The Dynamics of Discrimination: Theory and Evidence

Online Appendix

By J. Aislinn Bohren and Alex Imas and Michael Rosenberg*

September 2019

This Online Appendix contains the additional theoretical and empirical analysis referenced in the body of the paper.

A Alternative Ability Distributions

In this section, we consider alternative continuous ability distributions and a model with binary ability and quality. We show numerically that a belief reversal does not occur in the correctly specified model, and therefore, an analogue of Proposition 2 in the manuscript holds for these alternative ability distributions.

A.1 Continuous Ability Distributions

Suppose a worker has ability distributed according to $a \sim f_{a,\mu}(a)$ with parameter μ, and assume that the family of distributions $\{f_{a,\mu}(a)\}_{\mu\in\mathbb{R}}$ satisfies the monotone likelihood ratio property in μ. We will show that if there is a single type of evaluator with prior belief that males have a higher parameter μ than females, i.e. $\hat{\mu}_M > \hat{\mu}_F$ and it is common knowledge that all evaluators share these prior beliefs, then both the first and the second period evaluations are higher for males. In other words, no discrimination reversal occurs between the first and second period.

As before, each task has hidden quality $q_t = a + \varepsilon_t$, where $\varepsilon_t \sim N(0, 1/\tau_\varepsilon)$. Evaluator t observes the evaluations on past tasks and signal $s_t = q_t + \eta_t$ of the quality of the current task, where $\eta_t \sim N(0, 1/\tau_\eta)$, then reports evaluation $v(h_t, s_t, \mu) \equiv E_\mu[q|h_t, s_t]$.

*Bohren: University of Pennsylvania, 133 South 36th Street, Philadelphia, PA 19104, abohren@sas.upenn.edu. Imas: Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, aimas@andrew.cmu.edu. Rosenberg: CarGurus, 2 Canal Park, Cambridge, MA 02141, rosenberg.michael.m@gmail.com.

We first show that the first period evaluation is increasing in μ. The prior distribution of quality, denoted $f_{q,\mu}(q)$, is the convolution of the normally distributed prior distribution of ability and a normally distributed error term. The MLRP is preserved under convolution with a log-concave density and the normal distribution is log-concave. Therefore, the prior distribution of quality also satisfies the MLRP in μ. Suppose that the evaluator observes signal s_1. Then the posterior belief about quality conditional on signal s_1 is

$$f_{q,\mu}(q|s_1) = \frac{f_s(s_1|q)f_{q,\mu}(q)}{\int_Q f_s(s_1|q)f_{q,\mu}(q)dq}. \tag{1}$$

The MLRP is preserved under Bayesian updating when the likelihood function is independent of μ. Since $f_s(s_1|q)$ is independent of μ, the posterior $f_{q,\mu}(q|s_1)$ satisfies the MLRP in μ. By FOSD, $v(h_1, s_1, \mu) = E_\mu[q|s_1]$ is increasing in μ.

Note that the initial evaluation $v(h_1, s_1, \mu)$ is strictly increasing in s_1, and therefore, each signal s_1 maps to a unique evaluation. Further,

$$\begin{aligned} v(h_1, s_1, \mu) &= E_\mu[q|s_1] \\ &= \int_Q q f_{q,\mu}(q|s_1)dq \\ &= \frac{\int_Q \int_A q f_s(s_1|q)f_{q,\mu}(q|a)f_{a,\mu}(a)dadq}{\int_Q f_s(s_1|q)f_{q,\mu}(q)dq} \end{aligned} \tag{2}$$

where the third line follows from (1) and $f_{q,\mu}(q) = \int_A f_{q,\mu}(q|a)f_{a,\mu}(a)da$. Let $s(v, \mu)$ be the signal required to receive initial evaluation v, i.e. the solution to

$$v = \frac{\int_Q \int_A q f_s(s(v,\mu)|q)f_{q,\mu}(q|a)f_{a,\mu}(a)dadq}{\int_Q f_s(s(v,\mu)|q)f_{q,\mu}(q)dq}. \tag{3}$$

We next characterize how the evaluator in period two updates her belief about ability following history $h_2 = (v_1)$. Consider an evaluator with prior $f_{a,\mu}(a)$ who believes the first period evaluator had the same prior. The distribution of ability conditional on h_2 is

$$f_{a,\mu}(a|h_2) = \frac{f_s(s(v_1,\mu)|a)f_{a,\mu}(a)}{\int_A f_s(s(v_1,\mu)|a)f_{a,\mu}(a)da}. \tag{4}$$

Suppose that $f_{a,\mu}(a|h_2)$ satisfies the MLRP in μ. Then by the same reasoning as above, $f_{q,\mu}(q|h_2, s_2)$ satisfies the MLRP in μ and the second period evaluation $v(h_2, s_2, \mu) = E_\mu[q|h_2, s_2]$ is increasing in μ. Therefore, if the first and second period evaluators have common belief $\hat{\mu}_M > \hat{\mu}_F$, both the first and the

second period evaluations are higher for males, $v(h_1, s_1, \hat{\mu}_M) > v(h_1, s_1, \hat{\mu}_F)$ and $v(h_2, s_2, \hat{\mu}_M) > v(h_2, s_2, \hat{\mu}_F)$, and no discrimination reversal occurs.

By the above analysis, establishing that $f_{a,\mu}(a|h_2)$ satisfies the MLRP in μ is sufficient to rule out a reversal between the first and second period. This is equivalent to showing $\frac{\partial^2}{\partial\mu\partial a}\log f_{a,\mu}(a|h_2) > 0$ for all a, μ, h_2. Given that the denominator $\int_A f_s(s(v_1,\mu)|a) f_{a,\mu}(a)da$ is independent of a, this is equivalent to

$$\begin{aligned} & \frac{\partial^2}{\partial\mu\partial a}\left(\log f_s(s(v_1,\mu)|a) + \log f_{a,\mu}(a)\right) > 0 \\ i.e. \quad & \frac{\partial^2}{\partial\mu\partial a}\left(-\frac{\tau_\varepsilon\tau_\eta}{2(\tau_\varepsilon+\tau_\eta)}((s(v_1,\mu)-a)^2 + \log f_{a,\mu}(a)\right) > 0 \\ i.e. \quad & \frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon+\tau_\eta}\frac{\partial s(v_1,\mu)}{\partial\mu} + \frac{\partial^2}{\partial\mu\partial a}\log f_{a,\mu}(a) > 0 \end{aligned} \tag{5}$$

where the second line follows from the signal distribution $s|a \sim N(a, \frac{\tau_\varepsilon+\tau_\eta}{\tau_\varepsilon\tau_\eta})$. The first term $\frac{\partial s(v_1,\mu)}{\partial\mu}$ is negative, since when μ is higher, a lower signal is required to receive a given evaluation. The second term $\frac{\partial^2}{\partial\mu\partial a}\log f_{a,\mu}(a)$ is positive, since by assumption $f_{a,\mu}(a)$ satisfies the MLRP.

We numerically show that (5) holds for several classes of distributions by (i) numerically solving (3) for $s(v_1,\mu)$, and (ii) numerically calculating $\frac{\partial s(v_1,\mu)}{\partial\mu}$.

Exponential Distribution. The exponential distribution has density $f_{a,\mu}(a) = \frac{1}{\mu}e^{-a/\mu}$, where $a \in [0,\infty)$ and $E[a] = \mu$. Therefore, $\frac{\partial^2}{\partial\mu\partial a}\log f_{a,\mu}(a) = 1/\mu^2 > 0$ and the prior distribution satisfies the MLRP in μ. We show that (5) holds numerically for all parameters $\mu \in \{.01, .02, ..., 2.99, 3\}$ and $v \in \{-2, -1.99, ..., 5.99, 6\}$. Given that $\frac{\partial^2}{\partial\mu\partial a}\log f_{a,\mu}(a)$ is independent of a, (5) is also independent of a and the simulation holds for all $a \in [0,\infty]$. This numerically rules out a reversal when the prior distribution of ability follows the exponential distribution. See the Supplemental Appendix for the Matlab code to generate this simulation.

