The Dynamics of Discrimination: Theory and Evidence[*]

J. Aislinn Bohren[†] Alex Imas[‡] Michael Rosenberg[§]

January 2019

We model the dynamics of discrimination and show how its evolution can identify the underlying source. We test these theoretical predictions in a field experiment on a large online platform where users post content that is evaluated by other users on the platform. We assign posts to accounts that exogenously vary by gender and evaluation histories. With no prior evaluations, women face significant discrimination. However, following a sequence of positive evaluations, the direction of discrimination *reverses*: women's posts are favored over men's. Interpreting these results through the lens of our model, this dynamic reversal implies discrimination driven by biased beliefs.

KEYWORDS: Discrimination, Dynamic Behavior, Field Experiment
JEL: J16, D83, D9

[*] We thank Nageeb Ali, Linda Babcock, Michael Callen, Hanming Fang, Uri Gneezy, Polina Imas, Nicola Gennaioli, Gregor Jarosch, Emir Kamenica, Gene Kucher, George Loewenstein, Kristof Madarasz, Craig McIntosh, Margaret Meyer, Siqi Pan, Sally Sadoff, Lise Vesterlund, Leeat Yariv, and seminar and conference participants at the AEA Winter Meetings, Bocconi University, Carnegie Mellon, Drexel, ESSET Gerzensee, Harvard, SJDM Annual Meeting, NYU, Northwestern, Penn State, Princeton, Stanford SITE, Temple University, University of Chicago, University of Pennsylvania, UCLA, UCSD and University of Pittsburgh for helpful comments. We thank Mustafa Dogan, Daniel Hauser, Conrad Kosowski, Catherine Lillis, Suneil Parimoo, Jaesung Son, and Lucia Zhang for excellent research assistance. Finally, we would like to thank the team at Stack Overflow for helpful and illuminating discussions, and for providing us with the data for our analyses. The project received IRB approval at Carnegie Mellon and University of Pennsylvania. The experiment was pre-registered in the AEA RCT Registry (#AEARCTR-0000950).

[†] Carnegie Mellon University and University of Pennsylvania; corresponding author: abohren@sas.upenn.edu

[‡] Carnegie Mellon University

[§] Wayfair Inc.

1 Introduction

A rich literature has documented discrimination in a wide range of contexts (Bertrand and Duflo 2016). These empirical studies have mostly focused on static settings: individuals are evaluated based on the quality of a single piece of output or a single interaction, with no information on prior evaluations in similar contexts. As prior work has noted, it is difficult to identify the underlying source of discrimination from such static settings, as different sources generate the same patterns of observable behavior (Fang and Moro 2011). In this paper, we develop a theoretical framework to show how the dynamics of discrimination can be used to identify its underlying source, and test these predictions in a field experiment on a large online platform.

Consider a setting where individuals repeatedly perform tasks that generate output, and in the process, produce an observable history of evaluations on these tasks. For example, a man and a woman are employed at a firm and are promoted based on how managers evaluate their output. Their past promotions and performance evaluations correspond to the history of evaluations. Alternatively, workers contribute to crowdsourcing projects on a platform, such as GitHub. Each worker has an observable reputation score based on prior evaluations of his or her contributions. In such settings, when workers are starting out and lack evaluations of prior performance, initial discrimination occurs if a female worker's output is less likely to earn a promotion or receive a positive evaluation than a male's, despite the appearance of similar quality. Suppose new workers continue producing output, and receive similar sequences of evaluations. Does discrimination persist in this dynamic setting, is it mitigated, or does it even *reverse*?

The answer to this question depends critically on the underlying source of discrimination. If the source is belief-based – for example, the quality of output is imperfectly observed and evaluators believe that on average, men have higher abilities than women – then observing prior evaluations will reduce discrimination against women, relative to men with similar evaluations. This dynamic effect operates through two channels. First, prior evaluations provide signals of a worker's ability, which reduces the impact of perceived group statistics (e.g. beliefs about average ability) on how the worker's subsequent output is evaluated.[1] This mitigates discrimination between males and

[1] This is the channel typically considered in the literature on accurate statistical discrimination, i.e. belief-based discrimination with correct beliefs (e.g. Altonji and Pierret (2001)). The discrimination

females with similar evaluation histories. Second, and particular to a social learning setting, the informational content of these signals is *endogenously* determined by the behavior of prior evaluators. When initial beliefs favor men, to overcome this disparity, a woman needs to produce *higher* quality output than a man to receive a similar evaluation – for example, to be promoted or have her output accepted. This speeds up the mitigation of discrimination for evaluators who are aware of the higher standard for women. These evaluators may even come to believe that the woman is of higher ability than a man with a similar evaluation history – thereby favoring her output over the man's and *reversing* the direction of discrimination in subsequent periods. In fact, observing a reversal can help disentangle whether evaluators' models are correct or misspecified; we show theoretically that a reversal provides evidence for bias. In contrast to belief-based sources, if discrimination is caused by a taste or preference against rewarding or interacting with women (Becker 1957), then a woman who receives a similar sequence of evaluations to a man will continue to face discrimination in future periods.

