree}) \;=\; \frac{|\theta_i - \theta_j|}{\pi}.$$

Proof. Judge ℓ votes for candidate 1 iff $X_\ell := \cos(\theta_\ell)\, D_x + \sin(\theta_\ell)\, D_y > 0$, where $D_x = x_1 - x_2$ and $D_y = y_1 - y_2$ are independent $\mathcal{N}(0, 2)$ random variables. The pair (X_i, X_j) is bivariate normal with zero means and correlation $\rho = \cos(\theta_i - \theta_j)$. The radial symmetry of (D_x, D_y) implies that the disagreement region consists of two opposing wedges of opening angle $|\theta_i - \theta_j|$. ∎

A.2 Evaluator problem and first-order condition

Evaluator i chooses θ_i to minimise expected loss as defined above. Assuming an interior solution and differentiability almost everywhere, the first-order condition is:

$$\alpha(\theta_i - \omega_i) - \beta + \gamma\, \mathbb{E}_{\omega_j}\left[\text{sgn}(\theta_i - \theta(\omega_j))\right] = 0. \tag{A.2}$$

We look for a symmetric equilibrium in which evaluators adopt a monotone strategy $\theta(\omega)$. Depending on parameter values, corner solutions or bunching at the boundary may arise. We focus initially on the interior linear equilibrium for expositional clarity. (We also focus on parameter values such that $B > 0$, ensuring that the equilibrium is monotone in type. When $B < 0$, the equilibrium is non-monotone and unintuitive to interpret, so we do not consider this case.)

Under monotonicity of $\theta(\cdot)$, we have

$$\text{sgn}(\theta_i - \theta(\omega_j)) = \text{sgn}(\omega_i - \omega_j),$$

so the expectation depends only on the distribution of ω_j. This reflects the fact that, under monotone strategies, disagreement depends only on the relative ranking of types.

With $\omega \sim U[0, \pi/2]$, this expectation simplifies to:

$$\mathbb{E}_{\omega_j}\left[\text{sgn}(\omega_i - \omega_j)\right] = \frac{4}{\pi}\left(\omega_i - \frac{\pi}{4}\right). \tag{A.3}$$

A.3 Equilibrium

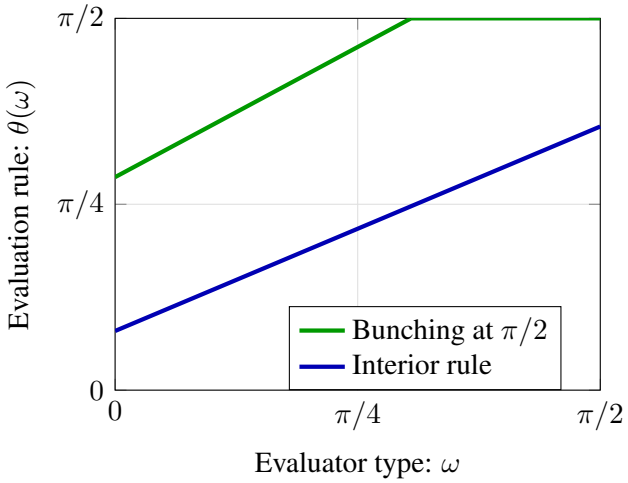
In equilibrium, evaluators trade off these forces when choosing their evaluation rule. We focus on symmetric equilibria in which strategies are monotone in type. The following proposition characterises the equilibrium.

Proposition 1. *There exists a symmetric equilibrium in which evaluators adopt a linear decision rule*

$$\theta(\omega) = A + B\omega,$$

for interior types, with possible bunching at the boundary. The coefficients are given by

$$B = 1 - \frac{4\gamma}{\pi\alpha}, \quad A = \frac{\beta}{\alpha} + (1 - B)\frac{\pi}{4}.$$



Note: The figure illustrates the equilibrium mapping $\theta(\omega)$. The interior rule is linear. When organisational pressure is sufficiently strong, the rule is truncated at the organisational benchmark $\pi/2$, generating bunching and a high-alignment outcome. In this region, further increases in β do not affect behaviour.

Figure A.1: Equilibrium evaluation rules

The linear rule above describes interior behaviour. However, the choice set is bounded: $\theta \in [0, \pi/2]$. For sufficiently high values of β relative to α and γ, the linear rule may imply $\theta(\omega) \geq \pi/2$ for a range of types. In this case, evaluators bunch at the boundary $\theta = \pi/2$, fully implementing the organisational criterion. Formally, the equilibrium takes the truncated form

$$\theta(\omega) = \min\left\{\frac{\pi}{2},\ A + B\omega\right\}.$$

This mapping is illustrated in Figure A.1.

A.4 Proof of Proposition 1

Substituting into the first-order condition:

$$\alpha(\theta_i - \omega_i) - \beta + \gamma \cdot \frac{4}{\pi}\left(\omega_i - \frac{\pi}{4}\right) = 0. \tag{A.4}$$

Rearranging:

$$\theta_i = \omega_i + \frac{\beta}{\alpha} - \frac{4\gamma}{\pi\alpha}\left(\omega_i - \frac{\pi}{4}\right). \tag{A.5}$$

Collecting terms:

$$\theta_i = \frac{\beta}{\alpha} + \frac{4\gamma}{\pi\alpha} \cdot \frac{\pi}{4} + \left(1 - \frac{4\gamma}{\pi\alpha}\right)\omega_i. \tag{A.6}$$

Simplifying:

$$\theta_i = A + B\omega_i, \tag{A.7}$$

where

$$B = 1 - \frac{4\gamma}{\pi\alpha}, \quad A = \frac{\beta}{\alpha} + (1 - B)\frac{\pi}{4}. \tag{A.8}$$

This establishes the linear equilibrium stated in Proposition 1.

A.5 Interpretation

The interior equilibrium can be interpreted as a weighted average. Evaluators choose a rule that combines their own preferred weighting ω_i with a common benchmark:

$$\theta(\omega_i) = \frac{\beta}{\alpha} + (1 - B)\frac{\pi}{4} + B\omega_i,$$

so that B governs the weight placed on individual preferences ω_i, while $1 - B$ captures the extent of convergence toward a shared benchmark (centered at $\pi/4$).

The parameters α, β, and γ map directly into this structure. Higher α increases reliance on individual preferences (raising B), higher β shifts evaluations toward the organisational benchmark (raising A), and higher γ compresses variation in evaluations by reducing B and inducing convergence across evaluators. Notably, β operates as a pure level shift, while both α and γ also affect dispersion and therefore indirectly influence the level of evaluations. These forces map closely to the experimental design: fatigue can be interpreted as increasing the effective weight on evaluators' own preferred criteria, which in the model corresponds to a higher effective α; organisational messages affect β; and features that increase the salience or cost of disagreement between evaluators affect γ.

The model can be interpreted as implying differing levels of alignment across evaluators. When organisational pressure is sufficiently strong relative to individual preferences, evaluators bunch at the boundary $\theta = \pi/2$, generating a high-alignment (pooling-like) outcome with low dispersion. In such cases, organisational criteria are already effectively implemented. This provides a natural interpretation of why the organisational-values treatment has little effect in early rounds of the experiment. Over time, decision fatigue can be interpreted as increasing the relative cost of deviating from the personal bliss point: α increases, judges rely more on their own preferences, and alignment declines. But an increase in organisational alignment – for example, through clearer emphasis to judges on organisational objectives – can generate a countervailing increase in β, sustaining a more aligned equilibrium. Heterogeneity in responses arises naturally in intermediate cases where some evaluators remain close to the organisational benchmark while others do not. In addition, the effect of disagreement aversion depends on the level of organisational alignment: when alignment is high, social pressure reinforces organisational objectives, while when alignment is weaker, judges may instead converge toward a common rule, reflecting social pressure to agree.

B Variables

Pair Level		
ASPIRE	**Expert**	**Comment**
Score for business concept	Likert 1-5	
Score for understanding market	Likert 1-5	
Score for growth strategy	Likert 1-5	
Score for presentation	Likert 1-5	
Score for business plan	Likert 1-5	
Score for business sense	Likert 1-5	
Overall score	Discrete 1-20	Not included for ML algorithms
Required Content Included	**Expert & Enumerator**	
Introduces themselves	Binary*	Not included for ML algorithms
Mentions business idea	Binary**	
Mentions target market	Binary	
Mentions competition	Binary	
Mentions operations	Binary	
Mentions business costs	Binary	
Other Expert Questions	**Expert**	
Is appearance appropriate	Likert 1-4	
Is the competitor confident	Likert 1-4	
Does the competitor seem arrogant	Likert 1-4	
How certain or convincing?	Likert 1-4	
Planned location of firm	Indicator	Not included for ML algorithms
Start or expand firm	Binary	Not included for ML algorithms
Other Enumerator Questions	**Enumerator**	
Ethnicity	Indicator	
Gives specific examples	Binary	
Is the competitor dressed formally	Likert 1-4	
Did the competitor take care of their outfit	Likert 1-4	
The competitor's general appearance is appropriate	Likert 1-4	
Is the competitor confident	Likert 1-4	
Does the competitor seem arrogant	Likert 1-4	
Does the competitor give examples of teamwork	Indicator	
Does the competitor mention their family	Indicator	
Does the competitor thank others	Indicator	
How old do you think the competitor is?	Discrete	
Emotions	**Proprietary Algorithm**	
Sum of positive and negative emotions	0-100	
Sum of absolute positive and negative emotions	0-100	
Duration of the submission	Discrete	

* No variation, ** Almost no variation.

Table A.1: Observable Characteristics of Submissions

Judge Level		
Judge Characteristics		
Hostile sexism (HS)	Continuous	
Benevolent sexism (BS)	Continuous	
Judge age	Discrete	
Judge gender	Binary	
Years of experience	Discrete	
Industry of firm	Indicator	
Position at firm	Indicator	
Questions About Experiment		
Extent to which judge cares about feedback committee (careFeedback)	Likert	
Extent to which judge considers votes of others (considerOthers)	Likert	
Select up to 5 judges important for perception (importanceOthers)	Indicator	
Importance of Characteristics for Proposal Quality		
Good concept (Concept)	Likert	
Understanding market (Market)	Likert	
Strategy for growth (Growth)	Likert	
Financial plan (Finance)	Likert	
Business plan (Plan)	Likert	
Presentation skill (Present)	Likert	
Business sense (Sense)	Likert	
Ninety-seven judges listed all these characteristics as 'Strongly agree' when asked about their importance..		

Table A.2: Judge-Level Characteristics

C Results Pre-Analysis Plan

We examine the effect of the three treatment arms using the following regression:

$$\text{WomanWins}_{jcp} = \alpha + \beta_1 \text{Info}_j + \beta_2 \text{Comm}_j + \beta_3 \text{Info}_j \cdot \text{Comm}_j + \mu_p + \varepsilon_{jcp} \tag{C.9}$$

Stata: `reghdfe womanWins info comm infoComm, absorb(pair) vce(cluster judge)`

The variables are defined as follows:

- WomanWins_{jcp}: Judge j on committee c votes for the woman in pair p.
- Info_j: Judge j has the information treatment (set to 1 for treatments 1 and 3).
- Comm_j: Judge j is on a committee (set to 1 for treatments 2 and 3).
- μ_p: Pair fixed effects.
- ε_{jp}: Standard errors are clustered at the judge level.

For our secondary hypothesis we do subgroup analysis by gender. We look at the treatment effect for male and female judges separately using the specification in equation C.9. We run regression C.9 using OLS on the sub-sample of male and female judges and then compare the coefficients testing the same set of hypotheses for the comparison of the two sets of parameters.

$$\text{WomanWins}_{jcp} = \alpha + \beta_1 \text{Info}_j + \beta_2 \text{Comm}_j + \beta_3 \text{Info}_j \cdot \text{Comm}_j + \mu_p + \varepsilon_{jcp} \tag{C.10}$$

Stata: `reghdfe womanWins info comm infoComm if gender == 1, absorb(pair) vce(cluster judge)`

$$\text{WomanWins}_{jcp} = \alpha + \beta_1 \text{Info}_j + \beta_2 \text{Comm}_j + \beta_3 \text{Info}_j \cdot \text{Comm}_j + \mu_p + \varepsilon_{jcp} \tag{C.11}$$

Stata: `reghdfe womanWins info comm infoComm if gender == 2, absorb(pair) vce(cluster judge)`

$$\text{WomanWins}_{jcp} = \alpha + \beta_1 \text{Info}_j + \beta_2 \text{Woman}_c + \beta_3 \text{Info}_j \cdot \text{Woman}_c + \mu_p + \varepsilon_{jp}, \tag{C.12}$$

where $\text{Woman}_c = 1$ if there is a woman on the committee.

Stata: `reghdfe womanWins info woman infoWoman if comm==1 & gender ==1, absorb(pair) vce(cluster judge)`[1]
The results from these regressions are presented in Table A.3.

	All judges b/se	Male judges b/se	Female judges b/se	Composition b/se
Org. Values	0.049	0.055	0.011	-0.025
	(0.03)	(0.04)	(0.06)	(0.05)
Social Image	0.022	0.032	-0.033	
	(0.03)	(0.04)	(0.06)	
Org. Values × Social Image	-0.054	-0.053	0.004	
	(0.04)	(0.05)	(0.09)	
Committee has at least one female member=1				-0.046
				(0.05)
Org. Values × Committee has at least one female member=1				0.058
				(0.08)
Pair FE	Yes	Yes	Yes	Yes
Control mean	0.455	0.453	0.469	0.497
N	2649	1978	630	1443

Notes: This table implements Equation C.9 and C.12. Standard errors are clustered at the judge level and the regression includes pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.3: The effects on the probability of a female candidate winning

C.1 The effect of the treatments on the probability of dissent

Finally, we examine the outcomes at a committee level, examining the probability of a unanimous decision for each of the treatment arms. We first run the same regression as in equation C.9, but now at a committee level with as dependent variable whether or not the committee's decisions was unanimous.