Beta Distribution. The beta distribution has density $f_{a,\alpha}(a) = \frac{1}{B(\alpha,\beta)}a^{\alpha-1}(1-a)^{\beta-1}$, where $a \in [0,1]$ and $E[a] = \alpha/(\alpha+\beta)$. Therefore, $\frac{\partial^2}{\partial\alpha\partial a}\log f_{a,\alpha}(a) = 1/a > 0$ and the prior distribution satisfies the MLRP in α. Letting μ correspond to α, note that for any β, the expected ability is increasing in α. We show that (5) holds numerically for all parameters $\alpha \in \{1, 1.05, ..., 2.95, 3\}$, $\beta \in \{1.5, 2, 2.5\}$, $a \in \{.02, .04, ..., .96, .98\}$ and $v \in \{-2, -1.98, ..., 2.98, 3\}$. This numerically rules out a reversal when the prior distribution of ability follows the beta distribution. See the Supplemental Appendix for the Matlab code to generate this simulation.

Gamma Distribution. The gamma distribution has density $f_{a,k}(a) = \frac{1}{\Gamma(k)\theta^k}a^{k-1}e^{-a/\theta}$, where $a \in (0,\infty)$ and $E[a] = k\theta$. Therefore, $\frac{\partial^2}{\partial k\partial a}\log f_{a,k}(a) = 1/a > 0$ and the

prior distribution satisfies the MLRP in k. Letting μ correspond to k, note that for any θ, the expected ability is increasing in k. This case is slightly different, as (5) does not hold for all a, k, θ and v. Since (5) is sufficient, but not necessary, for a reversal, we can also show that $E_k[a|h_2]$ is increasing in k, i.e. the posterior average ability is increasing in the parameter of interest k. We show that this holds numerically for all parameters $k \in \{1.1, 1.2, ..., 2.9, 3\}$, $\theta = 1$, $a \in \{0.1, 0.2, ..., 5.9, 6\}$ and $v \in \{-2, -1.9, ..., 7.9, 8\}$. This numerically rules out a reversal when the prior distribution of ability follows the gamma distribution. See the Supplemental Appendix for the Matlab code to generate this simulation.

A.2 Binary Ability and Quality Distributions

In this section we consider a model in which ability and quality are binary. Suppose a worker has ability $a \in \{L, H\}$ with $p_0 = Pr(H)$. Each task has hidden quality $q_t \in \{l, h\}$, where $\rho_a = Pr(h|a)$ and $\rho_H > \rho_L$. Let $\phi(p) \equiv Pr(h) = \rho_H p + \rho_L(1-p)$ denote the probability of high quality, given belief p about ability. As before, evaluator t observes the evaluations on past tasks and signal s_t of the quality of the current task. Assume $s_t \sim N(\mu, 1)$ when the quality is h and $s_t \sim N(0, 1)$ when the quality is l, where the latter mean is a normalization. Assume $\mu > 0$. The evaluator reports the probability that the quality is high, $v_t = Pr(q_t = h|s_t, h_t)$.

Given belief p that the worker has high ability, after observing signal s, the evaluator reports evaluation $v(s, p)$, where

$$\frac{v(s,p)}{1-v(s,p)} = \frac{f^h(s)}{f^l(s)} * \frac{\phi(p)}{1-\phi(p)}. \tag{6}$$

The probability of high quality $\phi(p)$ is strictly increasing in p, and therefore, the evaluation $v(s, p)$ is strictly increasing in p. Therefore, for a given signal, a higher belief about ability leads to a higher evaluation. The evaluation $v(s, p)$ is also strictly increasing in s. Therefore, each signal s maps to a unique evaluation $v(s, p)$. Let $s(v, p)$ be the signal required to receive evaluation v, given belief p that the worker is high ability. Given $v(s(v, p), p) = v$ and

$$\log \frac{f^h(s)}{f^l(s)} = s^2/2 - (s-\mu)^2/2 = \mu s - \mu^2/2,$$

from (6),

$$\log \frac{v}{1-v} = \log \frac{f^h(s(v,p))}{f^l(s(v,p))} + \log \frac{\phi(p)}{1-\phi(p)} = \mu s(v,p) - \mu^2/2 + \log \frac{\phi(p)}{1-\phi(p)}.$$

Solving for $s(v,p)$ yields

$$s(v,p) = \frac{1}{\mu}\log\frac{v}{1-v} - \frac{1}{\mu}\log\frac{\phi(p)}{1-\phi(p)} + \mu/2.$$

After observing evaluation v, the distribution of ability updates to

$$\frac{B(v,p)}{1-B(v,p)} = \frac{f^h(s(v,p))\rho_H + f^l(s(v,p))(1-\rho_H)}{f^h(s(v,p))\rho_L + f^l(s(v,p))(1-\rho_L)} * \frac{p}{1-p}. \tag{7}$$

By this reasoning, the initial evaluation $v(s,p_0)$ is increasing in p_0. Given posterior $p_1 = B(v,p_0)$, the next period evaluation $v(s,p_1)$ is increasing in p_1. Suppose that $B(v,p)$ is increasing in p, i.e. $\frac{d}{dp}B(v,p) > 0$. Then the next period evaluation $v(s,B(v,p_0))$ is also increasing in p_0. Therefore, both the initial evaluation and the second period evaluation are increasing in p_0. This rules out the possibility of a reversal: if $p_0^M > p_0^F$ for males and females, then following the same evaluation, $B(v,p_0^M) > B(v,p_0^F)$. Therefore, $v(s,p_0^M) > v(s,p_0^F)$ and $v(s,B(v,p_0^M)) > v(s,B(v,p_0^F))$. By recursive reasoning, this implies that there is no evaluation reversal between any periods t and $t+1$.

Therefore, to rule out reversals, it is sufficient to show that $\frac{d}{dp}B(v,p) > 0$. This is equivalent to showing that $\frac{d}{dp}\log\frac{B(v,p)}{1-B(v,p)} > 0$, which from (7) is equivalent to

$$\frac{d}{dp}\left[\log Pr(v|H) - \log Pr(v|L) + \log p + \log(1-p)\right] > 0 \tag{8}$$

where $Pr(v|a) = f^h(s(v,p))\rho_a + f^l(s(v,p))(1-\rho_a)$. This is equivalent to showing

$$\frac{\frac{d}{dp}Pr(v|H)}{Pr(v|H)} - \frac{\frac{d}{dp}Pr(v|L)}{Pr(v|L)} + \frac{1}{p} + \frac{1}{1-p} > 0, \tag{9}$$

where, given $\frac{df^h}{ds} = f^h(s)(\mu - s)$ and $\frac{df^l}{ds} = -f^l(s)s$,

$$\begin{aligned}\frac{d}{dp}Pr(v|a) &= \frac{ds(v,p)}{dp}\left[\frac{df^h(s(v,p))}{ds}\rho_a + \frac{df^l(s(v,p))}{ds}(1-\rho_a)\right] \\ &= \frac{ds(v,p)}{dp}\left[f^h(s(v,p))(\mu - s(v,p))\rho_a - f^l(s(v,p))s(v,p)(1-\rho_a)\right]\end{aligned}$$

and

$$\frac{ds(v,p)}{dp} = -\frac{\rho_H - \rho_L}{\mu\phi(p)(1-\phi(p))}.$$

We show that (9) holds numerically for all parameters $\rho_L \in \{.02, .04, ..., .96\}$,

$\rho_H \in \{\rho_L + .02, ..., .98\}$, $\mu \in \{0.5, 0.55, ..., 2.5\}$, $p \in \{.01, .02, ..., .99\}$ and $v \in \{.01, .02, ..., .99\}$. This numerically rules out reversals in the binary model. See the Supplemental Appendix for the Matlab code to generate this simulation.

B Alternative Models

In this section, we explore two alternative models: (i) coarse evaluations and (ii) shifting standards. We show that our main results from Section I in the manuscript extend to these settings.