Our theoretical framework formalizes the relationship between the dynamic pattern of discrimination, which is based on observable evaluations, and the sources of discrimination, which are unobservable and depend on underlying preferences and beliefs. The literature on belief-based sources has generally focused on correct beliefs (e.g. rational expectations), where evaluators are partial towards a group based on true differences in the underlying distributions of the relevant attribute (Altonji and Pierret 2001; Fang and Moro 2011; Knowles, Persico, and Todd 2001; Phelps 1972). However, recent research has demonstrated that systematic biases in judgment can lead to incorrect stereotypes against a particular group (Bordalo, Coffman, Gennaioli, and Shleifer 2016b; Fryer and Jackson 2008; Schwartzstein 2014). Therefore, we allow for three potential sources: (i) belief-based with correct beliefs, (ii) belief-based with incorrect, biased beliefs, and (iii) preference-based. We show that these sources make contrasting dynamic predictions. When discrimination is based on common knowledge of correct beliefs, then observing similar sequences of evaluations for a man and a woman will mitigate discrimination, but will never lead to a reversal. Therefore, observing a dynamic reversal provides evidence for a belief-based source with *bias*, since

literature in social psychology also discusses the role of individual-specific information in reducing reliance on using group statistics for judgment (see Fiske (1998) for review).

it is also inconsistent with standard preference-based sources. We also illustrate how one form of bias – where some evaluators hold incorrect stereotypes against a group and other evaluators are aware of these stereotypes – can lead to a dynamic reversal.

Our framework also formalizes how a second informational channel – the level of subjectivity in evaluation (Fiske, Bersoff, Borgida, Deaux, and Heilman 1991), modeled as the precision in signals of quality – provides further evidence to disentangle the source of discrimination. Specifically, decreasing the subjectivity of evaluations will mitigate belief-based discrimination, as beliefs about group statistics play a smaller role in assessing quality when signals of quality are more precise. But it will not affect preference-based discrimination, which will persist even if quality is perfectly observable. As we later discuss, identifying the underlying source of discrimination has significant implications for policy and welfare.

We test these theoretical predictions using a field experiment on a large online Q&A forum. The forum is a prominent resource for students and researchers in STEM fields – it has nearly 350,000 users, and belongs to a family of Q&A forums that has over 3 million questions asked and 4 million answers posted per year – which makes documenting the existence and source of gender discrimination in this setting particularly important. Users post mathematics questions or answers, and these posts are evaluated – voted up or down – by other users on the site. A user's reputation provides a summary statistic of evaluations of his or her past posts: higher reputation corresponds to more positive and fewer negative votes. Importantly, reputation is publicly observable and highly visible. Both the username and the level of reputation are prominently displayed adjacent to any post. Since reputation is generated by prior evaluations, this setting mirrors the social learning in our theoretical framework. Reputation is also valuable – it can be used as currency to *pay* other users for providing answers and it promotes users to higher ranks on the forum, opening the door to additional privileges. This includes privileges to "supervise" other users – for example, to edit, flag and close other users' posts. Similar to promotion decisions within a firm, evaluations are consequential because reputation gives users greater influence over the evaluators. Therefore, the link between evaluations and advancement on the forum mirrors many labor market settings.

In our experiment, we posted original mathematics questions on created accounts that exogenously vary in the gender of the username and the reputation of the user. Our setting is well-suited for exploring the dynamics of discrimination because we are able to exogenously vary the evaluation histories of users, as summarized by their publicly observable reputations. We posted half of the questions to novice accounts that did not have prior evaluations. We manually built the reputations of the remaining accounts by posting content until the reputations reached the top 25^{th} percentile on the forum. We then randomly reassigned the gender of the username to avoid endogeneity issues and ensure that the underlying informational content of reputation is the same for both genders. Finally, we posted the remaining questions to these advanced accounts. We compare the pattern of discrimination between novice and advanced question posts to test the dynamic predictions of the different sources of discrimination.