We then run the same regressions with as outcome variables whether or not the committee's decision was unanimously for a male candidate, and whether or not the committee's decision was unanimously for a female candidate. These regressions help us examine the channels through which the treatments affect judges' decisions. (For example, a high value for the committee treatment in the regression with as dependent variable UnanimousWomanWins would indicate judges in particular do not want to be found to be the only one dissenting by voting for the male candidate when the other judges' are expected to vote for the female candidate.)

We run the following three regressions:

[1] The condition & gender == 1 was not pre-specified erroneously. This means we focus on the effect of having a female committee member on male committee members.

$$\text{Unanimous}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \tag{C.13}$$

Stata: `reghdfe unanimous info comm infoComm, absorb(pair) vce(cluster committee)`

$$\text{UnanimousWomanWins}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \tag{C.14}$$

Stata: `reghdfe unanimousWomanWins info comm infoComm, absorb(pair) vce(cluster committee)`

$$\text{UnanimousManWins}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \tag{C.15}$$

Stata: `reghdfe unanimousManWins info comm infoComm, absorb(pair) vce(cluster committee)`

	Unanimous b/se	Unan. Female b/se	Unan. Male b/se
Org. Values	0.057	0.062*	-0.004
	(0.05)	(0.04)	(0.04)
Social Image	0.081*	0.060*	0.021
	(0.05)	(0.03)	(0.04)
Org. Values × Social Image	-0.078	-0.082*	0.003
	(0.07)	(0.05)	(0.05)
Pair FE	Yes	Yes	Yes
Control mean	0.318	0.129	0.190
N	871	871	871

Notes This table implements Equation C.13. It shows the effect of the different treatment conditions on the probability of unanimity among the grouped triplets of judges who – within each treatment arm – assessed the same pair of candidates. We show effects (i) on unanimity, (ii) on unanimity in favour of the male candidate, and (iii) on unanimity in favour of the female candidate. Standard errors are clustered at the committee level and the regression includes pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.4: The effect of the treatments on unanimity

C.2 The effect of committee composition on the probability of dissent

Finally, for the subsample of judges on committees, we test whether committee composition affects the probability of dissent at a committee level. To do this, we use the same dependent variables as in regression C.12, but now use as a dependent variable the probability of a unanimous decision.

$$\text{Unanimous}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Woman}_c + \beta_3 \text{Info}_c \cdot \text{Woman}_c + \mu_p + \varepsilon_{cp} \quad \text{(C.16)}$$

Stata: `reghdfe unanimous info woman infoWoman if comm==1, absorb(pair) vce(cluster committee)`

$$\text{UnanimousWomanWins}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \quad \text{(C.17)}$$

Stata: `reghdfe unanimousWomanWins info woman infoWoman if comm==1, absorb(pair) vce(cluster committee)`

$$\text{UnanimousManWins}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \quad \text{(C.18)}$$

Stata: `reghdfe unanimousManWins info woman infoWoman if comm==1, absorb(pair) vce(cluster committee)`

	Unanimous b/se	Unan. Female b/se	Unan. Male b/se
Org. Values	-0.009	-0.020	0.011
	(0.09)	(0.06)	(0.08)
Committee has at least one female member=1	0.137*	0.073	0.064
	(0.08)	(0.06)	(0.06)
Org. Values × Committee has at least one female member=1	-0.028	0.019	-0.046
	(0.13)	(0.09)	(0.10)
Pair FE	Yes	Yes	Yes
Control mean	0.327	0.148	0.180
N	444	444	444

Notes This table implements Equation C.16. It shows the effect of the information treatment and gender composition of the triplet on the probability of unanimity among the grouped triplets of judges who – within each treatment arm – assessed the same pair of candidates. We show effects (i) on unanimity, (ii) on unanimity in favour of the male candidate, and (iii) on unanimity in favour of the female candidate. Standard errors are clustered at the committee level and the regression includes pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.5: The effect of the organisational values and triplet composition on unanimity

D Details experiment

D.1 Protocol for the control group

This subsection details the protocol for the business plan competition for those judges not assigned to a treatment, (the control group). Following this we then detail how the protocol is different for those receiving the social image treatment. There are no further details to report on the organisational values treatment.

Judges are invited in groups of 16 to a hotel in central Addis Ababa. Once the judges arrive at the hotel and register, the assessments will start at a fixed time after which no new judges are included in the main sample. To start, the judges enter a room set up in a classroom style format and watch two videos. The first video explains the full format of the assessments and how the winner will be decided; the second video features a prominent Ethiopian businessman discussing the importance of access to start-up capital for aspiring entrepreneurs in Ethiopia (and, thus, the importance of the business plan competition). After watching these two videos, and after having had an opportunity to ask questions after the first video, the responses to which are strictly pre-defined, the assessments start.

The respondents then start the assessments in which they will watch the three-minute recordings of twelve pairs of candidates in the business plan competition. After watching each pair, which consists of one female and one male candidate, they are asked which of these candidates they want to cast their vote for for the business plan competition and how difficult it was to make this decision. The respondent answers these questions without showing the enumerator their answer.

These assessments provide the data for the main experiment, with as primary outcome whether the judge votes for the female candidate. Following the assessments, we then return to one of the pair of candidates and ask the judge to provide feedback. The enumerator observes, using the survey software, who the judge voted for for one pair of candidates. The respondent watches the recordings of these two candidates again, and is asked to write down reasons for their decision and to provide some feedback for the candidates both on some fixed dimensions and open-ended feedback. Finally, the judge is asked the following question for both a male and female judge:

> Please now imagine a [male/female] judge with [x] years of managerial experience working as [position] in [industry] has also assessed these candidates. We would like to ask you whom you think they would have voted for.

The purpose of this question is to elicit beliefs on how other judges would vote. The specific set-up is designed to mirror the committee treatment. These fields are filled based on the set of other judges who are invited to attend the same assessment session, i.e. if judges A and B both attend the same assessment session, judge A may here see these characteristics for judge B. Note that while the position, industry and years of experience are based on the actual characteristics, the gender may be changed to ensure the judge is asked about both a male and female judge.

We then run a "feedback session" with the individual judge in which they first, privately, write down reasons for their decision. They then give feedback on a number of fixed dimensions taken from Fafchamps and Quinn (2018) before being given the opportunity to provide some open-form feedback to the candidate. This feedback will then, after the conclusion of the full experiment, be shared with the candidate.

To finish the assessment day, judges are asked a number of closed form-questions about the experiment, about their relationships with other judges and on the ambivalent sexism inventory Glick and Fiske (1997);

see the separately attached document 'post-competition questions' for the full questionnaire. We implement this after the experiment in order to avoid the possibility of experimenter demand effects.
The entire experiment is conducted in Amharic.

D.2 Further details on the Social Image Treatment

The protocol for judges assigned to the social image treatment differs from that for individual judges as follows, and is explained in a separate explanatory video to that for individual judges:

- The room will be set up boardroom rather than classroom style so judges can see each other, and crucially the judges with whom they are on a committee.
- At the start of the assessment day, judges are asked to introduce themselves by telling other judges their name, industry and their position at their company. The judges are asked not to share any additional information.
- Judges still go through the decisions individually, but know they are deciding together with two other judges in the room. Before they watch the videos for each pair of candidates, they are shown photo CV's of these two judges. This includes, beyond a photo of the judge, their name, industry and position at their company. They are also reminded their decisions may be shared with these judges in the feedback sessions.
- The judges are on a different random triplet of judges for each pairwise assessment.
- After the twelve assessments, they also re-watch one of the pairs of candidates' responses. They are then asked who they think *the other two judges on this triplet* voted for. They again see the exact same set of characteristics as those the individual judges see based on the photo CV, and now additionally know this judges' name and what they look like.
- The three judges then come together and are told who each of them voted for for the pair of candidates whose videos they just re-watched. They are asked to each give reasons for their decisions to each other as the judges not receiving the social image treatment judges did individually. They then, together, provide the same feedback as an individual judge to the candidate.
- The judges know that in the case of a split decision, the candidates will get respectively 2 and 1 point, *i.e.* one point for each vote that is cast for them.

To ensure judges on the social image treatment are aware of the section at the end of the experiment in which their decisions are made public to other judges. To do so, the feedback session is clearly highlighted in the video at the start of the assessments explaining the protocol. In this video, actors playing judges go through the full experiment and are shown discussing their decisions (with the sound muted; the respondents know these are actors). Beyond this, judges are reminded every time they are shown the judges on their triplet that they may have to give reasons for their decisions to these two judges.

D.3 Randomization

This section details how randomisation is done for the experiment.

Assignment of judges to treatments

Judges are first invited to participate as a judge in the business plan competition. We aim to over-sample judges due to high expected non-attendance, thus intending to invite 320 judges in total aiming to have 240 actually attend including 180 male and 60 female judges. After inviting the judges, we randomly assign them to a treatment assigning 20 women and 60 men to each treatment. We thus assign 80 judges assigned to each treatment. Once a judge is assigned to one treatment, they will not be swapped across treatments. To specify how randomisation is done, once we have a sample of judges we will randomly split judges across the treatments using Stata's "cut" command based on a random uniform variable with no duplicates. Using this command, the set of male and female judges are separately assigned to the four treatments resulting in a quarter of the male and female judges in the sample being assigned to each treatment.

Order of sessions

We will run five or six sessions for each treatment arm over the course of two weeks, running two sessions each day. Every two days, fours sessions, one for each of the four treatment arms, will be conducted during this time (except on Sunday). This ensures ex-ante attendance for each treatment arm is expected to be the same, and we impose the same constraints for inclusion on individual and committee judges.

Assignment of judges to sessions

We flexibly assign judges to a specific assessment day. Judges are randomly assigned to and invited for a specific assessment day, but if they cannot attend this slot they are offered to attend one of the other four slots for their treatment.
For each assessment day, we aim to include at least ten male and two female judges. To achieve this, we invite eleven male and four or five female judges to each assessment day. We will conduct the experiment if more than eight judges show up, if fewer than eight judges attend the assessments will be rescheduled.
Note that to do the feedback sessions at the end of the sessions for the committee treatment, we require a multiple of three judges to attend for each judge can be included in a three-member assessment. To maximise the number of judges we can include, we allow for some judges not to do a feedback session for the committee treatment if they cannot do so as part of a three-member committee.

Assignment of judges to triplets

The final element of randomisation is due to the difficulty in predicting how many judges will attend. On the day, each judges is randomly assigned a "judge number" from one to the number of judges attending. These judges numbers have been randomly pre-assigned to be members of a specific set of triplets assessing specific pair of candidates. This method allows us to flexibly deal with attrition in a logistically achievable way.

Assignment of candidates to competitions

Finally, we randomly assign candidates to competitions. We first stratify by gender, to then randomly assign five male and five female candidates to each competition. As the candidates are not present at the

competition, there are no logistical constraints in this randomisation.

D.4 Invitation judges

I am calling on behalf of EconInsights, an Ethiopian research company. Your company has participated in several surveys with us as part of a research project of the University of Oxford over the past years. Most recently, you have helped us rank aspiring managers and entrepreneurs based on hypothetical vignettes. As we told you at the time, the individuals who performed best in the vignettes as an entrepreneur have now been invited for a business plan competition.
We would like to draw on your personal expertise as a successful member of the Ethiopian business community, and invite you to be a judge in the business plan competition. In this role, you would be watching videos of young individuals pitching business plan proposals. We would like you to help us decide who should get a 50.000 Birr grant towards their business plan.
We would ask for you to join us for around three to four hours to do these assessments. We would pay you X Birr for your time in addition to covering transport expenses. You would not have to prepare anything for these assessments. If you are willing to participate, we will call you in the near future to schedule a time and place sometime in the coming month for this competition. Would you be willing to participate?
If yes: We would like to introduce you to the other judges using your professional credentials. Our records say these are as follows:

- Name [X]
- Company [X]
- Position [X]
- Years of experience as a manager [X]

Is this the correct (q1) name, (q2) company, (q3) position and (q4) years of experience? [Enumerator: Update this in our records with follow-up question if not.]

D.5 Invitation candidates

Dear X,
I am calling on behalf of EconInsight. As you will remember, you were invited to participate in a management challenge where you got to see different scenarios and we recorded you responding to these scenarios on X date. As promised, we showed these recordings to human resources managers of different firms to determine who would participate in a business plan competition designed to support promising young Ethiopian entrepreneurs. Based on the assessment we obtained from HR managers, we're happy to announce that you have been selected to present your business ideas for the business plan competition. We would like to congratulate you, as your good performance means that you can potentially obtain a reward of 50,000 Birr if your business plan is selected by independent judges. I will now offer you some key details to help you prepare your business plan. In this competition, you will go up against nine other candidates for a chance to win the 50,000 Birr prize. To participate, we will ask you to come to our studio

to record a three minutes business proposal before the 11th of March. Our judges, experienced HR managers, will look at your entry in the following month, and we will aim to inform you of the results by the end of April. If you do not show up to the Studio to record your business plan by March 11th , you will not be able to participate and you lose the opportunity to win the prize .

For the competition, we will ask you to prepare a three-minute pitch for your business, also giving a brief introduction of yourself. This can either be for an existing business, or a plan for a new business. In your entry, you should split your presentation between introducing yourself, your business idea (opportunity), target market, potential competition, operations and cost of business. This should all be done at most in three minutes. Do you have any further questions?