B.1 Coarse Evaluations

Set-up. Suppose that the set-up is identical to Section I.A in the manuscript, except that evaluations are binary – the evaluator chooses to either upvote or downvote a post, $v_t \in \{0, 1\}$. The evaluator receives a payoff of $q - c_g$ from upvoting a task from a worker of gender g and quality q, where, as before, c_g is a taste parameter with $c_M = 0$ and $c_F \geq 0$, and receives a payoff of 0 from downvoting a task.

The definitions of preference-based and belief-based partiality remain the same. We slightly adjust the definition of discrimination to account for the binary action space. A voting strategy specifies the set of signals that map into each type of vote. We say discrimination occurs at history h if there exists a set of signals on which females and males receive different votes. As before, define

$$D(h, s) \equiv v(h, s, M) - v(h, s, F).$$

Definition 1 (Discrimination)**.** *A female (male) faces* discrimination *at history h if $D(h, s) \geq 0$ ($D(h, s) \leq 0$) for all s, with a strict inequality for a positive measure of signals.*

Decision Rule. The evaluator maximizes her expected payoff by choosing $v_t = 1$ iff

$$E[q_t | h_t, s_t, g] \geq c_g, \tag{10}$$

where the expectation is taken with respect to the posterior distribution of quality, conditional on (h_t, s_t, g). Note that $E[q_t | s_t, h_t, g]$ is strictly increasing in s_t, since $f_{s|q}$ satisfies the MLRP with respect to q. Therefore, the optimal evaluation strategy can be represented as a cut-off rule on the signal. A task gets an upvote if the signal $s_t \geq \overline{s}(h_t, g)$ for some cut-off $\overline{s}(h_t, g)$. Discrimination can be represented in terms of the signal cut-off: a female faces discrimination at history h_t

if $\overline{s}(h_t, F) > \overline{s}(h_t, M)$, with an analogous definition for males. The set of signals on which discrimination occurs is an interval with measure $\overline{s}(h_t, F) - \overline{s}(h_t, M)$.

Initial Discrimination. As in Section I in the manuscript, the posterior belief about quality after observing signal s_1 is normal,

$$q_1|s_1 \sim N\left(\frac{\tau_q\hat{\mu}_g + \tau_\eta s_1}{\tau_q + \tau_\eta}, \frac{1}{\tau_q + \tau_\eta}\right).$$

The evaluator chooses $v_1 = 1$ if

$$\frac{\hat{\mu}_g\tau_q + s_1\tau_\eta}{\tau_q + \tau_\eta} \geq c_g,$$

or

$$s_1 \geq \overline{s}(\hat{\mu}_g, c_g) \equiv c_g\left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) - \hat{\mu}_g\left(\frac{\tau_q}{\tau_\eta}\right).$$

The cut-off is increasing in c_g and decreasing in $\hat{\mu}_g$. All of the initial discrimination results easily extend to the coarse evaluation setting. In particular, initial discrimination occurs if and only if $c_F > 0$ or $\hat{\mu}_M > \hat{\mu}_F$. As $\tau_\eta \to \infty$, $\overline{s}(h_1, g) \to c_g$. Therefore, initial discrimination persists as evaluations become perfectly objective if and only if evaluators have preference-based partiality, $c_F > 0$.

Impossibility of Reversal. For simplicity, we focus on how workers are evaluated in period $t = 2$, conditional on receiving an accept vote in period $t = 1$. We first consider a setting in which all evaluators have identical preferences and prior beliefs about ability, and have accurate beliefs about the preferences and prior beliefs of other evaluators. In the second period, the evaluator chooses $v_2 = 1$ if $E[q_2|v_1 = 1, s_2, g] \geq c_g$. Computing $E[q_2|v_1 = 1, s_2, g]$ is more challenging than in the first period, as the posterior belief about ability is no longer normally distributed, and therefore, neither is the posterior belief about quality q_2. By Lemma 1, we know that the belief about ability conditional on an upvote in the first period, $\{f_{\hat{\mu}}(a|v_1 = 1)\}_{\hat{\mu}\in\mathbb{R}}$, satisfies the MLRP in the prior $\hat{\mu}$. By Lemma 2, the MLRP is preserved under convolution with a normal error term, and hence, $E_{\hat{\mu}}[q_2|v_1 = 1, s_2, g]$ is increasing in $\hat{\mu}$. Therefore, when evaluators have belief-based partiality and a worker receives an upvote in the first period, there is no belief reversal in ability or expected quality in the second period, and hence, no discrimination reversal.

Proposition 1. *Suppose all evaluators have the same prior beliefs about the distributions of ability, a correct model of the beliefs and preferences of other evaluators, and belief-based partiality. Then there is no discrimination reversal in the second period, following an upvote in the first period.*

Therefore, the impossibility of a reversal also holds when evaluations are coarse.

Proof of Proposition 1. Suppose a worker has prior expected average ability $\hat{\mu}_g = \mu$. Let $f_\mu(a)$ denote the prior distribution of ability for this worker, and let $f_\mu(a|v_1 = 1)$ denote the posterior distribution, conditional on observing an upvote on the first post, $v_1 = 1$. By assumption, $f_\mu(a)$ is the normal distribution with mean μ and precision τ_a. After observing $v_1 = 1$, the public belief about ability is updated to

$$f_\mu(a|v_1 = 1) = \frac{P_\mu(v_1 = 1|a) f_\mu(a)}{\int_\infty^\infty P_\mu(v_1 = 1|a) f_\mu(a) da},$$

where $P_\mu(v_1 = 1|a)$ is the likelihood function that determines the informativeness of an upvote in the first period. This likelihood function is an equilibrium object that depends on gender and prior beliefs. □

Lemma 1. *The family of posterior beliefs about ability following an upvote in the first period, $\{f_\mu(a|v_1 = 1)\}_{\mu\in\mathbb{R}}$, satisfies the MLRP in μ.*

Proof. Since the prior belief about ability is normal, $f_\mu(a) = \sqrt{\tau_a}\phi(\sqrt{\tau_a}(a - \mu))$, where ϕ is the p.d.f. of the standard normal distribution. Therefore, $\{f_\mu(a)\}_{\mu\in\mathbb{R}}$ is MLR ordered in μ, by property of the normal distribution. The likelihood function depends on the cut-off rule $\overline{s}$,

$$\begin{aligned}
P_\mu(v_1 = 1|a) &= P_\mu(s_1 \geq \overline{s}|a) \\
&= P_\mu(a + \varepsilon_1 + \eta_1 \geq \overline{s}|a) \\
&= P_\mu(\varepsilon_1 + \eta_1 \geq \overline{s} - a|a) \\
&= P_\mu(\varepsilon_1 + \eta_1 \geq \overline{s} - a) \qquad \text{since } \varepsilon_1, \eta_1 \perp a \\
&= 1 - \Phi\left(\sqrt{\tau_{\varepsilon\eta}}(\overline{s} - a)\right) \qquad \text{since } \varepsilon_1 + \eta_1 \sim N(0, 1/\tau_{\varepsilon\eta}) \\
&= \Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a - \overline{s})\right) \qquad \text{since } 1 - \Phi(x) = \Phi(-x)
\end{aligned}$$

where Φ is the c.d.f of the standard normal distribution, and $\tau_{\varepsilon\eta} \equiv \frac{\tau_\varepsilon \tau_\eta}{\tau_\varepsilon + \tau_\eta}$. Therefore, for cut-off rule $\overline{s}(\mu, c)$, the likelihood ratio of the posterior distribution of ability is

$$\begin{aligned}
\frac{f_\mu(a|v_1 = 1)}{f_\mu(a'|v_1 = 1)} &= \frac{P_\mu(v_1 = 1|a)}{P_\mu(v_1 = 1|a')} \cdot \frac{f_\mu(a)}{f_\mu(a')} \\
&= \frac{\Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a - \overline{s}(\mu, c))\right)}{\Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a' - \overline{s}(\mu, c))\right)} \cdot \frac{\phi(\sqrt{\tau_a}(a - \mu))}{\phi(\sqrt{\tau_a}(a' - \mu))}. \qquad (11)
\end{aligned}$$