We also posted answers to other users' posts from a second set of novice accounts that exogenously vary the gender of the username. This allows us to test the comparative static on how the level of subjectivity involved in judging posts (e.g. the precision of the signal) affects discrimination. While the forum's guidelines for voting on questions are based on fairly subjective criteria – whether the question is interesting, useful, or well-researched – the guideline for voting on answers is clear-cut – whether the answer is correct or not. If the source is preference-based, this distinction will not affect discrimination: our model predicts similar levels of discrimination for both question and answer posts. In contrast, if discrimination is belief-based, then our model predicts that reducing subjectivity will mitigate it: answer posts will face less discrimination than question posts.[2]

We measure discrimination as the difference in reputation earned or net votes on posts by accounts with male versus female usernames. We find no significant discrimination on answer posts: answers posted by females with no prior evaluations earned a similar amount of reputation and received a similar number of positive votes as answers posted by males with no evaluations. In contrast, we find that females face *significant* initial discrimination when the judgment of quality is more subjective: questions posted

[2] An evaluator who has a preference against women but does not want to appear discriminatory – either to himself or others – may also discriminate less on objective quality dimensions, as such discrimination is more obvious (e.g. moral wiggle room (Dana, Weber, and Kuang 2007)). Two features of our experimental setting suggest that this phenomenon may be less likely to emerge: (i) evaluators are anonymous, removing the motivation to signal to others; and (ii) discrimination mostly occurs along the margin of choosing whether to upvote or not evaluate a post. We observe few downvotes, and moral wiggle room is typically conceptualized as an avoidance of salient negative actions.

to female accounts with no prior evaluations are evaluated less favorably – earning less reputation and fewer positive votes – than questions posted to similar male accounts. Directly comparing questions and answers produces a significant interaction, indicating greater discrimination against females when judgments of quality are more subjective. This is consistent with belief-based but not preference-based discrimination. We also find significant discrimination on questions posted to advanced accounts, but the direction of discrimination *reverses*: questions posted to advanced female accounts earn *more* reputation than those posted by similarly advanced males. This produces a significant interaction effect between the user's rank on the forum (Novice or Advanced) and gender. Interpreting these results through the lens of our model suggests that initial discrimination is belief-based, with bias playing a role in the evaluation process.

In addition to our experimental results, we exploit two additional data sources: a proprietary dataset that contains additional information about the users who evaluated the content from our experiment, and a large observational dataset of all posts on the forum. We used these datasets to run additional robustness tests and to rule out other potential explanations for the observed reversal, such as gender differences in attrition or the variance of ability. We also compare discrimination by type of post and reputation in the observational data. We find analogous patterns to the experiment, including both the dynamic reversal between questions posted by novice and advanced users and the lack of discrimination for answers.

The findings presented here highlight the importance of studying discrimination in dynamic settings, as discrimination in favor of a certain group – or a lack thereof – at any given stage can either be a function of or precursor to discrimination against that same group at a different stage. Both in academic and popular discourse, a common argument used to illustrate the *lack* of discrimination against a group is to point to individuals from that group who have made it to positions of prominence. Our theoretical framework and empirical evidence highlight the flaw of this argument: if individuals are aware that members of a group face discrimination at an earlier stage, there may be Bayesian foundations for favoring members of that group at later stages. For example, in a much-discussed paper, Williams and Ceci (2015) find that accomplished female academics in STEM fields are favored over male academics. The authors state that 'these results suggest it is a propitious time for women launching careers in academic science.' In contrast, other work has found significant discrimination against female students in STEM (Moss-Racusin, Dovidio, Brescoll, Graham, and Handelsman 2012; Reuben, Sapienza, and Zingales 2014). While these sets of findings appear contradictory, our results suggest that discrimination in favor of accomplished female professors may actually be a function of discrimination *against* women earlier in the pipeline.

Our conceptual and experimental framework can be applied to many other labor market settings. Settings where individuals offer a product and can be identified by their gender and prior history of evaluations are becoming progressively more widespread and economically important. Stack Exchange, GitHub, TaskRabbit, Upwork and AirBnB are just a few examples of such platforms. Our framework is also relevant for settings in which prior work has found reversals of discrimination between the hiring and promotion stages *within* a firm, and it provides a possible explanation for the "female leadership premium" (again, within a firm) that has been documented in the management literature (see the Related Literature section for further discussion).

Our results are also useful for assessing the welfare consequences of discrimination. While the welfare implications of discrimination driven by preferences or correct beliefs are unclear, the welfare implications of discrimination caused by biased beliefs are more straightforward – in our setting, biased beliefs lead to distorted evaluations. Even if a discrimination reversal occurs, so that women eventually receive higher evaluations than men with similar *evaluation* histories, these women still receive lower evaluations than men with similar *output quality* histories. In other words, the reversal does not offset initial discrimination: a woman who is favored over a man with a similar evaluation history should receive an even higher evaluation than she does, given correct beliefs about her expected ability. Perhaps even more importantly, women may inefficiently stagnate at lower stages than men with similar abilities due to initial discrimination. That is, women and men with similar output histories will not achieve the same level of success – the women will be systematically *underrated*. Therefore, hiring and promotion decisions based on these evaluations will be suboptimal, particularly when future evaluators are also biased or are not aware of the bias of prior evaluators.[3]

Finally, our results highlight the importance of considering the dynamic impact of