If you are interested, you simply need to agree that we can record your proposal and play them for our judges, primarily human resources managers; we would also use your answers, anonymously, as part of our research on business plan competitions and committees in Ethiopia. We will not share with the judges any of the answers you have given in any previous questionnaires. We would like you to come to the studio as soon as possible. Could you attend on X date? Our Studio is located around Meskel flower road, in a building which hosts Kezira advertising on the 6th floor. [note to self: include other landmarks]. Of course, feel free to get in touch with me if you would like me to give you the precise location of the Studio. Also note that we will fully cover your transport expenses to make the trip to the Studio.

We will give you a follow-up call to remind you your appointment day in the day before you are scheduled to come to the studio. Is this the best phone number to reach you?

D.6 Summary statistics judges

Table A.6: Judge-level summary statistics

	Overall	Control	Org. Values	Social Image	Both Treatments	p-value
Gender (1=male)	0.741	0.702	0.793	0.722	0.750	0.702
Judge age	41.41	38.91	42.62	41.15	43.04	0.246
Experience in current position (years)	6.12	6.49	6.81	5.93	5.27	0.474
Total experience (years)	19.68	19.04	19.95	19.91	19.78	0.981
Formal management education	0.770	0.772	0.793	0.722	0.804	0.696
At least a BA	0.782	0.772	0.810	0.806	0.732	0.719
HR	0.407	0.491	0.345	0.403	0.393	0.452
Administration	0.342	0.281	0.431	0.306	0.357	0.327
Finance	0.128	0.140	0.103	0.125	0.143	0.918
Other	0.123	0.088	0.121	0.167	0.107	0.565
Most senior manager or owner	0.329	0.298	0.431	0.278	0.321	0.281
Finance and administration	0.148	0.158	0.155	0.125	0.161	0.933
HR Manager	0.325	0.386	0.259	0.333	0.321	0.545
Other	0.198	0.158	0.155	0.264	0.196	0.362

Notes This table displays the average characteristics of judges in each of the treatment arms. The department of the manager is the main department in which the manager operates in their firm, the other category contains all departments in which fewer than 5% of managers work. Columns two to six report these characteristics by treatment, and column seven reports the p-value for a Wald test for the hypothesis that these means are equal across the treatment arms.

D.7 Summary statistics expert rankings

	Male candidates	Female candidates	Difference	p-value
Expert 1 Score	10.743	10.055	0.688	0.005
Expert 2 Score	14.873	15.270	-0.397	0.168
Average Score	12.808	12.662	0.146	0.518
Expert 1 choice	0.498	0.380	-0.118	0.052
Expert 2 choice	0.401	0.515	0.114	0.067
Expert favourite	0.460	0.473	0.013	0.841
Unanimous expert favourite	0.283	0.257	-0.025	0.597

Notes: This table describes the expert scores for male and female candidates, and tests whether the two HR experts assess male and female candidates to be similar on average. The table reports Expert 1 Score and Expert 2 Score (the average raw score from 1-20 for each of the experts), the average score (averaged by candidate over the two experts), the probability of being Expert 1 and Expert 2 favourite, the probability of being the "expert favourite", i.e. preferred by the expert based on the average score, and the "unanimous expert favourite, i.e. preferred by both experts. Standard errors are clustered at the judge level and the regression includes pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.7: Summary statistics of expert rankings

E Additional Results

E.1 Appendix Table and Figure: Continuous rounds

	(1) Woman Wins	(2) Expert-favoured Candidate Wins	(3) Expert-favoured Woman Wins	(4) Expert-favoured Man Wins
roundJudge	-0.013*	-0.006	-0.009**	0.003
	(0.007)	(0.006)	(0.004)	(0.005)
Org. Values × roundJudge	0.019**	0.014*	0.016***	-0.003
	(0.008)	(0.008)	(0.006)	(0.006)
Social Image × roundJudge	0.009	0.008	0.008	0.000
	(0.008)	(0.007)	(0.005)	(0.005)
Org. Values & Social Image × roundJudge	0.015*	0.009	0.011**	-0.002
	(0.008)	(0.008)	(0.006)	(0.006)
Org. Values	-0.075	-0.053	-0.060	0.006
	(0.062)	(0.057)	(0.040)	(0.044)
Social Image	-0.034	-0.040	-0.034	-0.006
	(0.060)	(0.056)	(0.037)	(0.043)
Org. Values & Social Image	-0.078	-0.059	-0.063	0.004
	(0.062)	(0.057)	(0.039)	(0.045)
Pair FE	Yes	Yes	Yes	Yes
Control mean (early rounds)	0.486	0.576	0.278	0.299
N	2,649	2,649	2,649	2,649

Notes: Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.8: The treatment effect interacted with rounds (continuous)

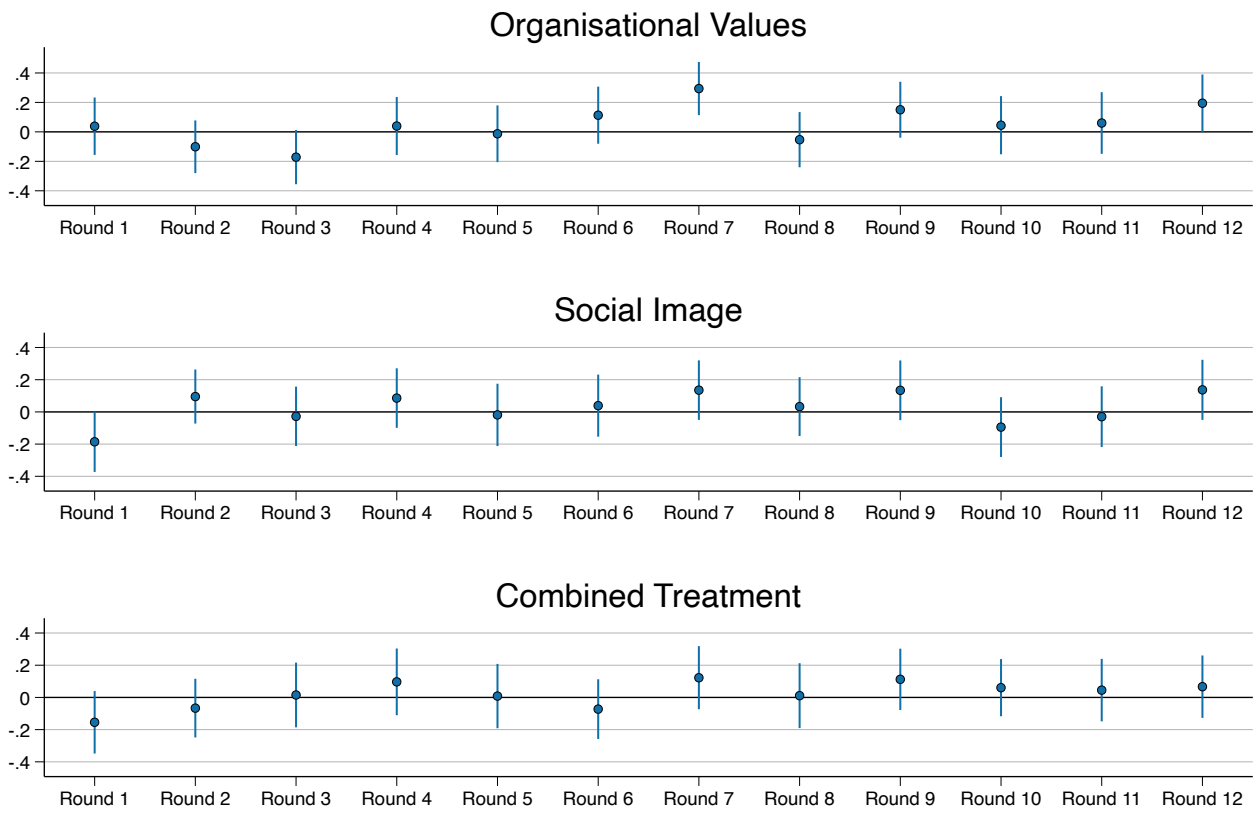


Figure A.2: Treatment effect by individual round

Here we report the coefficients β from: $\texttt{womanWins}_{jp} = \sum_{r=1}^{12} \alpha_r \cdot [\texttt{round}_{jp} = r] + \sum_{r=1}^{12} \sum_{t=2}^{4} \beta_{rt} [\texttt{round}_{jp} = r] \cdot \textit{Treat}_j + \mu_p + \varepsilon_{jp}$.

	Male Score	Female Score	Score Diff.	Share Fem. Fav.	N
1	12.776	12.487	-0.288	0.455	156
2	12.723	12.620	-0.103	0.492	179
3	12.812	12.723	-0.089	0.462	186
4	12.944	12.658	-0.286	0.439	187
5	12.787	12.624	-0.162	0.459	157
6	12.777	12.693	-0.084	0.486	179
7	12.781	12.794	0.013	0.471	187
8	12.965	12.674	-0.291	0.435	184
p-value (joint)	0.988	0.986	0.985	0.954	.

Notes: Each row corresponds to one of eight treatment-arm $\times$ round-half cells: rows 1–4 cover the early rounds (1–6) of the control, Organisational Values, Self-Image, and combined treatments, respectively; rows 5–8 cover the late rounds (7–12) of the same four arms. "Male Score" and "Female Score" are the mean expert evaluation scores of the male and female candidates in each cell; "Score Diff." is the female-minus-male mean score; "Share Fem. Fav." is the fraction of pairs in which the female candidate has the higher average expert score; N is the number of distinct candidate pairs contributing to the cell. The sample is collapsed to the pair level within each cell so that each pair contributes once. The final row reports the p-value of a one-way ANOVA F-test of equality of the cell means across all eight cells. Statistical significance is denoted by * ($p < 0.10$), ** ($p < 0.05$), and *** ($p < 0.01$).

Table A.9: Balance of expert scores across treatment arms and rounds

F Robustness to the choice of sample

Our experiment featured two sessions that slightly deviated from the pre-analysis plan:

(i). For the very first set of assessments, only six judges participated—fewer than the minimum of eight we intended. This session was part of the control group, meaning judges decided privately and individually, reducing the likely impact. However, because randomization was implemented *as if* there were eight judges, the session produced a significant number of incomplete triplets.

(ii). In a separate control-group session, only male judges participated in a group of fourteen members.

In our main analysis we drop pairwise assessments only when fewer than three judges within one treatment arm assess a given candidate pair. This occurred in the first session described above and in a few other assessments where judges ran out of time. The objective is to maintain the same set of assessments for both judge-level and committee-level analyses.

This section shows that all main results are robust to alternative sampling choices. We consider three restrictions:

(i). **Main**: Our primary sample specification.

(ii). **Fully Covered**: Only pairs observed in all four treatment arms.

(iii). **Excl. First Day**: Dropping the session with only six participating judges (i).

The all-male assessment day (ii) is retained in all specifications; its exclusion does not alter any result.

Table A.10 consolidates robustness checks across subsamples. Columns (1)–(3) correspond to the three sample restrictions, and Panels A–C report treatment effects on three outcomes: (A) voting for a female candidate, (B) voting for the expert-favoured female candidate (restricting to pairs where the female candidate is expert-preferred), and (C) reaching a unanimous decision among the three judges assessing a pair.

All statistically significant parameter estimates in the main analysis retain the same sign across every robustness subsample, and nearly all remain significant at the 10% level or better. Point estimates are generally stable, with variation in significance levels reflecting changes in sample size and clustering structure.

Figure A.3 shows that the distribution of expert score differences is virtually identical across treatment arms, confirming that randomisation achieved balance in the quality composition of candidate pairs.

These sample restrictions do not affect the composition results. The changes pertain only to control-group sessions or assessments that do not form complete triplets. The composition analysis includes only complete triplets in the social image and combined treatment arms; consequently, the composition sample is unchanged across these restrictions.

F.1 Main treatment effects across subsamples

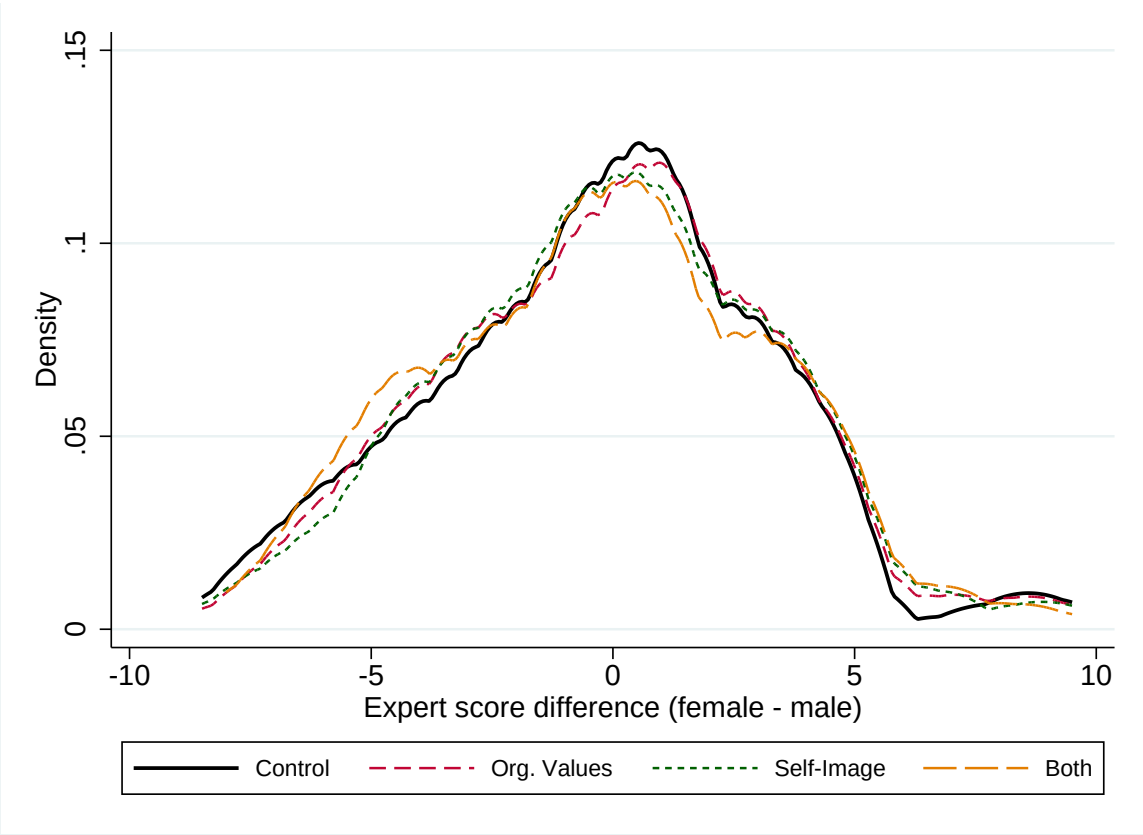
F.2 Late-round heterogeneity across subsamples