The goal is to show that (11) is increasing in μ for $a > a'$, i.e. the posterior belief satisfies the MLRP. The first term on the RHS is decreasing in μ, since an upvote is more informative for lower μ (or higher c), and the second term on the RHS is increasing in μ, since the prior belief satisfies the MLRP in μ. The posterior belief will satisfy the MLRP iff for all a and μ,

$$\frac{\partial^2}{\partial a \partial \mu} \log P_\mu(v_1 = 1|a) + \log f_\mu(a) \geq 0. \tag{12}$$

Recall $\overline{s}(\mu, c) = c\left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) - \mu\left(\frac{\tau_q}{\tau_\eta}\right)$. Computing the first term of (12),

$$\begin{aligned}
\frac{\partial^2}{\partial a \partial \mu} \log P_\mu(v_1 = 1|a) &= \frac{\partial^2}{\partial a \partial \mu} \log \Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a - \overline{s}(\mu, c))\right) \\
&= \frac{\partial}{\partial a} \frac{\phi\left(\sqrt{\tau_{\varepsilon\eta}}(a - \overline{s})\right)}{\Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a - \overline{s})\right)} \times \left(-\frac{\partial \overline{s}}{\partial \mu}\right) \sqrt{\tau_{\varepsilon\eta}} \\
&= \frac{-\Phi(x)\phi(x)x - \phi(x)^2}{\Phi(x)^2} \times \left(-\frac{\partial \overline{s}}{\partial \mu}\right) \tau_{\varepsilon\eta} \\
&= -\left(\frac{\phi(x)x}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2}\right)\left(\frac{\tau_q \tau_{\varepsilon\eta}}{\tau_\eta}\right),
\end{aligned}$$

where $x \equiv \sqrt{\tau_{\varepsilon\eta}}(a - \overline{s}(\mu, c))$ and $-\frac{\partial \overline{s}}{\partial \mu} = \tau_q/\tau_\eta$. Computing the second term of (12)

$$\begin{aligned}
\frac{\partial^2}{\partial a \partial \mu} \log f_\mu(a) &= \frac{\partial^2}{\partial a \partial \mu} \log \phi(\sqrt{\tau_a}(a - \mu)) \\
&= \frac{\partial}{\partial a} \frac{\tau_a(a - \mu)\phi(\sqrt{\tau_a}(a - \mu))}{\phi(\sqrt{\tau_a}(a - \mu))} \\
&= \frac{\partial}{\partial a} \tau_a(a - \mu) \\
&= \tau_a.
\end{aligned}$$

Therefore, need to show that for all x,

$$\begin{aligned}
&\tau_a - \left(\frac{\phi(x)x}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2}\right)\left(\frac{\tau_q \tau_{\varepsilon\eta}}{\tau_\eta}\right) \geq 0 \\
&\Leftrightarrow \tau_x - \left(\frac{\phi(x)x}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2}\right) \geq 0,
\end{aligned} \tag{13}$$

where $\tau_x \equiv \frac{\tau_a \tau_\eta}{\tau_q \tau_{\varepsilon\eta}}$. From Stack Exchange[1], we know that

$$\left(\frac{\phi(x)x}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2} \right) \leq 1.$$

From the definition of τ_x,

$$\begin{aligned} \tau_x &\equiv \frac{\tau_a \tau_\eta}{\tau_q \tau_{\varepsilon\eta}} \\ &= \frac{(\tau_a + \tau_\varepsilon)(\tau_\eta + \tau_\varepsilon)}{\tau_\varepsilon^2} \\ &= \frac{\tau_a \tau_\eta}{\tau_\varepsilon^2} + \frac{\tau_\eta}{\tau_\varepsilon} + \frac{\tau_a}{\tau_\varepsilon} + 1 \\ &\geq 1. \end{aligned}$$

Therefore, (13) holds for all x. Therefore, for all $a > a'$, (11) is increasing in μ and $\{f_\mu(a|v=1)\}_{\mu \in \mathbb{R}}$ satisfies the MLRP. □

Given Lemma 1, for $\mu > \mu'$, $f_\mu(a|v=1)$ first-order stochastically dominates $f_{\mu'}(a|v=1)$. Therefore, $E_\mu[a|v_1=1]$ is increasing in μ, and there is no belief reversal about ability in the second period. Lemma 2 establishes that the posterior distribution of quality following an upvote in the first period and signal s_2 in the second period, $g_\mu(q_2|v_1=1,s_2)$, also satisfies the MLRP in the prior belief μ.

Lemma 2. *The posterior distribution of quality, following an upvote in the first period and signal s_2 in the second period, $\{g_\mu(q_2|v_1=1,s_2)\}_{\mu\in\mathbb{R}}$, satisfies the MLRP in μ.*

Proof. From Lemma 1, $\{f_\mu(a|v_1=1)\}_{\mu\in\mathbb{R}}$ satisfies the MLRP. Since $q_2 = a + \varepsilon_2$, the prior distribution of second period quality, $g_\mu(q_2|v_1=1)$, is the convolution of $f_\mu(a|v_1=1)$ and $f_\varepsilon(\varepsilon)$, where f_ε denotes the density of ε. From Theorem 2.1(d) in **?**, the MLRP is preserved when an independent random variable with a log-concave density function is added to a family of random variables that satisfy the MLRP. Since $a \perp \varepsilon$ and f_ε is a log-concave density (the normal distribution is log concave), the family of distributions $\{g_\mu(q_2|v_1=1)\}_{\mu\in\mathbb{R}}$ satisfies the MLRP. Therefore,

$$\frac{\partial^2}{\partial q \partial \mu} \log g_\mu(q_2|v_1=1) > 0,$$

[1] https://math.stackexchange.com/questions/2337419/property-of-standard-normal

which also means that

$$\frac{\partial^2}{\partial q \partial \mu} \log g_\mu(q_2|v_1 = 1, s_2) > 0,$$

since the likelihood function (the distribution of $s_2|q_2$) is independent of μ, and the denominator is independent of q_2. Therefore, for any signal s_2, the posterior belief about quality $\{g_\mu(q_2|v_1 = 1, s_2)\}_{\mu \in \mathbb{R}}$ also satisfies the MLRP. □

The MLRP implies FOSD, which implies that for any signal s_2, $E_\mu[q_2|v_1 = 1, s_2]$ is increasing in μ. Therefore, there is no belief reversal about quality in the second period. Hence, discrimination does not reverse between the first and second period.

B.2 Shifting Standards

Suppose that the evaluator's payoff also depends on the seniority of the worker, as measured by the worker's *reputation* $r(h_t) \equiv \sum_{n=1}^{t-1} v_n$, which is the sum of the worker's past evaluations. She receives a payoff of $(v - (q - c(r) - c_g))^2$ from reporting evaluation v on a task of quality q from a worker of gender g and reputation r, where $c : \mathbb{R} \to \mathbb{R}_+$ is the *benchmark of evaluation* for a worker with reputation r and, as above, c_g is a taste parameter with $c_M = 0$. Assume that $c(r)$ is weakly increasing in r to capture the idea that as reputation increases, a worker receives additional privileges or promotions, and the benchmark to promote the worker increases with the worker's seniority. Normalize the initial benchmark to $c(0) = 0$, and assume that $c(r) = 0$ for all $r < 0$, so that workers who produce negative quality do not receive a more lenient benchmark.