Our main finding is that gender bias emerges with evaluation fatigue: in late rounds (rounds 7–12), control-group judges are significantly less likely to select the female candidate, while the organisational values

Panel A: Voting for female candidate			
	(1) Main b/se	(2) Fully covered b/se	(3) No ctrl day 1 b/se
Org. Values	0.049	0.061*	0.051
	(0.03)	(0.04)	(0.03)
Self-Image	0.022	0.013	0.025
	(0.03)	(0.04)	(0.03)
Org. Values & Self-Image	0.017	0.047	0.020
	(0.03)	(0.04)	(0.03)
Control mean	0.457	0.439	0.456
N	2649	1443	2619
Panel B: Voting for expert-favoured female candidate			
	(1) Main b/se	(2) Fully covered b/se	(3) No ctrl day 1 b/se
Org. Values	0.095**	0.103*	0.098**
	(0.05)	(0.06)	(0.05)
Self-Image	0.039	0.041	0.041
	(0.04)	(0.05)	(0.05)
Org. Values & Self-Image	0.017	0.050	0.020
	(0.05)	(0.06)	(0.05)
Control mean	0.554	0.545	0.556
N	1221	651	1206
Panel C: Unanimous decision			
	(1) Main b/se	(2) Fully covered b/se	(3) No ctrl day 1 b/se
Org. Values	0.057	0.084	0.057
	(0.05)	(0.06)	(0.05)
Self-Image	0.081*	0.119**	0.079
	(0.05)	(0.06)	(0.05)
Org. Values & Self-Image	0.060	0.072	0.057
	(0.05)	(0.06)	(0.05)
Control mean	0.317	0.299	0.324
N	867	481	856

Notes: This table reports OLS treatment effects across three sample restrictions. The dependent variable is: voting for the female candidate (Panel A, column headers from fragment), voting for the expert-favoured female candidate (Panel B), and a unanimous committee decision (Panel C). All specifications include candidate-pair fixed effects. Standard errors clustered at the judge level (Panels A–B) or committee level (Panel C) are reported in parentheses. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.10: Robustness of main treatment effects to alternative sample restrictions



Notes: Each line shows the kernel density of the expert score difference (female minus male candidate) for a different treatment arm. The distributions nearly perfectly overlap, confirming that randomisation achieved balance in the quality composition of candidate pairs across treatments.

Figure A.3: Distribution of expert score differences by treatment arm

treatment fully counteracts this decline. Here we show that this result is robust across all three sample restrictions defined above.

Table A.11 restricts the sample to late rounds and shows treatment effects on voting for the female candidate across the three subsamples. Table A.12 reports the full treatment $\times$ late-round binary interaction, and Table A.13 reports the treatment $\times$ continuous-round interaction. In all cases, the late-round treatment effects are stable across subsamples.

	(1) Main b/se	(2) Fully covered b/se	(3) No ctrl day 1 b/se
Org. Values	0.112***	0.132***	0.114**
	(0.04)	(0.05)	(0.04)
Self-Image	0.047	0.022	0.047
	(0.04)	(0.05)	(0.04)
Org. Values & Self-Image	0.080*	0.100*	0.080*
	(0.04)	(0.05)	(0.05)
Control mean	0.424	0.391	0.427
N	1298	717	1280

Notes: This table restricts the sample to late rounds (evaluation rounds 7–12) and shows treatment effects on voting for the female candidate. Standard errors are clustered at the judge level and all regressions include pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.11: Robustness: Treatment effects in late rounds across subsamples

	(1) Main b/se	(2) Fully covered b/se	(3) No ctrl day 1 b/se
Org. Values	-0.019	-0.013	-0.014
	(0.04)	(0.05)	(0.04)
Self-Image	-0.005	-0.009	0.000
	(0.04)	(0.05)	(0.04)
Org. Values & Self-Image	-0.033	-0.016	-0.027
	(0.04)	(0.05)	(0.04)
Second half	-0.105**	-0.122**	-0.100**
	(0.05)	(0.06)	(0.05)
Org. Values × Second half	0.137**	0.147**	0.130**
	(0.06)	(0.07)	(0.06)
Self-Image × Second half	0.055	0.042	0.050
	(0.05)	(0.07)	(0.05)
Org. Values & Self-Image × Second half	0.101*	0.123	0.096
	(0.06)	(0.08)	(0.06)
Control mean	0.457	0.439	0.456
N	2649	1443	2619

Notes: This table interacts the treatment indicators with a binary late-round indicator (=1 for rounds 7–12). Standard errors are clustered at the judge level and all regressions include pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.12: Robustness: Treatment × late-round interaction across subsamples

	(1) Main b/se	(2) Fully covered b/se	(3) No ctrl day 1 b/se
Org. Values	-0.075	-0.055	-0.070
	(0.06)	(0.08)	(0.06)
Self-Image	-0.034	-0.040	-0.029
	(0.06)	(0.08)	(0.06)
Org. Values & Self-Image	-0.078	-0.068	-0.073
	(0.06)	(0.08)	(0.06)
roundJudge	-0.013*	-0.014*	-0.012*
	(0.01)	(0.01)	(0.01)
Org. Values × roundJudge	0.019**	0.018*	0.019**
	(0.01)	(0.01)	(0.01)
Self-Image × roundJudge	0.009	0.008	0.009
	(0.01)	(0.01)	(0.01)
Org. Values & Self-Image × roundJudge	0.015*	0.018	0.014*
	(0.01)	(0.01)	(0.01)
Control mean	0.457	0.439	0.456
N	2649	1443	2619

Notes: This table interacts the treatment indicators with the continuous evaluation round number. Standard errors are clustered at the judge level and all regressions include pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.13: Robustness: Treatment × continuous round interaction across subsamples

F.3 Sensitivity to expert favourite definition

Our analyses frequently condition on whether the female candidate is expert-favoured, defined as having a higher average expert score. Table A.14 shows that the main treatment effects are robust to alternative definitions: a strict threshold requiring a score difference exceeding half a standard deviation, and a requirement that both individual experts agree on the female candidate's superiority.
Table A.15 extends this by including the treatment $\times$ late-round interaction under each definition. The emergence of bias in late rounds and its prevention by the organisational values treatment are stable across expert-favourite operationalisations.

	(1) Avg. score b/se	(2) Strict (>0.5 SD) b/se	(3) Both experts b/se
Org. Values	0.095**	0.130**	0.036
	(0.05)	(0.05)	(0.06)
Self-Image	0.039	0.043	0.033
	(0.04)	(0.05)	(0.05)
Org. Values & Self-Image	0.017	0.029	-0.009
	(0.05)	(0.05)	(0.06)
Control mean	0.554	0.595	0.644
N	1221	723	648

Notes: Column (1) restricts to pairs where the female candidate has a higher average expert score. Column (2) requires the score difference to exceed 0.5 standard deviations. Columns (3) and (4) restrict to pairs where expert 1 or expert 2, respectively, individually favours the female candidate. Standard errors are clustered at the judge level and all regressions include pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.14: Sensitivity of treatment effects to expert favourite definition

	(1) Avg. score b/se	(2) Strict (>0.5 SD) b/se	(3) Both experts b/se
Org. Values	-0.037	-0.014	-0.112
	(0.06)	(0.07)	(0.08)
Self-Image	-0.033	-0.018	-0.024
	(0.06)	(0.07)	(0.08)
Org. Values & Self-Image	-0.080	-0.071	-0.085
	(0.06)	(0.08)	(0.08)
Second half	-0.186***	-0.171**	-0.184**
	(0.06)	(0.08)	(0.08)
Org. Values × Second half	0.268***	0.279***	0.278***
	(0.08)	(0.09)	(0.11)
Self-Image × Second half	0.142*	0.114	0.103
	(0.08)	(0.10)	(0.10)
Org. Values & Self-Image × Second half	0.193**	0.192*	0.137
	(0.09)	(0.10)	(0.11)
Control mean	0.554	0.595	0.644
N	1221	723	648

Notes: Each column uses a different expert favourite definition (see Table A.14). The specification interacts treatment indicators with a late-round indicator (rounds 7–12). Standard errors are clustered at the judge level and all regressions include pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.15: Expert favourite definition sensitivity: treatment × late round

F.4 Pre-specified factorial specification across subsamples

The pre-analysis plan specified a factorial decomposition of the treatment into an organisational values component (Info) and a social image component (Comm), with their interaction. Table A.16 shows this specification applied to the probability of voting for the female candidate across the three subsamples. Table A.17 applies the same specification to the probability of a unanimous committee decision. The factorial decomposition is stable across all three sample restrictions.

	(1) Main b/se	(2) Fully covered b/se	(3) No ctrl day 1 b/se
Org. Values	0.049	0.061*	0.051
	(0.03)	(0.04)	(0.03)
Self-Image	0.022	0.013	0.025
	(0.03)	(0.04)	(0.03)
Org. Values × Self-Image	-0.054	-0.027	-0.056
	(0.04)	(0.05)	(0.04)
Control mean	0.457	0.439	0.456
N	2649	1443	2619

Notes: This table uses the pre-specified factorial decomposition (Org. Values × Social Image) rather than treatment dummies. Standard errors are clustered at the judge level and all regressions include pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.16: Robustness: PAP factorial specification across subsamples (female wins)

	(1) Main b/se	(2) Fully covered b/se	(3) No ctrl day 1 b/se
Org. Values	0.057	0.084	0.057
	(0.05)	(0.06)	(0.05)
Self-Image	0.081*	0.119**	0.079
	(0.05)	(0.06)	(0.05)
Org. Values × Self-Image	-0.078	-0.132	-0.078
	(0.07)	(0.09)	(0.07)
Control mean	0.317	0.299	0.324
N	867	481	856

Notes: This table uses the pre-specified factorial decomposition (Org. Values × Social Image) rather than treatment dummies, with unanimity as the outcome. Standard errors are clustered at the committee level and all regressions include pair fixed effects. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.17: Robustness: PAP factorial specification across subsamples (unanimous)

G External Validation: Gender Bias in Candidate Evaluation

In a separate experiment, HR managers from the same pool of firms evaluated young professionals for their suitability for entrepreneurship and a management position in their firm through video vignettes.[2] Three features of this experiment make it an ideal setting for our purposes. First, candidates were randomly assigned to triplets, so the gender composition of each group is orthogonal to candidate quality. Second, each triplet was independently assessed by two HR managers, and the order in which managers encountered the five triplets was randomised. This means we can compare evaluations of the *same* candidates, in the *same* triplet, by two managers who encounter that triplet at different points in the sequence—one as her first assessment, another as her fifth. Any difference in how they evaluate female candidates within the triplet is therefore attributable to sequence position, not to candidate or group characteristics. Third, the full set of assessments took each manager approximately one hour, providing sufficient scope for fatigue to accumulate across rounds.

[2] We omit the firms from which we include managers in this experiment's control group from this analysis, and show the results are robust to including them.

	All		Male Manager		Female Favourite	
	(1) By Round	(2) Linear	(3) By Round	(4) Linear	(5) By Round	(6) Linear
Round 2 × Female	0.008		-0.014		-0.111	
	(0.062)		(0.110)		(0.093)	
Round 3 × Female	-0.100*		-0.220*		-0.233**	
	(0.059)		(0.113)		(0.094)	
Round 4 × Female	-0.027		-0.122		-0.071	
	(0.061)		(0.111)		(0.092)	
Round 5 × Female	-0.152**		-0.256**		-0.283***	
	(0.061)		(0.112)		(0.095)	
Round × Female		-0.034***		-0.062***		-0.051**
		(0.013)		(0.023)		(0.020)
Female × Male Mgr			0.013	-0.006		
			(0.101)	(0.097)		
Round 2 × Female × Male Mgr			0.036			
			(0.136)			
Round 3 × Female × Male Mgr			0.168			
			(0.140)			
Round 4 × Female × Male Mgr			0.107			
			(0.141)			
Round 5 × Female × Male Mgr			0.146			
			(0.136)			
Female × Male Mgr × Round				0.036		
				(0.029)		
Female × Fem. Favourite					-0.180**	-0.118
					(0.088)	(0.090)
Round 2 × Female × Fem. Favourite					0.228*	
					(0.117)	
Round 3 × Female × Fem. Favourite					0.229*	
					(0.120)	
Round 4 × Female × Fem. Favourite					0.102	
					(0.117)	
Round 5 × Female × Fem. Favourite					0.214*	
					(0.120)	
Female × Fem. Favourite × Round						0.030
						(0.026)
N	15,506	15,506	13,408	13,408	14,808	14,808
Firm × Vignette × Sector FE	Yes	Yes	Yes	Yes	Yes	Yes
Candidate FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. var.	0.342	0.342	0.341	0.341	0.342	0.342

Notes: OLS estimates. The dependent variable is an indicator equal to one if the candidate is ranked first among three candidates in the vignette. *Female* is an indicator for female candidates. *Round* indicates the presentation order of the vignette (1–5). Columns (1)–(2) pool all managers; columns (3)–(4) interact with an indicator for male managers; columns (5)–(6) interact with an indicator for whether the female candidate has the highest leave-one-out expert score in the group. Odd-numbered columns include round as a factor variable with Round 1 as the omitted category; even-numbered columns model round linearly. All specifications include firm × vignette × sector and candidate fixed effects. Standard errors clustered at the firm level in parentheses. The sample excludes firms whose managers participated in the committee experiment control group. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.18: Effect of candidate gender on ranking outcomes (excl. committees control group)

Table A.18 reports the ranking results. Table A.19 repeats the same specifications including firms from the committee experiment control group; the estimates are qualitatively unchanged.

	All		Male Manager		Female Favourite	
	(1) By Round	(2) Linear	(3) By Round	(4) Linear	(5) By Round	(6) Linear
Round 2 × Female	0.006		0.061		-0.090	
	(0.059)		(0.106)		(0.089)	
Round 3 × Female	-0.102*		-0.149		-0.229**	
	(0.056)		(0.106)		(0.089)	
Round 4 × Female	-0.026		-0.078		-0.057	
	(0.057)		(0.102)		(0.085)	
Round 5 × Female	-0.123**		-0.118		-0.202**	
	(0.060)		(0.108)		(0.091)	
Round × Female		-0.028**		-0.038*		-0.037*
		(0.013)		(0.022)		(0.019)
Female × Male Mgr			0.073	0.052		
			(0.095)	(0.091)		
Round 2 × Female × Male Mgr			-0.047			
			(0.132)			
Round 3 × Female × Male Mgr			0.083			
			(0.132)			
Round 4 × Female × Male Mgr			0.064			
			(0.130)			
Round 5 × Female × Male Mgr			0.015			
			(0.132)			
Female × Male Mgr × Round				0.014		
				(0.028)		
Female × Fem. Favourite					-0.155*	-0.089
					(0.082)	(0.085)
Round 2 × Female × Fem. Favourite					0.180	
					(0.112)	
Round 3 × Female × Fem. Favourite					0.218*	
					(0.114)	
Round 4 × Female × Fem. Favourite					0.086	
					(0.109)	
Round 5 × Female × Fem. Favourite					0.139	
					(0.116)	
Female × Fem. Favourite × Round						0.019
						(0.025)
N	17,172	17,172	14,926	14,926	16,444	16,444
R^2	0.214	0.213	0.233	0.233	0.219	0.218
Firm × Vignette × Sector FE	Yes	Yes	Yes	Yes	Yes	Yes
Candidate FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. var.	0.342	0.342	0.342	0.342	0.342	0.342

Notes: Same specification as Table A.18, but including firms whose managers participated in the committee experiment control group. See Table A.18 notes for full details. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.19: Effect of candidate gender on ranking outcomes (incl. committees control group)

HR managers also give a Likert score to each aspiring manager. The ranking results do not carry over to these scores. Table A.20 shows no effect of candidate gender on continuous ratings, consistent with evidence that comparative judgments reveal biases that absolute ratings do not.

	All		Male Manager		Female Manager	
	(1) By Round	(2) Linear	(3) By Round	(4) Linear	(5) By Round	(6) Linear
Round 2 × Female	0.019		0.111		-0.068	
	(0.115)		(0.157)		(0.215)	
Round 3 × Female	-0.133		-0.069		0.082	
	(0.109)		(0.144)		(0.244)	
Round 4 × Female	-0.055		0.071		-0.366	
	(0.111)		(0.148)		(0.233)	
Round 5 × Female	-0.076		0.091		-0.211	
	(0.119)		(0.166)		(0.251)	
Round × Female		-0.024		0.013		-0.074
		(0.025)		(0.034)		(0.052)
N	17,131	17,131	10,754	10,754	4,141	4,141
R^2	0.493	0.492	0.517	0.517	0.658	0.658
Mean dep. var.	3.319	3.319	3.338	3.338	3.323	3.323

Notes: OLS estimates. The dependent variable is the Likert score (1–5) assigned by the HR manager, pooling managerial and entrepreneurial ability ratings. Columns (1)–(2) pool all managers; columns (3)–(4) restrict to male managers; columns (5)–(6) restrict to female managers. Odd-numbered columns include round as a factor variable with Round 1 as the omitted category; even-numbered columns model round linearly. All specifications include firm × vignette × domain and candidate fixed effects. Standard errors clustered at the firm level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.20: Effect of candidate gender on Likert scores

H Machine Learning Procedures

This section describes the machine-learning procedures used to examine treatment effect heterogeneity.

H.1 Variables and variable construction

Each observation is a judge-level assessment of a candidate pair. The outcome is `womanWins`, an indicator equal to one when the judge selects the female candidate. Treatments are coded as in the main analysis: 1 is control, 2 is organizational values, 3 is social image, and 4 is the combined treatment. Early rounds are rounds 1–6 (`roundJudge` ≤ 6); late rounds are rounds 7–12.
The covariate matrix combines several families of predictors, summarised in Table A.21.

Family	Representative variables and role
Identifiers and design variables	`pair`, `treatment`, `judge`, and `roundJudge`. These columns define the treatment comparisons, the early-versus-late split, and the clustering or subsampling units used by the algorithms.
Baseline judge and session variables	`shareWomen`, `shareWomenCommittee`, `judgeAge`, `oldestJudge`, `youngestJudge`, `ageDiff`, `HS`, `BS`, and `femaleJudge`. These variables capture judge characteristics and the local committee environment.
Expert-evaluation variables	`overallExp`, `ConceptExp`, `MarketExp`, `GrowthExp`, `FinanceExp`, `PlanExp`, `SenseExp`, and `PresentExp`. These variables summarize expert assessments of candidate quality and presentation.
Text and language variables	`valenceMean`, `engagementMean`, `agentic_per_1000`, `communal_per_1000`, `genbit_avg_bias`, and the `empath_*` measures. These variables capture sentiment, linguistic content, and gendered language in submissions.
Presentation and content variables	`appearanceAppropriateExp`, `confidenceExp`, `arrogantExp`, `certainExp`, `mentionsIdea`, `mentionsMarket`, `dressedFormalEnu`, `careOutfitEnu`, `generalAppearance`, `confidenceEnu`, `teamworkEnu`, `Articulate`, and `specificExamples`. These variables measure visible presentation style, business content, and other candidate characteristics.

Table A.21: Variables used in the machine-learning procedures

Candidate-level covariates—expert evaluation sub-scores, text and language features, presentation assessments, business content indicators, and candidate age—are encoded as *standardised female-minus-male differences* within each pair. A positive value indicates that the female candidate scores higher on that dimension. Judge and session variables (demographics, sexism scales, and committee gender composition) are not pair-differenced: they vary across judges but are constant across the candidate pairs that a given judge evaluates.

All continuous covariates are standardised to mean zero and unit variance before estimation. Identifiers, design variables, and binary judge-level variables (`shareWomen`, `shareWomenCommittee`, `femaleJudge`) are left unstandardised. The stability-selection procedure uses only the behavioural, textual, and presentation features (excluding identifiers and treatment information).

H.2 Abadie-style judge-level heterogeneity procedure

We use the endogenous stratification approach of Abadie, Chingos, and West (2018) to examine heterogeneity in treatment effects by predicted baseline performance. The key idea is to construct a prognostic score—the predicted probability that a female candidate wins absent any treatment effect—and then estimate grouped average treatment effects (GATEs) for observations with low versus high predicted performance. Our implementation exploits the temporal structure of the experiment: because the main reduced-form results show that treatment effects are concentrated in late rounds (rounds 7–12), while early rounds (rounds 1–6) exhibit no detectable treatment effects, we use early-round data to train the prognostic model and apply it out-of-sample to late rounds.

Training data and justification. Within each repetition, the prognostic model is trained on the auxiliary half of early-round observations, pooling across treatment arms (drawing the training sample after pooling on all early-round observations (approximately 1,300 judge-pair decisions) rather than on the control group alone (approximately 325 observations) yields a substantially more stable prognostic model). This choice is justified by the null treatment effects in early rounds: since treatment assignment does not affect outcomes in rounds 1–6 (see Table 3, Panel B), all early-round data is effectively pre-treatment with respect to the fatigue mechanism that generates late-round bias. The early-round null also serves as an internal validation: when the trained model is applied to the held-out early-round sample, it finds no treatment effect heterogeneity, confirming that the prognostic score does not generate spurious patterns where none exist.

Covariate selection. The prognostic model excludes the overall expert evaluation score (`overallExp`), which is the single strongest predictor of `womanWins`. Including it would cause the prognostic grouping to collapse to the expert-favoured subgroup analysis already reported in the reduced-form results. The model retains the expert sub-scores (concept, market potential, growth, finance, plan quality, sense, and presentation), expert assessments of presentation style (appearance, confidence, arrogance, certainty), as well as candidate age, text and language features, enumerator-coded measures, and business content indicators. This covariate set captures a richer, multi-dimensional profile of candidate quality—including specific dimensions of the business proposal and presentation—while avoiding the single summary score that drives the reduced-form heterogeneity results. Because all candidate-level covariates are female-minus-male differences (see Section H.1), the prognostic score predicts relative female success within each pair.

Sample splitting. Within each repetition, the procedure draws a random 50/50 split of early-round candidate pairs. Even though the outcome is at the judge level, the split is performed over unique pair identifiers

to ensure that all observations from a given candidate pair are assigned to the same subsample. The auxiliary half is used to train the prognostic model; the complementary half provides honest out-of-sample early-round estimates.

Model estimation. On the auxiliary subsample, we estimate an adaptive elastic net for the binary outcome `womanWins` using `lassoglm`. The covariate matrix consists of columns 16 onward from the exported ML data file—expert sub-scores, age, text/language features, expert and enumerator presentation assessments, business content indicators, and interpersonal measures (see Table A.21)—excluding only the overall expert score in column 15. We perform a grid search over 25 values of the elastic-net mixing parameter $\alpha \in [0.01, 1]$ and 25 values of the regularisation parameter $\lambda \in [\exp(-10), \exp(-1)]$, using five-fold cross-validation with two Monte Carlo replications to evaluate deviance at each grid point. The (α, λ) pair minimising cross-validated deviance is selected. We then refit the model using only the covariates with non-zero coefficients from the initial regularised fit, estimating an unpenalised logistic regression on the selected variables. This two-step procedure implements an adaptive elastic net that yields more precise coefficient estimates on the retained predictors.

Scoring and grouping. The fitted model is applied to three target samples: (i) the complementary half of early-round pairs (out-of-sample), (ii) the full late-round sample (fully out-of-sample), and (iii) the two combined. Let $\hat{m}_i$ denote the predicted probability that the female candidate wins for observation i. In each target sample, observations are partitioned into two groups at the median of $\hat{m}_i$, producing a low-predicted-performance group and a high-predicted-performance group.

Treatment effect estimation. Within each group g, the grouped average treatment effect for treatment arm t relative to the control group is:

$$\widehat{\tau}_{g,t} = \mathbb{E}[Y_i \mid G_i = g, T_i = t] - \mathbb{E}[Y_i \mid G_i = g, T_i = 1],$$

where Y_i is `womanWins`, $T_i = 1$ denotes the control group, and G_i is the predicted-performance group. Standard errors for the within-group means are clustered at the judge level. Treatment-specific and control-group means within each group are reported alongside the GATEs.

Inference. The procedure is repeated across 100 random pair splits on each of 4 parallel workers, yielding 400 draws. Point estimates are the mean across draws; confidence intervals are the 5th and 95th percentiles of the empirical distribution. This accounts for uncertainty from the random partition.

Interpretation. This procedure is an exploratory heterogeneity exercise rather than a formal structural estimator. It asks whether treatment effects differ systematically between observations that appear weaker or stronger according to a model trained in the pre-fatigue early rounds. The late-round estimates are the primary output, as they are fully out-of-sample relative to the training data. The early-round estimates serve as a placebo check.

H.3 Chernozhukov-style generic machine learning procedure

We complement the Abadie prognostic-score approach with the generic machine learning (GML) framework of Chernozhukov, Demirer, Duflo, and Fernández-Val (2025) to test for and characterise heterogeneity in the effect of organizational values messaging. Whereas the Abadie procedure stratifies by predicted *baseline* performance, the GML framework directly estimates individual-level conditional average treatment effects (CATEs) and tests whether they vary systematically across observations. The procedure yields three outputs: a best linear predictor (BLP) test for heterogeneity, grouped average treatment effects (GATES) that summarize treatment effects by predicted effect size, and classification analysis (CLAN) that describes the characteristics of the most- and least-affected groups.

Treatment comparison. Because organisational values messaging is the active treatment margin, we pool the two arms that include it (treatment arms 2 and 4, both with `info` $= 1$) as the treated group, compared against the control group (treatment arm 1).

Estimation sample. The sample is restricted to late rounds (`roundJudge` > 6), where the treatment effect is concentrated, yielding approximately 900 observations.

Sample splitting. Within each repetition, unique candidate pairs are randomly partitioned into auxiliary and main halves. All observations from a given candidate pair are assigned to the same subsample. The auxiliary half is used to train treatment-arm-specific outcome models; the main half is used for treatment effect estimation. This pair-level split ensures that information from a given candidate pair does not appear in both the training and estimation samples.