The optimal evaluation strategy is to report

$$v(h_t, s_t, g) = \frac{\tau_{q,t} \hat{\mu}_g(h_t) + \tau_\eta s_t}{\tau_{q,t} + \tau_\eta} - c(r(h_t)) - c_g, \tag{14}$$

where $\hat{\mu}_g(h_t)$ is the expected ability of the worker, conditional on history h_t. Fixing $\hat{\mu}_g(h_t)$ and s_t, as the worker's reputation increases, he or she receives a lower evaluation for the same expected quality. Note that shifting standards will have no effect on discrimination, since the benchmark of evaluation term cancels between females and males, $D(h_t, s_t) = \left(\frac{\tau_{q,t}}{\tau_{q,t} + \tau_\eta}\right)(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)) + c_F$.

A positive initial evaluation (i.e. above average, $v_1 > \hat{\mu}_g$) impacts the *standard* faced by a worker – the signal required to receive a given evaluation – in two ways: it increases the evaluator's belief about the worker's ability, and it increases the benchmark of evaluation. A positive evaluation is *good news* about ability: the distribution of ability following a positive evaluation first order stochastically

dominates the prior distribution of ability. Since expected quality is equal to expected ability, and the signal required to earn a given evaluation is decreasing in expected quality, increasing the expected ability while holding reputation constant results in a lower standard. However, a positive evaluation also increases the worker's reputation, and therefore, the benchmark of evaluation. Holding the belief about ability fixed, higher reputation workers face stricter standards. Therefore, the overall effect of a positive evaluation on standards is ambiguous.

We say a worker faces *shifting standards* if, conditional on receiving a positive initial evaluation, the worker faces a stricter standard in period 2 – a higher signal is required to receive any evaluation, relative to the signal required for the same evaluation in period 1. Let $s(v, h, g)$ denote the signal required for a worker with history h and gender g to receive evaluation v.

Definition 2. *A worker faces* shifting standards *following evaluation v_1 if the initial evaluation is positive, $v_1 > \hat{\mu}_g$, but the worker subsequently faces a stricter standard, $s(v, v_1, g) > s(v, \emptyset, g)$ for all $v \in \mathbb{R}$.*

Shifting standards implies that the positive evaluation's negative impact on the benchmark of evaluation outweighs the positive impact on the belief about the worker's expected quality. Note that the definition is required to hold at all evaluations $v \in \mathbb{R}$, but this is not restrictive, as given $h_2 \supset h_1$, $s(v, h_2, g) - s(v, h_1, g)$ is independent of v. Therefore, the definition either holds at all evaluations or at no evaluations. For any positive initial evaluation v_1, it is straightforward to show that there exists a cut-off $\overline{c}$ such that if the new benchmark of evaluation exceeds this cut-off, $c(v_1) > \overline{c}$, a worker faces shifting standards.

Standards unambiguously rise after a negative initial evaluation, $v_1 < \hat{\mu}_g$. A negative evaluation is bad news about the worker's ability, and either raises or maintains the initial benchmark of evaluation.

C Additional Empirical Analysis

C.1 Example Question and Answer Posts

The following screenshots of a randomly selected question and answer post illustrate how users create content on the forum. These posts are not part of our experiment.

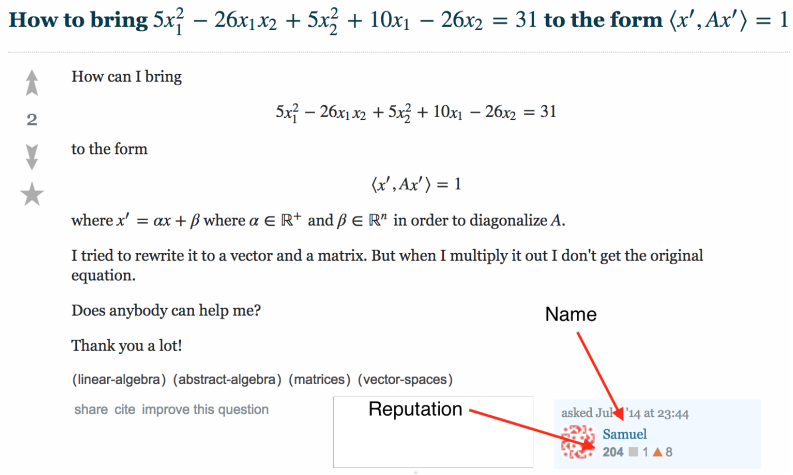

Figure 1. Question Post

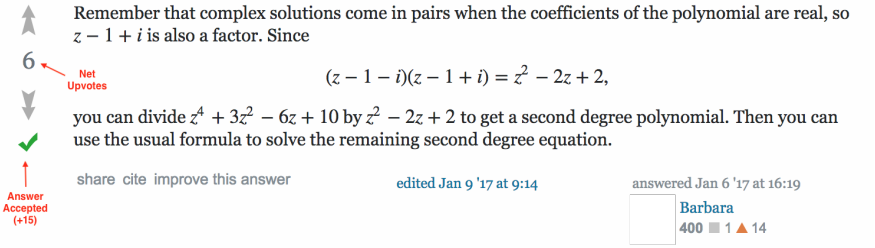

Figure 2. Answer Post

C.2 Robustness

Upvotes Only. The following tables present analogous regressions to Tables 1 and 2 in the manuscript, using number of upvotes as the dependent variable.

Table 1. Subjectivity: Effect of Gender on Evaluation of Novice Answers and Questions (Upvotes Only)

	Answers (1)	Questions (2)	Answers & Questions (3)
Male	-0.20	0.57	-0.20
	(.17)	(.27)	(.23)
Question			0.17
			(.23)
Male*Question			0.77
			(.32)
Constant	0.81	0.97	0.81
	(.12)	(.19)	(.16)
# Obs	135	135	270

Standard errors from OLS regressions reported in parentheses; Male=1 if male username, 0 otherwise; Question=1 if question post, 0 if answer; Novice accounts only.

Table 2. Dynamics: Effect of Gender on Evaluation of Novice and Advanced Questions (Upvotes Only)

	Novice (1)	Advanced (2)	Novice & Advanced (3)
Male	0.57	-0.64	0.57
	(.27)	(.27)	(.27)
Advanced			0.45
			(.27)
Male*Advanced			-1.20
			(.38)
Constant	0.97	1.42	0.97
	(.19)	(.19)	(.19)
# Obs	135	138	273

Standard errors from OLS regressions reported in parentheses; Male=1 if male username, 0 otherwise; Advanced=1 if Advanced account, 0 otherwise.

First Vote Only. The following tables present parallel regressions to Tables 1 and 2 in the manuscript, using only the first vote on a post in our experiment from each evaluator.

Table 3. Subjectivity: Effect of Gender on Evaluation of Novice Answers and Questions (First Vote Only)

	Answers Only		Questions Only		Answers & Questions	
	Δ Rep	Net Votes	Δ Rep	Net Votes	Δ Rep	Net Votes
	(1)	(2)	(3)	(4)	(5)	(6)
Male	−1.15	−0.28	2.17	0.44	−1.15	−0.28
	(.82)	(.16)	(1.07)	(.22)	(.96)	(.19)
Question					−0.42	−0.13
					(.96)	(.19)
Male*Question					3.32	0.72
					(1.35)	(.27)
Constant	3.55	0.70	3.13	0.57	3.55	0.70
	(.58)	(.12)	(.76)	(.15)	(.68)	(0.14)
# Obs	135	135	135	135	270	270

Standard errors from OLS regressions reported in parentheses; Male=1 if male username, 0 otherwise; Question=1 if question post, 0 if answer; Novice accounts only.

Table 4. Dynamics: Effect of Gender on Evaluation of Novice and Advanced Questions (First Vote Only)

	Advanced		Novice & Advanced		
	Δ Rep	Net Votes	Δ Rep	Net Votes	Binary
	(1)	(2)	(3)	(4)	(5)
Male	−2.58	−0.51	2.17	0.44	0.10
	(1.14)	(.23)	(1.12)	(.23)	(.08)
Advanced			1.64	0.35	0.02
			(1.11)	(.22)	(0.08)
Male*Advanced			−4.75	−0.95	−0.28
			(1.57)	(.32)	(.11)
Constant	4.77	0.93	3.13	0.57	0.44
	(0.81)	(.16)	(.79)	(.16)	(.06)
# Obs	138	138	273	273	273

Standard errors from OLS regressions reported in parentheses; Male=1 if male username, 0 otherwise; Advanced=1 if Advanced account, 0 otherwise.