Model estimation. On the auxiliary subsample, we estimate separate elastic net logistic regressions for each treatment arm, modelling $\Pr(Y_i = 1 \mid X_i, T_i = t)$ with an ℓ_1/ℓ_2 penalty. The covariate vector X_i consists of the candidate-pair and judge-level variables described in Table A.21, excluding the overall expert score (to avoid collapsing heterogeneity to a single dimension). We search over a grid of 10 values of $\alpha \in [0.01, 1]$ and 10 values of $\lambda \in [\exp(-10), \exp(-1)]$, selecting the pair that minimises three-fold cross-validated deviance (with two Monte Carlo replications). We then refit an unpenalised logistic regression on the covariates with non-zero coefficients from the initial fit (post-selection refit).

Predicted treatment effects. The fitted arm-specific models are applied to the main subsample to obtain predicted outcomes under each treatment arm. The predicted CATE for each observation is $S_i = \hat{m}(X_i, T = \text{treated}) - \hat{m}(X_i, T = \text{control})$, where $\hat{m}$ is the logistic prediction from the relevant arm-specific model.

BLP test for heterogeneity. Following Chernozhukov et al. (2025), we estimate the best linear predictor of the CATE by regressing outcomes on the predicted baseline $\hat{m}(X_i, T = \text{control})$, the predicted CATE, the treatment indicator, and their interaction:

$$Y_i = \alpha_0 + \alpha_1 \hat{m}(X_i, T{=}0) + \tfrac{1}{2}\alpha_2 S_i + \beta_1 (D_i - p) + \beta_2 (D_i - p)(S_i - \bar{S}) + \varepsilon_i,$$

where D_i is the treatment indicator and $p = 2/3$ is the design-implied treatment probability (two pooled arms versus one control arm). The coefficient β_1 captures the average treatment effect; β_2 captures whether predicted heterogeneity translates into actual heterogeneity. A significant β_2 rejects homogeneous effects. Standard errors are clustered at the judge level.

GATES. Observations in the main subsample are sorted by S_i and divided into equally sized groups. In our primary specification we use two groups (a median split); as a robustness check we also report results with three groups (terciles). Within each group, we estimate the average treatment effect relative to control by regressing Y_i on group-specific treatment indicators, controlling for the predicted baseline. Monotonicity across groups is imposed following Chernozhukov et al. (2025).

CLAN. For each covariate, we estimate its mean within each GATES group. Comparing the characteristics of the most-affected group (highest predicted treatment effect) with the least-affected group reveals which observable features distinguish observations that benefit most from the treatment.

Inference. The procedure is repeated 250 times with independent random sample splits. For each split, 90% confidence intervals are constructed as $\hat{\theta} \pm 1.645 \times \hat{\sigma}$, where $\hat{\sigma}$ is the clustered standard error from the split-specific regression. The reported point estimate is the median of $\hat{\theta}$ across splits; the reported confidence bounds are the medians of the split-specific lower and upper bounds, following the variational inference approach in Chernozhukov et al. (2025).

H.4 Machine learning results

H.4.1 Abadie-style heterogeneity results

Table 3 in the main paper presents the Abadie-style grouped treatment effects. The prognostic model is trained on all early-round observations and applied out-of-sample to late rounds (Panel C), the complementary half of early rounds (Panel B), and all rounds combined (Panel A). The model partitions observations at the median of the predicted probability that the female candidate wins.

H.4.2 Chernozhukov-style heterogeneity results

Table A.22 presents the BLP test for treatment effect heterogeneity, Figure A.4 shows the GATES estimates, and Table A.23 characterises the most- and least-affected groups. The BLP does not reject homogeneous treatment effects at the 90% level ($\hat{\beta}_2 = 0.04$, CI $[-0.22, 0.31]$), but the significant loading on the CATE proxy ($\hat{\alpha}_2 = 1.36$, CI $[0.95, 1.76]$) indicates that the machine-learning predictions capture meaningful variation in treatment effects. The GATES estimates are consistent with this: the above-median group experiences a treatment effect of 10.4 percentage points, more than twice the below-median group (4.3 pp). The CLAN analysis (Table A.23) reveals a clear pattern: the most-affected group consists of candidate pairs where the female candidate scores higher on expert assessments of concept, market understanding, and business sense, is rated as more articulate and provides more specific examples, and mentions operational details more frequently—but scores lower on appearance. The treatment appears

to help most when there is a substantive merit case for the female candidate that fatigued judges might otherwise overlook in favour of surface impressions.

	Estimate
α_0 (Constant)	0.103^{**} $[0.003,\ 0.203]$
α_1 (Baseline proxy)	0.742^{**} $[0.537,\ 0.946]$
α_2 (CATE proxy)	1.358^{**} $[0.945,\ 1.764]$
β_1 (ATE)	0.075 $[-0.004,\ 0.154]$
β_2 (Heterogeneity)	0.035 $[-0.219,\ 0.305]$

Notes: This table reports the best linear predictor (BLP) of the conditional average treatment effect on the probability that the female candidate wins, following Chernozhukov et al. (2025). The treated group pools treatment arms 2 (organisational values) and 4 (both treatments); the control group is treatment arm 1. The coefficient β_1 captures the average treatment effect; β_2 captures the loading on predicted heterogeneity, where a significant β_2 rejects the null of homogeneous effects. The estimation sample is restricted to late rounds (rounds 7–12). GATES groups are defined by a median split of predicted treatment effects. Standard errors are clustered at the judge level. Point estimates are medians across 250 random sample splits; 90% confidence intervals in brackets. ** denotes that the 90% confidence interval excludes zero.

Table A.22: Chernozhukov BLP: Test for treatment effect heterogeneity

H.4.3 Tercile-split results

As a robustness check, we repeat the GATES and CLAN analyses using terciles (three groups) rather than a median split. The results are qualitatively similar: the GATES estimates show a monotonic gradient across groups, with the top tercile experiencing the largest treatment effect. The CLAN identifies fewer significant covariate differences with the finer partition, consistent with reduced statistical power from smaller group sizes.

	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**

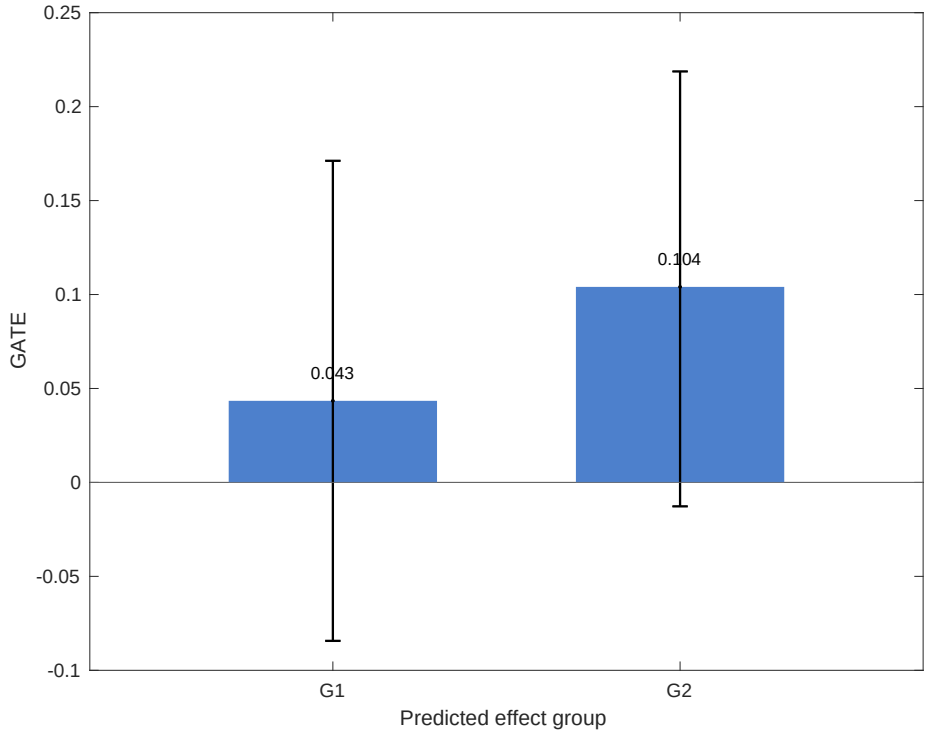
Notes: This table reports mean covariate values within the least-affected (G1) and most-affected (G2) GATES groups from the Chernozhukov et al. (2025) classification analysis (CLAN). All covariates are standardised differences between the female and male candidate in the pair, with positive values indicating 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.

Table A.23: CLAN: Characteristics of most and least affected groups

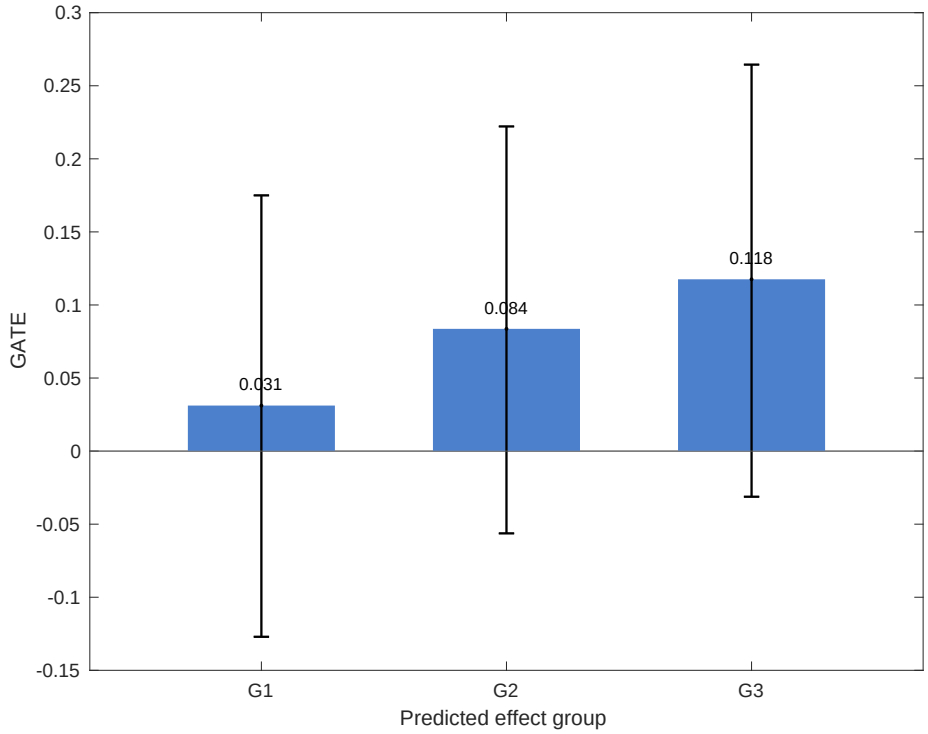
	Estimate
α_0 (Constant)	0.157^{**} $[0.036,\ 0.264]$
α_1 (Baseline proxy)	0.673^{**} $[0.436,\ 0.926]$
α_2 (CATE proxy)	1.103^{**} $[0.710,\ 1.495]$
β_1 (ATE)	0.076 $[-0.010,\ 0.160]$
β_2 (Heterogeneity)	0.104 $[-0.159,\ 0.340]$

Notes: Same specification as Table A.22, with three GATES groups (terciles) rather than a median split. ** denotes that the 90% confidence interval excludes zero.

Table A.24: Chernozhukov BLP: Tercile-split specification



Notes: This figure reports grouped average treatment effects (GATES) on the probability that the female candidate wins, following Chernozhukov et al. (2025). Observations are divided into two equally sized groups at the median of the predicted individual treatment effect S_i. Group 1 contains observations with the lowest predicted effect; Group 2 the highest. Monotonicity is imposed. The treated group pools treatment arms 2 and 4 (organisational values); the control group is treatment arm 1. The estimation sample is restricted to late rounds (rounds 7–12). Error bars show 90% confidence intervals with standard



Notes: Same specification as Figure A.4, with observations divided into three equally sized groups (terciles) by S_i. Group 1 contains observations with the lowest predicted effect; Group 3 the highest.

Figure A.5: GATES: Tercile-split grouped average treatment effects

	G1	G3	Diff
Overall score (expert)	-0.02	0.15	0.17
Concept (expert)	-0.12	0.20	0.32**
Market (expert)	-0.03	0.10	0.14
Growth (expert)	0.09	0.16	0.07
Finance (expert)	0.02	0.10	0.08
Plan (expert)	-0.05	0.14	0.19
Business sense (expert)	0.01	0.14	0.13
Presentation (expert)	-0.06	0.12	0.18
Age	-0.01	-0.01	-0.01
Valence (text)	-0.03	-0.05	-0.02
Engagement (text)	0.09	0.06	-0.03
Agentic language	-0.02	-0.01	0.02
Communal language	-0.08	0.02	0.11
Gender bias (text)	-0.05	0.14	0.18
Appearance (expert)	0.14	-0.17	-0.31**
Confidence (expert)	-0.07	0.15	0.22
Arrogance (expert)	-0.10	-0.03	0.06
Certainty (expert)	0.01	0.16	0.15
Confident, woman only (expert)	-0.04	0.05	0.09
Mentions idea	-0.06	-0.04	0.02
Mentions market	-0.05	-0.08	-0.03
Mentions competition	-0.14	0.06	0.20
Mentions operations	-0.21	0.21	0.42**
Mentions cost	-0.08	-0.09	-0.00
Dressed formally (enum.)	-0.05	0.04	0.09
Care with outfit (enum.)	0.01	0.06	0.05
General appearance (enum.)	0.09	-0.02	-0.11
Confidence (enum.)	-0.19	0.06	0.25**
Arrogance (enum.)	0.10	-0.15	-0.24
Teamwork (enum.)	-0.08	0.11	0.20
Mentions family (enum.)	-0.15	-0.00	0.14
Certainty (enum.)	-0.06	0.12	0.17
Articulate	-0.08	0.06	0.14
Specific examples	-0.13	0.22	0.35**
Gender bias, woman (text)	-0.07	0.09	0.16

Notes: Same specification as Table A.23, with three groups (terciles of predicted treatment effects) rather than a median split. ** denotes that the 90% confidence interval for the G3−G1 difference excludes zero.