C.3 Voter Characteristics

Table 5. Voter Characteristics by Post Type

	Voter Reputation (1)	Voter Gender: % Female (2)
Answers	16679	0.14
	(2040)	(.04)
Questions: All	18836	0.10
	(1254)	(.02)
Questions: Novice	17957	0.11
	(1684)	(.03)
Questions: Advanced	19839	0.09
	(1877)	(.03)

Note: Standard errors reported in parentheses; voter reputation winsorized at 90 percent.

C.4 Observational Data Analysis

We next present the analysis of the observational data described in Section II.D in the manuscript.

C.4.1 Description of Algorithm to Code Gender.

? developed the algorithm to code gender and validated its accuracy through secondary data collection on online Q&A forums. The algorithm uses look-up tables with the frequencies of first names by gender and country. For example, while John and Claire are common male and female names, respectively, across countries, Andrea is a common male name in Italy and a common female name in Germany. We preprocessed the data to obtain (*name, country*) tuples for each user when such information is available. The preprocessed data is then fed into a Python tool that classifies the tuple as 'male,' 'female,' or 'x' (when gender cannot be inferred). The tool uses an iterative process that first employs country-specific look-up tables, and if that does not lead to a resolution, switches to common conventions for usernames (**?**). **?** collected additional data from users on the forum to validate the tool, demonstrating a level of precision greater than 90%. The algorithm and associated data files are publicly available on GitHub at https://github.com/tue-mdse/genderComputer.

C.4.2 Attrition

In the following analysis, we use the logged measure of the reputation earned on a post for interpretability. All results also hold with the unlogged measure.

In Table 6, we run a probit regression, regressing a dummy for whether a user generated a second post on the inferred gender of the username, the log of the reputation earned on the first post and their interaction. Column (1) presents the results pooling question and answer first posts, and also includes a dummy for whether the first post was a question, Column (2) presents the results for question first posts only and Column (3) does the same for answers first posts only. Neither the gender variable nor the interaction is significant in any of the specifications.

Table 6. Likelihood of Generating a Second Post

	Pooled	Questions	Answers
	(1)	(2)	(3)
Male	-0.10	0.15	-0.53
	(.23)	(.28)	(.42)
Reputation First Post	0.58	0.62	0.51
	(.05)	(.06)	(.09)
Reputation First Post * Male	0.03	-0.03	0.13
	(.05)	(.07)	(.10)
First Post Question	-0.13		
	(.01)		
Constant	-2.47	-2.76	-2.18
	(.21)	(.25)	(.39)
# Obs	85,354	71,868	13,486

Standard errors from probit regressions reported in parentheses; Second Post=1 if user posts a second time, 0 otherwise; Male=1 if male username, 0 otherwise; First Post Question = 1 if the first post was an question, 0 otherwise.

In Table 7, we split the reputation earned on the first post into quartiles. We again run a probit regression, regressing a dummy for whether a user generated a second post on the inferred gender of the username, a dummy for the quartile of reputation earned on the first post, and the interaction of the gender dummy with each quartile dummy. We again do not observe a significant main effect of gender nor of the interaction with reputation quartile.[2]

[2]The results are robust to different size reputation bins. Coefficients on gender and the interactions are not significant for alternative numbers of bins. When a coefficient does approach

Table 7. Likelihood of Generating a Second Post (Quartiles)

	Pooled (1)	Questions (2)	Answers (3)
Male	0.03	0.09	0.01
	(.05)	(.06)	(.14)
Reputation First Post Q2	0.24	0.20	0.49
	(.05)	(.06)	(.14)
Reputation First Post Q3	0.36	0.31	0.68
	(.05)	(.05)	(.14)
Reputation First Post Q4	0.54	0.47	0.93
	(.05)	(.06)	(.14)
Reputation First Post Q2 * Male	-0.06	-0.10	-0.05
	(.06)	(.06)	(.15)
Reputation First Post Q3 * Male	-0.02	-0.08	0.05
	(.06)	(.06)	(.15)
Reputation First Post Q4 * Male	0.02	-0.05	0.06
	(.06)	(.06)	(.15)
First Post Question	-0.17		
	(.01)		
Constant	-0.38	-0.49	-0.70
	(.05)	(.05)	(.13)
# Obs	85,354	71,868	13,486

Standard errors from probit regressions reported in parentheses; Second Post=1 if user posts a second time, 0 otherwise; Male=1 if male username, 0 otherwise; First Post Question = 1 if the first post was a question, 0 otherwise. The first reputation quartile is the omitted variable across all specifications.

In Table 8, we repeat the analysis from Table 6 for the likelihood of generating a third through tenth post, pooling questions and answers. Column (t) presents the results for the probit regression that regresses whether a user generated a post t on the gender dummy, log of reputation earned on the previous post (post $t-1$), their interaction, and whether the previous post was a question. Neither the coefficient on the gender dummy nor the interaction is significant in any of the specifications $t = 3, ..., 10$.

significance, if anything, its sign suggests that women who earned a low initial reputation are more likely to generate a second post than males who earned a low initial reputation – a form of differential attrition that would generate larger subsequent discrimination against females, not the reversal that we observe.

Table 8. Likelihood of Generating Post t

	Post 3	Post 4	Post 5	Post 6	Post 7	Post 8	Post 9	Post 10
Male	0.49	-0.24	0.30	-0.81	-1.13	1.03	1.10	0.88
	(0.39)	(0.51)	(0.60)	(0.67)	(0.78)	(0.97)	(1.06)	(1.16)
Rep. Previous Post	0.83	0.65	0.69	0.44	0.52	0.88	0.87	0.83
	(0.09)	(0.11)	(0.14)	(0.15)	(0.17)	(0.22)	(0.24)	(0.27)
Rep. Previous Post * Male	-0.12	0.06	-0.07	0.21	0.27	-0.23	-0.27	-0.20
	(0.10)	(0.12)	(0.15)	(0.16)	(0.19)	(0.24)	(0.26)	(0.28)
Previous Post Question	-0.19	-0.25	-0.29	-0.25	-0.30	-0.33	-0.28	-0.25
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Constant	-2.80	-1.72	-1.70	-0.63	-0.77	-2.21	-2.11	-1.90
	(0.36)	(0.47)	(0.56)	(0.62)	(0.72)	(0.91)	(1.01)	(1.10)
# Obs	37,781	25,848	20,344	16,964	14,519	12,772	11,433	10,339

Standard errors from probit regressions reported in parentheses; Post # = 1 if user posts after posting for Post (# − 1), 0 otherwise; Male=1 if male username, 0 otherwise; Previous Post Question = 1 if the previous post was an question, 0 otherwise.

Finally, in Table 9, we pool all posts in the same regression. Column (1) presents the results for the probit regression that regresses whether a user generated a post t on the gender dummy, log of total reputation earned on all previous posts, their interaction, and whether the previous post was a question. In Columns (2) and (3), we include dummies for the post number in the sequence – this controls for how many posts it took to generate the total reputation. In Column (3), we also control for the log of reputation earned on the previous post and its interaction with gender. This allows for the possibility that the evaluation of a user's most recent post is more salient for his or her decision to post again relative to earlier performance. Standard errors are clustered at the individual level. As can be seen from Table 9, neither the coefficient on the gender dummy nor on the interaction with total reputation is significant in any of the specifications. In Column (3), the coefficient on the interaction of gender and reputation earned on previous post also is not significant.