Table A.25: CLAN: Tercile-split characteristics

I Sequential Decision Patterns

A natural concern is that the emergence of gender bias in late rounds reflects sequential decision patterns rather than cognitive fatigue. Two well-documented mechanisms could generate this pattern: the *gambler's fallacy*, whereby decision-makers overcorrect after streaks of similar decisions (Chen, Moskowitz, & Shue, 2016), and *sequential contrast effects*, whereby the quality of the preceding case distorts the assessment of the current one (Radbruch & Schiprowski, 2025). We present a series of tests showing that the late-round treatment interaction is not materially attenuated by controls for prior decision patterns.

Empirical approach. Table A.26 reports our main specification—treatment arms interacted with a late-round indicator (rounds 7–12)—augmented with progressively richer controls for sequential decision patterns. Column 1 reproduces the baseline result. Column 2 adds an indicator for whether the judge voted for the female candidate in the previous observed assessment, following the baseline specification in Chen et al. (2016): $Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \text{Controls} + \varepsilon_{it}$. Under the gambler's fallacy, $\beta_1 < 0$, implying judges who just voted for a woman are less likely to do so again. Column 3 adds the cumulative count of prior votes for women (excluding the current round), analogous to the moving-average controls that Chen et al. (2016) use to absorb decision-maker heterogeneity without the Nickell bias induced by individual fixed effects. Column 4 adds the lagged expert score difference of the preceding pair, which captures the sequential contrast effect channel emphasised by Radbruch and Schiprowski (2025): if a high-quality female candidate in the previous pair creates a contrast that depresses the current female candidate's chances, this variable should absorb it.

Streaks test. Column 5 implements the streaks test from Chen et al. (2016, Section 2.2), which discriminates between competing models of sequential decision-making. We include indicators $I(Y_{t-2}, Y_{t-1})$ for the two previous decisions—$I(1,1)$, $I(0,1)$, and $I(1,0)$—with two consecutive votes against the female candidate $I(0,0)$ as the omitted category. The gambler's fallacy predicts $\beta_{(1,1)} < \beta_{(0,1)} < \beta_{(1,0)} < 0$: longer streaks of affirmative decisions should produce stronger reversals because they are perceived as increasingly unlikely under random variation. By contrast, a pure sequential contrast effect predicts that only the most recent case matters, so $\beta_{(1,1)} \approx \beta_{(0,1)}$. Column 6 includes all sequential controls simultaneously.

Results. The organisational values $\times$ late-round interaction coefficient is remarkably stable across all six specifications, ranging from 0.117 to 0.140 (Table A.26). None of the sequential control variables are individually significant at conventional levels, with the exception of a weakly significant lagged score difference (-0.005, $p < 0.10$) in Column 4, which loses significance entirely when combined with other controls in Column 6. The lagged vote coefficient is small and insignificant (-0.015, SE $= 0.020$), providing no evidence of gambler's fallacy. The cumulative prior votes coefficient is similarly null (0.005, SE $= 0.007$), providing no indication of quota-like behaviour. In the streaks test (Column 5), all three streak indicators are small, insignificant, and similar in magnitude (-0.021 to -0.035), inconsistent with the gambler's fallacy prediction of $\beta_{(1,1)} < \beta_{(0,1)} < \beta_{(1,0)}$.

Pair quality autocorrelation. A prerequisite for attributing sequential patterns to behavioral bias rather than quality ordering is that the sequence of pairs evaluated by each judge is not systematically correlated

with pair quality (Chen et al., 2016, Table A.VI). In our experiment, the assignment of pairs to positions within a session is random, so any autocorrelation in pair quality across the sequence has occurred purely by chance and is by construction orthogonal to treatment assignment. Consistent with this, simple regressions without pair fixed effects reveal only mild and negligible autocorrelation in lagged score differences. Our pair fixed effects absorb the quality of the current pair, and explicitly conditioning on lagged pair quality in Column 4 does not materially change the treatment $\times$ late-round interaction, confirming that the ordering of pair quality does not drive our results.

Comparison with the contrast effects literature. Two features of our design and results are difficult to reconcile with a sequential contrast effects interpretation. First, Radbruch and Schiprowski (2025) find that contrast effects *weaken* over the evaluation sequence: the autocorrelation in their admission process drops from approximately 10 percentage points in early slots to 3 percentage points in slots 10–12 (their Figure 4). Our finding is the opposite—bias is absent in early rounds and emerges only in late rounds—which is inconsistent with a contrast effect that should be strongest when the previous candidate's memory is most salient. Second, our design requires judges to choose between a male and female candidate presented *simultaneously*, unlike the sequential single-candidate evaluations in Radbruch and Schiprowski (2025). A contrast effect in our setting would require gender-specific comparison with the previous pair's female candidate, without symmetric adjustment for the male candidate—a mechanism that the existing literature does not document.

	(1) Baseline	(2) + Lag vote	(3) + Cum. votes	(4) + Lag score	(5) Streaks	(6) All seq.
Org. Values	-0.019	-0.020	-0.019	-0.019	0.003	-0.018
	(0.043)	(0.046)	(0.042)	(0.046)	(0.054)	(0.046)
Self-Image	-0.005	0.039	-0.006	0.040	0.031	0.039
	(0.040)	(0.044)	(0.040)	(0.043)	(0.051)	(0.043)
Org. Values & Self-Image	-0.033	-0.001	-0.033	-0.000	0.009	-0.001
	(0.042)	(0.045)	(0.042)	(0.044)	(0.053)	(0.044)
Second half=1	-0.105**	-0.095**	-0.118**	-0.094**	-0.089*	-0.108**
	(0.045)	(0.045)	(0.047)	(0.044)	(0.051)	(0.047)
Org. Values × Second half=1	0.137**	0.140**	0.134**	0.138**	0.117*	0.137**
	(0.057)	(0.060)	(0.058)	(0.059)	(0.067)	(0.060)
Self-Image × Second half=1	0.055	0.015	0.053	0.016	0.023	0.012
	(0.054)	(0.055)	(0.054)	(0.054)	(0.061)	(0.055)
Org. Values & Self-Image ×	0.101*	0.069	0.099*	0.069	0.058	0.067
Second half=1	(0.057)	(0.058)	(0.058)	(0.057)	(0.063)	(0.058)
Voted for woman (t−1)		-0.015				-0.013
		(0.020)				(0.023)
Cum. prior votes for women			0.005			0.007
			(0.007)			(0.008)
Lagged score diff. (t−1)				-0.005*		0.000
				(0.003)		(0.005)
Woman favoured (t−1)						-0.045
						(0.036)
I(1,1): two votes for woman					-0.021	
					(0.031)	
I(0,1): against then for					-0.035	
					(0.029)	
I(1,0): for then against					-0.032	
					(0.028)	
Control mean (early)	0.486	0.470	0.486	0.470	0.475	0.470
N	2,649	2,404	2,649	2,404	2,159	2,404
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports OLS estimates of treatment effects on the probability that the female candidate wins, interacted with a late-round indicator (rounds 7–12). Column 1 reproduces the baseline specification. Columns 2–4 progressively add controls for sequential decision patterns: the judge's vote in the previous observed assessment (Col. 2), the cumulative number of prior votes for female candidates (Col. 3), and the lagged expert score difference of the preceding pair (Col. 4). Column 5 follows Chen et al. (2016) by including streak indicators $I(Y_{t-2}, Y_{t-1})$ for the two previous decisions; the omitted category is two consecutive votes against the female candidate. Under the gambler's fallacy, the coefficient on $I(1,1)$ should be more negative than $I(0,1)$, which should be more negative than $I(1,0)$; under pure contrast effects, only the most recent decision matters, so $I(1,1) \approx I(0,1)$. Column 6 includes all sequential controls simultaneously. The omitted treatment category is Control; the omitted round category is early rounds (1–6). Columns 2, 4, and 6 lose approximately 245 observations (one per judge) because lagged variables are undefined for each judge's first observed assessment; Column 3 retains the full sample; Column 5 loses approximately 490 observations due to the second lag. 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 A.26: Robustness: Controlling for sequential decision patterns

I.1 Robustness: candidate presentation order

Each pair of candidates is presented to the committee in a randomly determined order: independently for each triplet that assesses it (although the order is fixed within a triplet), the platform shuffles which of the male and female candidate is shown first. The unconditional rate at which the first-shown candidate is selected is approximately 51%.

Table A.27 reports treatment effects on the probability that the first-shown candidate is selected, conditional on pair fixed effects. Pooled coefficients (columns (1) and (2)) average across two within-pair forces of opposite sign: in pairs where the female is shown first, any treatment-induced shift toward the female candidate raises selection of the first-shown candidate; in pairs where the male is shown first, the same shift lowers it. Columns (3) and (4) decompose these by conditioning on the gender of the first-shown candidate, isolating the treatment effect on within-order selection. The combined Org. Values $\times$ Self-Image arm raises selection of the first-shown female by 10.4 percentage points (column (4)) but does not change selection of the first-shown male, indicating that treatment-induced changes in gender preference operate in the female-first sub-sample without an offsetting movement when the male is shown first.

	First candidate wins			
	All rounds	By late round	Male shown first	Female shown first
Org. Values	0.029	0.038	-0.025	0.091**
	(0.028)	(0.038)	(0.045)	(0.043)
Self-Image	0.019	0.034	-0.033	0.063
	(0.027)	(0.036)	(0.055)	(0.048)
Org. Values & Self-Image	0.079***	0.110***	0.013	0.104**
	(0.025)	(0.037)	(0.047)	(0.048)
Late Rounds		0.024		
		(0.042)		
Org. Values $\times$ Late Rounds		-0.019		
		(0.054)		
Self-Image $\times$ Late Rounds		-0.029		
		(0.054)		
Org. Values & Self-Image $\times$ Late Rounds		-0.062		
		(0.057)		
Pair FE	Yes	Yes	Yes	Yes
Control mean	0.510	0.510	0.562	0.471
N	2,649	2,649	1,275	1,374

Notes: The outcome is an indicator equal to one if the first-shown candidate of a pair is selected by the judge. Column (1) pools all rounds; column (2) interacts treatment with a late-round indicator (rounds 7–12); columns (3) and (4) restrict the sample to pairs where the male and female candidate is shown first, respectively. All specifications include pair fixed effects and standard errors clustered at the judge level. Pooled coefficients in columns (1)–(2) average across two within-pair forces of opposite sign and should be interpreted as asymmetric treatment effects by candidate-order rather than primacy effects in levels. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.27: Treatment effects on selection of the first-shown candidate

Table A.28 examines whether the treatment effects on the probability of a female candidate winning dif-

fer by which candidate is shown first. Column (1) reports the full-sample interaction, column (2) restricts to late rounds (where the main result is concentrated), and column (3) reports the triple interaction with round timing. The interaction of the combined treatment with first-shown-female is positive and marginally significant in the pooled specification (10.5 percentage points), suggesting that the treatment effects on female selection are slightly larger when the female candidate is presented first; the late-round and triple-interaction columns are imprecise. The level treatment effects (rows for each treatment without interaction) remain consistent with the headline result reported in the main text.

	Woman wins		
	All rounds	Late rounds only	All rounds
Org. Values	0.011	0.102*	-0.076
	(0.041)	(0.057)	(0.058)
Self-Image	0.019	0.056	-0.018
	(0.046)	(0.063)	(0.063)
Org. Values & Self-Image	-0.034	0.035	-0.102*
	(0.042)	(0.061)	(0.058)
Woman shown first	-0.012	0.015	-0.030
	(0.043)	(0.068)	(0.054)
Org. Values × Woman shown first	0.070	0.020	0.105
	(0.054)	(0.082)	(0.071)
Self-Image × Woman shown first	0.004	-0.021	0.021
	(0.060)	(0.096)	(0.074)
Org. Values & Self-Image × Woman shown first	0.105*	0.096	0.139*
	(0.057)	(0.090)	(0.075)
Late Rounds			-0.131**
			(0.060)
Org. Values × Late Rounds			0.177**
			(0.075)
Self-Image × Late Rounds			0.082
			(0.075)
Org. Values & Self-Image × Late Rounds			0.142*
			(0.074)
Late Rounds × Woman shown first			0.044
			(0.073)
Org. Values × Late Rounds × Woman shown first			-0.072
			(0.101)
Self-Image × Late Rounds × Woman shown first			-0.047
			(0.098)
Org. Values & Self-Image × Late Rounds × Woman shown first			-0.078
			(0.101)
Pair FE	Yes	Yes	Yes
Control mean (male-first)	0.438	0.383	0.438
N	2,649	1,298	2,649

Notes: The outcome is an indicator equal to one if the female candidate of the pair wins. Column (1) interacts treatment with an indicator for the female candidate being shown first (full sample). Column (2) restricts the sample to late rounds (rounds 7–12). Column (3) reports the full triple interaction of treatment, late-round, and first-shown-female. All specifications include pair fixed effects and standard errors clustered at the judge level. The control mean refers to the male-shown-first sub-sample. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.28: Treatment-effect heterogeneity on female selection by candidate-order

J Treatment Effect Heterogeneity by Judge Gender

This section examines whether the main treatment effects differ by the gender of the evaluating judge. We estimate the treatment $\times$ timing interaction separately for male and female judges:

$$Y_{jcp} = \alpha + \sum_k \beta_k \,\text{Treatment}_k + \gamma \,\text{Late round} + \sum_k \delta_k \,(\text{Treatment}_k \times \text{Late round}) + \mu_p + \varepsilon_{jcp} \quad \text{(J.19)}$$

where j indexes judges, c committees, and p candidate pairs; μ_p denotes pair fixed effects. We estimate Equation (J.19) separately for male judges (gender $= 1$) and female judges (gender $= 2$), clustering standard errors at the judge level.

Table A.29 presents results using the binary second-half indicator (rounds 7–12). The fatigue-induced decline in female candidate selection and its prevention by the organisational values treatment are concentrated among male judges. For male judges, the organisational values treatment effect in late rounds is large and significant ($p = 0.005$ for female wins, $p < 0.001$ for expert-favoured female wins). Female judges show no significant treatment effects in either early or late rounds.

Table A.30 replaces the binary second-half indicator with the continuous evaluation round number:

$$Y_{jcp} = \alpha + \sum_k \beta_k \,\text{Treatment}_k + \gamma \,\text{Round} + \sum_k \delta_k \,(\text{Treatment}_k \times \text{Round}) + \mu_p + \varepsilon_{jcp} \quad \text{(J.20)}$$

The continuous-round specification confirms that the organisational values treatment effect for male judges grows monotonically with evaluation round, while no such pattern exists for female judges.

	(1) Woman Wins	(2) Expert-favoured Woman Wins	(3) Expert-favoured Man Wins
Org. Values	-0.033	-0.037	-0.010
	(0.047)	(0.031)	(0.032)
Social Image	0.019	-0.005	-0.027
	(0.047)	(0.029)	(0.035)
Org. Values & Social Image	-0.032	-0.054*	-0.018
	(0.049)	(0.030)	(0.035)
Late Round	-0.112**	-0.100***	0.021
	(0.052)	(0.030)	(0.037)
Org. Values × Late Round	0.169**	0.164***	-0.010
	(0.065)	(0.043)	(0.045)
Social Image × Late Round	0.040	0.061	0.010
	(0.061)	(0.038)	(0.044)
OV & SI × Late Round	0.112*	0.118***	-0.003
	(0.064)	(0.043)	(0.048)
Female judge	-0.001	-0.009	-0.018
	(0.076)	(0.048)	(0.048)
Org. Values × Female	0.070	0.099	0.022
	(0.113)	(0.066)	(0.072)
Social Image × Female	-0.092	-0.035	0.073
	(0.095)	(0.064)	(0.064)
OV & SI × Female	-0.008	0.069	0.057
	(0.092)	(0.059)	(0.062)
Late Round × Female	0.022	0.054	0.014
	(0.105)	(0.070)	(0.067)
Org. Values × Late Round × Female	-0.156	-0.180*	-0.008
	(0.154)	(0.103)	(0.095)
Social Image × Late Round × Female	0.054	0.013	-0.027
	(0.125)	(0.089)	(0.079)
OV & SI × Late Round × Female	-0.044	-0.117	-0.037
	(0.136)	(0.088)	(0.087)
Pair FE	Yes	Yes	Yes
Control mean (early, male)	0.485	0.262	0.322
p: Org. Values late (male)	0.005	0.000	0.554
p: Social Image late (male)	0.191	0.062	0.604
p: Combined late (male)	0.101	0.050	0.566
p: Org. Values late (female)	0.591	0.563	0.881
p: Social Image late (female)	0.777	0.571	0.529
p: Combined late (female)	0.762	0.791	0.999
N	2,649	2,649	2,649

Notes: OLS estimates of treatment effects on committee decision outcomes, interacted with evaluation timing, separately by judge gender. The dependent variable in columns 1 is an indicator for the female candidate being selected; column 2 indicates whether an expert-favoured woman is selected; column 3 indicates whether an expert-favoured man is selected. Treatment indicators are each interacted with *Second half*, an indicator for rounds 7–12, and the judges' gender. "Control mean (early, male)" reports the dependent variable mean for men in the control group in rounds 1–6. All specifications include candidate-pair fixed effects. Standard errors clustered at the judge level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.29: Treatment effects by evaluation timing and judge gender (binary)

	(1) Female Wins	(2) Male Expert Pick Wins	(3) Female Expert Pick Wins
Org. Values	-0.090	-0.008	-0.095**
	(0.069)	(0.050)	(0.046)
Social Image	0.009	-0.034	-0.017
	(0.068)	(0.052)	(0.042)
Org. Values & Social Image	-0.089	-0.005	-0.100**
	(0.070)	(0.053)	(0.045)
Round	-0.013*	0.003	-0.011**
	(0.007)	(0.005)	(0.004)
Org. Values × Round	0.022**	-0.001	0.022***
	(0.009)	(0.007)	(0.006)
Social Image × Round	0.005	0.002	0.007
	(0.009)	(0.006)	(0.006)
OV & SI × Round	0.018*	-0.002	0.016***
	(0.009)	(0.007)	(0.006)
Female judge	-0.005	-0.014	-0.024
	(0.114)	(0.076)	(0.070)
Org. Values × Female	0.078	0.064	0.169*
	(0.171)	(0.121)	(0.091)
Social Image × Female	-0.156	0.100	-0.066
	(0.142)	(0.092)	(0.093)
OV & SI × Female	0.041	0.033	0.133
	(0.138)	(0.094)	(0.084)
Round × Female	0.002	0.000	0.007
	(0.015)	(0.010)	(0.010)
Org. Values × Round × Female	-0.013	-0.007	-0.025*
	(0.023)	(0.016)	(0.015)
Social Image × Round × Female	0.014	-0.006	0.006
	(0.019)	(0.012)	(0.013)
OV & SI × Round × Female	-0.011	0.001	-0.019
	(0.020)	(0.013)	(0.012)
Pair FE	Yes	Yes	Yes
Control mean (early, male)	0.485	0.322	0.262
N	2,649	2,649	2,649

Notes: Same specification as Table A.29, replacing the binary second-half indicator with the continuous evaluation round number (1–12). See Table A.29 notes for full details. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.30: Treatment effects by evaluation round and judge gender (continuous)

K Long-term labour market outcomes

Around 18 months following the competition, we collect follow-up data to assess what happened to the candidates after the competition. We collect data from 89 of the 100 candidates on their employment status, labour market outcomes, and some questions relating to their current business (plans). To understand whether judges recommended different types of candidates across treatments, we run the following regression pooling the treatments, where the observation is the labour market outcomes of individual i:

$$\begin{aligned} Y_i = &\alpha + \beta_1 \texttt{Score\_Control}_i + \beta_2 \texttt{Score\_Organisational\_Values}_i + \\ &\beta_3 \texttt{Score\_Social\_Image}_i + \beta_4 \texttt{Score\_Both\_treatments}_i + \\ &\zeta \texttt{Winner}_i + \varepsilon_i \end{aligned}$$

Where `Score_X` is the number of pairwise comparisons a candidate won in treatment arm X and `Winner` an indicator for having won the prize. Table A.31 describes the results from this regression,[3] including in Y an indicator for transitioning out of self-employment, self-employment, wage-employment, total income, total investment in their business since the competition, the number of steps they have taken to establish a business and whether or not they have taken out a business loan since the business. These measures aim to broadly capture the labour-market outcomes of the young professionals following the competition.

[3] We exclude the transition into self-employment as a dependent variable due to the small number of respondents that do so.

Table A.31: Long-term outcomes and competition performance

	Transition into self-employment	Self-employed	Wage-employed	Total income (ETB1000)	Total investment (ETB1000)	Total steps taken	Has loan
Score control	0.026*	-0.005	0.004	0.018	0.027	0.057	-0.023
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.22)	(0.02)
Score org. values	0.006	0.018	0.007	-0.004	0.025	0.123	0.030
	(0.02)	(0.02)	(0.02)	(0.05)	(0.07)	(0.27)	(0.02)
Score social image	-0.039**	-0.045*	-0.002	-0.009	-0.064	-0.170	-0.014
	(0.02)	(0.02)	(0.02)	(0.05)	(0.07)	(0.31)	(0.02)
Score both treatments	0.007	0.018	-0.005	0.049	0.017	0.009	-0.025
	(0.01)	(0.02)	(0.02)	(0.06)	(0.08)	(0.30)	(0.02)
Winner	0.048	0.206	0.125	-0.056	0.315	0.766	0.407**
	(0.13)	(0.20)	(0.16)	(0.47)	(0.61)	(2.23)	(0.20)
Sample mean	0.135	0.371	0.753	30.604	158.348	6.258	0.348
N	89	89	89	89	89	89	89

Notes: This table describes the relationship between labor market outcomes and candidate performance in each of the treatment arms. The transition into self-employment is a dummy equal to one if the respondent was not in self-employment before but is after the competition. Self-employed and wage-employed are dummies. Total income is monthly income out of any source (in ETB 1000). Total investment is the amount the respondent has invested in their business since the competition (in ETB 1000), estimated using Poisson regression. Total steps taken is the number of steps (from a fixed list) a respondent has taken starting a business. Has loan is an indicator for whether the respondent has a business loan. All monetary variables are winsorized at the 95th percentile. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.32: Long-term outcomes and competition performance by round period

	Transition into self-employment	Self-employed	Wage-employed	Total income (ETB1000)	Total investment (ETB1000)	Total steps taken	Has loan
				Panel A: Early rounds			
Score control	0.012	-0.003	-0.001	0.009	-0.003	0.129	-0.017
	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)	(0.16)	(0.01)
Score org. values	0.002	0.019	-0.002	0.035	0.043	0.197	0.025
	(0.01)	(0.02)	(0.01)	(0.03)	(0.04)	(0.17)	(0.02)
Score social image	-0.017	-0.013	0.006	-0.003	-0.016	-0.126	-0.008
	(0.01)	(0.02)	(0.02)	(0.04)	(0.06)	(0.19)	(0.01)
Score both treatments	-0.006	-0.008	-0.005	0.002	-0.025	-0.110	-0.022*
	(0.01)	(0.01)	(0.01)	(0.04)	(0.05)	(0.15)	(0.01)
Winner	0.052	0.188	0.142	-0.013	0.337	0.725	0.422**
	(0.13)	(0.20)	(0.16)	(0.48)	(0.58)	(2.12)	(0.20)
Sample mean	0.133	0.367	0.756	30.412	156.589	6.189	0.344
N	89	89	89	89	89	89	89
				Panel B: Late rounds			
Score control	0.013	-0.012	0.001	0.020	0.035	-0.151	-0.020
	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)	(0.16)	(0.01)
Score org. values	-0.005	-0.007	0.005	-0.032	-0.003	-0.143	-0.003
	(0.01)	(0.02)	(0.01)	(0.04)	(0.05)	(0.19)	(0.02)
Score social image	-0.016	-0.027*	-0.008	0.005	-0.048	-0.016	-0.007
	(0.01)	(0.02)	(0.02)	(0.03)	(0.05)	(0.20)	(0.01)
Score both treatments	0.003	0.025	0.006	0.046	0.028	0.190	0.003
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.20)	(0.02)
Winner	0.045	0.241	0.126	0.067	0.328	1.142	0.387*
	(0.14)	(0.19)	(0.14)	(0.44)	(0.57)	(2.12)	(0.20)
Sample mean	0.133	0.367	0.756	30.412	156.589	6.189	0.344
N	89	89	89	89	89	89	89

Notes: This table describes the relationship between labor market outcomes and candidate performance in each of the treatment arms, separately for early rounds (1–6) and late rounds (7–12). The transition into self-employment is a dummy equal to one if the respondent was not in self-employment before but is after the competition. Self-employed and wage-employed are dummies. Total income is monthly income out of any source (in ETB 1000). Total investment is the amount the respondent has invested in their business since the competition (in ETB 1000), estimated using Poisson regression. Total steps taken is the number of steps (from a fixed list) a respondent has taken starting a business. Has loan is an indicator for whether the respondent has a business loan. All monetary variables are winsorized at the 95th percentile. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The results, presented in Table A.31, show that overall, performance in the competition has limited predictive power for long-term labor market outcomes. This finding aligns with previous studies, such as McKenzie and Sansone (2019), which suggest that accurately predicting the future success of aspiring entrepreneurs based on competition performance is inherently difficult.

Despite the limited predictive power across most outcomes, there are a few notable patterns. Candidates who performed well in the control treatment (i.e. received high scores in the absence of any organisational or social pressure) are significantly more likely to transition into self-employment. Interestingly, those who scored well amongst control group judges also tended to invest more in their businesses. This suggests that the control group judges were slightly better at identifying candidates likely to grow their businesses. In contrast, candidates who performed well amongst judges in the social-image treatment are less likely to transition into self-employment and to be in self-employment.

Overall, the results indicate that while competition performance may not be a strong predictor of long-term success in various labor market outcomes, certain treatment-specific effects do emerge. High performers in the control group appear to have moved away from self-employment while those remaining invest relatively more in business activities. Conversely, those who do well in the social-image treatment are more likely to remain self-employed, though their ventures are less financially successful. Given the large number of regressions done here these results are unlikely to survive rigorous multiple-hypothesis testing.

References

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

Bursztyn, L., Egorov, G., Haaland, I., Rao, A., & Roth, C. (2023). Justifying dissent. *The Quarterly Journal of Economics*, *138*(3), 1403–1451.

Chen, D. L., Moskowitz, T. J., & Shue, K. (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.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernández-Val, I. (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. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19303 doi: https://doi.org/10.3982/ECTA19303

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

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

Glick, P., & Fiske, S. T. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of women quarterly*, *21*(1), 119–135.

McKenzie, D., & Sansone, D. (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in nigeria. *Journal of Development Economics*, *141*, 102369. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304387818305601 doi: https://doi.org/10.1016/j.jdeveco.2019.07.002

Prendergast, C. (1993). A theory of "yes men"'. *The American Economic Review*, 757–770.
Radbruch, J., & Schiprowski, A. (2025). Interview sequences and the formation of subjective assessments. *Review of Economic Studies*, *92*(2), 1226–1256.