Table 9. Likelihood of Generating Next Post

	(1)	(2)	(3)
Male	0.01	0.02	0.02
	(.01)	(.01)	(.17)
Total Reputation	0.05	0.24	0.11
	(.00)	(.01)	(.01)
Total Reputation * Male	0.00	-0.00	0.00
	(.00)	(.00)	(.00)
Reputation Previous Post			0.49
			(.04)
Reputation Previous Post * Male			-0.00
			(.04)
Previous Post Question	-0.31	-0.21	-0.18
	(.01)	(.01)	(.01)
Constant	0.04	-0.98	-2.50
	(.01)	(.03)	(.16)
Post Number Dummies	No	Yes	Yes
# Obs	235,354	235,354	235,354

Standard errors from probit regressions reported in parentheses, clustered at the user level; Next Post=1 if user posts a subsequent post, 0 otherwise; Male=1 if male username, 0 otherwise; Previous Post Question = 1 if the previous post was an question, 0 otherwise. Post Number refers to whether or not dummies corresponding to the post's position in the sequence are included.

C.4.3 Autocorrelation in Error Process for Quality

Suppose individual i has ability a_i. From Section I.A in the manuscript, i's answer in period t has quality

$$q_{i,t} = a_i + \varepsilon_{i,t},$$

and μ_g denotes the population average ability for a new user of gender g. Negative autocorrelation in the error process corresponds to $Cor(\varepsilon_{i,t}, \varepsilon_{i,t+1}) < 0$. If there is sufficiently negative autocorrelation, then observing above average quality in period t (i.e. $q_1 > \mu_g$) leads to expected quality that is below average in period $t+1$ (i.e. $E[q_2|q_1] < \mu_g$). If females have lower expected ability than males, then observing above average quality in $t+1$ is more informative about ability and could possibly generate a belief reversal in expected ability. This is because a given level of high quality in the subsequent period is more informative for females than males following a previous observation of similar high quality. Note that an error process with negative autocorrelation is distinct from mean-reverting quality. For any form of correlation in $\varepsilon_{i,t}$ (e.g. positive, negative or none), if $q_{i,t} > \mu_{g_i}$, then $E[q_{i,t+1}|q_{i,t}] < q_{i,t}$ and lower quality is expected in the subsequent period. Therefore, negative autocorrelation in the errors is not required to generate mean-reverting quality.

Answers. If answer quality is observable, then each evaluation corresponds to the report of quality, $v_{i,t} = q_{i,t}$ for answer post t from user i. We can write the quality as

$$q_{i,t} = \mu_{g_i} + (a_i - \mu_{g_i}) + \varepsilon_{i,t}.$$

Consider the following random effects regression:

$$v_{i,t} = \beta_0 + \beta_1 * \mathbf{1}_{g_i=M} + u_i + e_{i,t}, \tag{15}$$

where gender g_i is the gender of user i. Then $\mu_F = \beta_0$, $\mu_M = \beta_0 + \beta_1$, $a_i - \mu_{g_i} = u_i$ and $\varepsilon_{i,t} = e_{i,t}$.

We use the Wooldridge test for serial correlation in panel data to test for serial correlation in the error $\varepsilon_{i,t}$ (**?**). We compiled a panel dataset consisting of all answer posts from users with reputation 1 to 250, which is the relevant reputation range for our experiment. Following specification (15), we first ran a random effects regression of the reputation earned on an answer post on a dummy for gender. We then tested the estimated residuals $\hat{e}_{i,t}$ for autocorrelation using the xtserial program in Stata. Under the null hypothesis of no first-order autocorrelation, we found an F-statistic of $F(1, 7972) = 0.277$ and $Prob > F = 0.5988$. Therefore, we do not observe significant autocorrelation in the error process for the quality of answer posts.

Questions. Question posts have an added layer of complication, as the reported evaluation $v_{i,t}$ is a combination of the signal of quality of question post t and the current belief about the ability of user i when he or she posts question t, which we denote by $\mu_{g_i,t}$. From (A5) in the manuscript,

$$\begin{aligned} v_{i,t} &= \frac{1}{\tau_q(t)+\tau_\eta}(\tau_q(t)\mu_{g_i,t}+\tau_\eta s_{i,t}) \\ &= \frac{1}{\tau_q(t)+\tau_\eta}(\tau_q(t)\mu_{g_i,t}+\tau_\eta a_i+\tau_\eta(\varepsilon_{i,t}+\eta_{i,t})) \end{aligned}$$

where $\tau_q(t) \equiv \tau_a(t)\tau_\varepsilon/(\tau_a(t)+\tau_\varepsilon)$ and the second line follows from signal $s_{i,t} = a_i+\varepsilon_{i,t}+\eta_{i,t}$. From (A8) in the manuscript, the current belief $\mu_{g_i,t}$ is an additively separable function of past evaluations and the prior belief μ_g. We do not directly observe $\mu_{g_i,t}$, but we can proxy the past evaluations component of it with the current reputation score. Consider the following random effects regression:

$$\begin{aligned} v_{i,t} = \quad & \beta_0+\beta_1*\mathbf{1}_{g_i=M}+\beta_2*R_{i,t}+\beta_3*R_{i,t}*\mathbf{1}_{g_i=M} \\ & +\beta_4*NumPosts_{i,t}+\beta_5*NumPosts_{i,t}*\mathbf{1}_{g_i=M}+u_i+e_{i,t}, \end{aligned} \quad (16)$$

where $R_{i,t}$ is cumulative reputation of user i when he/she posts question t and $NumPosts_{i,t}$ is the number of posts (questions and answers) that generated $R_{i,t}$. Note that index t refers to question t, not post t, since we are restricting attention to questions. Similar to the case of answers, the random effect is the difference between individual and population ability (i.e. the prior μ_g), β_0 is a function of the prior belief about average female ability and $\beta_0+\beta_1$ is a function of the prior belief about average male ability. The reputation terms capture the past evaluation component of current beliefs, while β_0 and β_1 capture the prior belief component of current beliefs (recall the current belief is an additive function of these two components).

As in the case of answers, we ran a random effects regression on questions posts using specification (16), then tested the estimated residuals for autocorrelation. Under the null hypothesis of no first-order autocorrelation, we found an F-statistic of $F(1,7972) = 51.947$ and $Prob > F = 0.0000$. Therefore, we reject the null hypothesis of no first-order autocorrelation. Next, we use the estimated residuals to run the regression:

$$\hat{e}_{i,t} = \rho\hat{e}_{i,t-1}+error_{i,t},$$

in order to determine the direction of autocorrelation. The estimated correlation is positive, with coefficient $\hat{\rho} = .076$ and standard error .003. Therefore, this is not consistent with an error process that exhibits negative autocorrelation.

C.4.4 Gender Differences in Evaluations.

Next, we examine gender differences in evaluations in the observational dataset. As in our experiment, we focus on the evaluation of questions posted to novice and advanced accounts, and the evaluation of answers posted to novice accounts. We define posting to novice and advanced accounts similar to the experiment. A novice post corresponds to posting a question or answer to an account with no prior reputation or posts. An advanced post corresponds to posting a question to an account that has attained a reputation of at least 100 points but not more than 240 (the approximate range in our experiment); importantly, the question has to be the *first* post to the account once it reaches this reputation threshold.

This analysis comes with several important caveats that are discussed in the text, including that the number of posts that generated a user's reputation is relevant for inferring ability, as different numbers of posts can result in similar reputations. We control for this issue in our experiment through randomization; it is less straightforward to control for in the observational data. We attempt to address the issue by running specifications where the advanced accounts required 20 or fewer posts to reach their respective reputation levels. A user earning the average number of upvotes per post would need to post approximately 20 questions to attain 100 reputation points.[3]

We find that the evaluation patterns by gender across the different types of posts are similar to those documented in the experiment, although the effect sizes vary and are often smaller. For the evaluation of answers (Table 10), we regress reputation points earned per answer post (ΔRep) on inferred gender. We restrict attention to answers posted to accounts with a reputation less than 240 (Column (1)), answers posted to novice accounts (Column (2)), and answers posted to novice accounts during the timeframe of the experimental study (Column (3)). Across these three specifications, we find no significant evidence of gender discrimination.

For the evaluation of novice questions (Table 11), we regress reputation points earned per question post (ΔRep) on inferred gender for questions posted by novice users. We run the analysis on questions posted to all novice accounts (Column (1)), questions posted to novice accounts during the timeframe of the experimental study (Column (2)) and questions posted to novice accounts for users who also posted after reaching at least 100 reputation points (Column (3)). Restricting attention to users who eventually earn at least 100 reputation points allows us to focus on users who are presumably posting higher quality

[3]The results are robust to limiting the analysis to 10 or fewer posts, which is the number of answers an average user would need to post to attain 100 reputation points. Increasing or decreasing the number of posts, including the variable in the regression, or not controlling for it at all does not qualitatively change the results.

Table 10. Evaluation of Answers: ΔRep

	Reputation < 240	Novice	Novice - Experiment Window
	(1)	(2)	(3)
Male	0.09	0.23	-0.66
	(.28)	(.33)	(.46)
Constant	7.44	7.94	6.32
	(.26)	(.31)	(.42)
# Obs	19,983	10,760	3,533

Standard errors from OLS regressions reported in parentheses; Male=1 if male username, 0 otherwise.

content, given the reputation they eventually earn.[4] Across all specifications, we find that questions posted by novice accounts with female usernames earn fewer reputation points than those posted by novice accounts with male usernames. The magnitude of this difference is larger for the specifications that restrict attention to the users who eventually reach 100 reputation points.

Table 11. Evaluation of Questions Posted by Novice Users: ΔRep

	All	Experiment Window	Reach 100
	(1)	(2)	(3)
Male	0.58	0.30	3.56
	(.11)	(.14)	(2.04)
Constant	7.92	5.05	20.58
	(.10)	(.12)	(1.85)
# Obs	72,896	26,092	5,927

Standard errors from OLS regressions reported in parentheses; Male=1 if male username, 0 otherwise.

Lastly, we look at questions posted to advanced accounts (Table 12). We regress reputation points earned per question post (ΔRep) on inferred gender for questions posted to all advanced accounts (Column (1)), questions posted to advanced accounts that required 20 or fewer posts to reach their respective

[4]The specification in (3) lifts the restriction of the user attaining a maximum reputation of 240, since this additional restriction cuts the sample substantially. The results reported in specifications (1) and (2) are robust to lifting this restriction as well, with coefficients on the Male dummy increasing to 1.33 and 0.45, respectively.

reputation levels (Column (2)), and questions posted to advanced accounts during the timeframe of the experimental study (Column (3)). Across all specifications we find that questions posted to advanced accounts with female usernames are favored over those posted to advanced accounts with male usernames.

Table 12. Evaluation of Questions Posted by Advanced Users: ΔRep

	All	< 20 Posts	Experiment Window
	(1)	(2)	(3)
Male	-0.88	-1.58	-1.63
	(.55)	(.68)	(.88)
Constant	9.3	10.59	7.65
	(.49)	(.61)	(.75)
# Obs	2,123	1,599	531

Standard errors from OLS regressions reported in parentheses; Male=1 if male username, 0 otherwise.

D Stereotyping

In Section I in the manuscript, we established that a dynamic reversal of discrimination can arise when some evaluators hold beliefs that females are of lower average ability than they actually are, and other evaluators are aware of these incorrect beliefs. In this section, we use publicly available statistics from the observational dataset to explore one potential mechanism that could lead to such biased beliefs.

? develop a framework in which biased stereotypes arise and persist due to 'representativeness', a well-documented cognitive heuristic used to simplify complex probability judgments (**?**). When assessing the frequency of a type in a particular group, an individual who uses this heuristic focuses on the *relative* likelihood of that type with respect to a reference group, rather than assessing the absolute frequency of the type. The type that is most frequently found in one group relative to another, e.g. the frequency of Floridians over 65 relative to the frequency of people over 65 in the rest of the country, is *representative* of that group. The heuristic exaggerates the perceived frequency of the representative type in the respective group, and as a result, distorts beliefs about the associated type distribution. Specifically, a 'kernel of truth' in the relative frequency – that the proportion of seniors is higher amongst Floridians than in the rest of the US – may lead to a biased stereotype about absolute frequencies – that most

Floridians are seniors.[5]

Let t represent a user's quintile in the ability distribution, $t \in T = \{1^{st}, ..., 5^{th}\}$. A type t is 'representative' of group g, in relation to the comparison group $-g$, if the likelihood ratio $\pi_{t,g}/\pi_{t,-g}$ is high, where $\pi_{t,g}$ is the probability that a worker from group g is in quantile t. The 'representative' type corresponds to the most salient difference between groups; it is the first type to come to mind when using the heuristic to form beliefs, and leads to overweighting of the perceived frequency of the type within the group. Specifically, **?** define the stereotyped belief as

$$\pi^{st}_{t,g} \equiv \pi_{t,g} \frac{\left(\frac{\pi_{t,g}}{\pi_{t,-g}}\right)^{\theta}}{\sum_{s \in T} \pi_{s,g} \left(\frac{\pi_{s,g}}{\pi_{s,-g}}\right)^{\theta}}, \tag{17}$$

where $\theta \geq 0$ corresponds to the extent of the belief distortion. Incorrect stereotypes are most likely to form when there are group differences in the frequency of a particular type, but the overall type distributions are largely the same. This is consistent with recent empirical work that finds support for the model (**???**).

Here, we explore how 'representativeness' can lead to biased beliefs in our setting. We examine the distribution of users' reputation earned per answer post over the *entire* range of reputations at time of posting. Since we do not observe evidence for discrimination on answers posted to low reputation accounts in either the experiment or the observational data, we use the evaluation of answers as a proxy for ability. We divide the distribution of reputation earned per answer post into quintiles by gender. The distributions are fairly similar across male and female usernames: the median corresponds to the 3^{rd} quintile for both male and female users, with the mean equal to 2.97 for males and 2.87 for females. The difference in means is fairly small, representing 6% of a standard deviation of the average quintile position, and is only marginally significant. However, using these means as estimates of the perceived means of ability ($\hat{\mu}_F$ and $\hat{\mu}_M$ from the theory model), we see that even mild belief distortions due to 'representativeness' quickly exacerbate this small underlying difference.

Figure 3 illustrates the difference between perceived means of males and females as a function of the degree of distortion θ caused by the stereotype heuristic. While the perceived means are fairly similar when the distortion is minimal (θ=0), under moderate levels of distortion (for example, $\theta = 2.5$ estimated in prior studies (**?**)), the difference in perceived means triples to nearly half a quintile. As shown in Section I in the manuscript, if even a small proportion of individuals hold such distorted beliefs, this can lead to a dynamic reversal of discrimination.

[5]This stereotype is incorrect – the overall age distribution of Floridians is quite similar to the rest of the country, and the majority of Floridians are under 65.

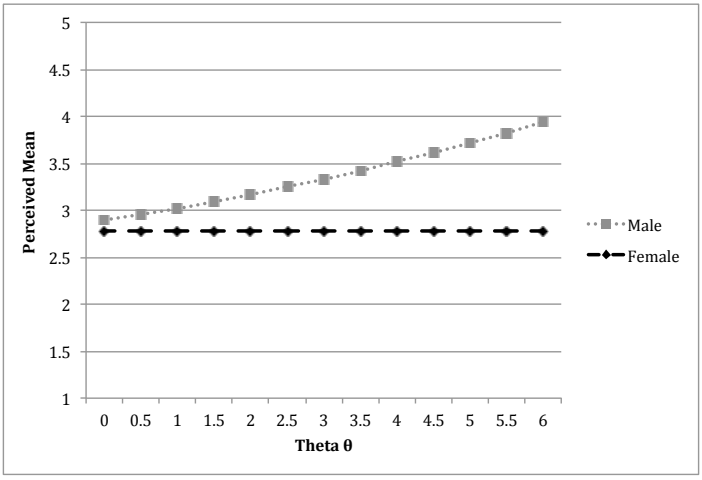

Figure 3. Subjective average ability by gender $\hat{\mu}_g$ as function of θ.