[3] More generally, in learning settings with no action interdependence, an individual with an incorrect belief about the prior distribution of ability or informational content of evaluations will make suboptimal choices, relative to an individual with correct beliefs. This contrasts with settings with action interdependence, in which the effect of incorrect beliefs is ambiguous: when correct beliefs lead to a market failure, it may be possible for incorrect beliefs to improve the market outcome.

interventions that aim to reduce discrimination, particularly in regards to how these interventions impact beliefs. For example, evidence suggests that individuals systematically *overestimate* the prevalence of affirmative action policies and the extent to which they lower evaluation standards (Kravitz and Platania 1993). An intervention that leads to perceived lenient standards at one stage will impact assessments at later stages, and can even lead to greater subsequent discrimination. This highlights the importance of accurately informing the population who evaluates members of a target group about the scope of such interventions.

Related Literature. Discrimination has been documented in a wide range of settings, including hiring (Bartos, Bauer, Chytilova, and Matejka 2016; Riach and Rich 2006), housing (Ewens, Tomlin, and Wang 2014), and service markets (Gneezy, List, and Price 2012). It has also been documented against group identities based on race (Bertrand and Mullainathan 2004; Parsons, Sulaeman, Yates, and Hamermesh 2011), ethnicity (Fershtman and Gneezy 2001; Milkman, Akinola, and Chugh 2012) and gender (Goldin and Rouse 2000; Moss-Racusin et al. 2012). The few studies that use observational data to attempt to identify the source of discrimination typically compare the evaluations of a state (for example, whether output is accepted or rejected) to the true underlying value of that state (i.e. whether the output was actually high or low quality). For example, Knowles et al. (2001) compare decisions of law enforcement to search a motor vehicle to the success rate of the search; similarities in success rates across races led the authors to conclude that higher search rates for African-American drivers are due to statistical rather than preference-based discrimination (see Anwar and Fang (2006); Arnold, Dobbie, and Yang (2017) for similar tests). In recent work, Sarsons (2017) uses an event study approach for matched samples of surgeons to explore belief-based gender discrimination in physician referrals. She concludes that the observed pattern of gender discrimination is not consistent with Bayesian learning with respect to accurate beliefs about the distribution of surgeon ability. However, in many observational settings, it is difficult or impossible to construct matched samples or to observe the true value of the underlying state at an individual level. Further, observational data often faces endogeneity issues that preclude the causal identification of discrimination.

Due to endogeneity issues, many researchers have employed field experiments to study discrimination. Field experiments have been successful in causally identifying the incidence of discrimination, but most cannot identify the source of this discrimination (Bertrand and Duflo 2016). One notable exception is List (2004), who documents that minorities receive inferior initial and final offers when bargaining in a market for sports cards. He supplements these data with a series of artefactual and framed field experiments to identify the source of this discrimination. Data from dictator games, market and auction experiments provide support for belief-based discrimination, and rule out preference-based sources. This method demonstrates how eliciting the true value of the underlying state for different groups (e.g. the distribution of reservation prices across races) can identify the source of discrimination in a static setting. We provide a complementary approach that illustrates how dynamic data and variation in the subjectivity of judgement can be used to achieve the same goal.

Our findings shed light on the mechanism behind previously documented discrimination reversals. In labor market settings, Booth, Francesconi, and Frank (1999); Groot and van den Brink (1996); Lewis (1986); Petersen and Saporta (2004) find discrimination against women at the initial hiring stage for promotable jobs, but conditional on being hired, they find that women are more likely to be promoted. Rosette and Tost (2010) document a female leadership premium, showing that in contrast to women at lower levels within an organization, women in high positions are seen as more effective than men at similar positions.[4] In a field experiment, Ayalew, Manian, and Sheth (2018) show that workers are more likely to follow a man's advice than a woman's; however, this result reverses when they are informed that the woman or man has achieved a high level position in a job outside of the experiment. In the art market, Bocart, Gertsberg, and Pownall (2018) document that while female artists are less likely to transition from primary to secondary art markets, those who do command a 4.4% premium on artworks sold. In academia, Mengel, Sauermann, and Zolitz (2017) find that junior female instructors systematically receive lower teaching evaluations compared to male instructors for similar courses, but at the senior level, female instructors receive higher evaluations than male instructors. While these results could be driven by institutional factors, our theoretical and empirical findings suggest that the reversals

[4] Leslie, Manchester, and Dahm (2017) argue that this leadership premium extends to perceived potential as well. In their paper, women who are perceived to be able to rise through the ranks are judged to add more value to the company than men with similarly high potential; in contrast, low potential women are judged to add less value than low potential men. Importantly, substantially *fewer* women are judged as being able to rise through the ranks than men. See also Beaman, Chattopadhyay, Duflo, Pande, and Topalova (2009) for the effect of exposure to female leaders on perceived effectiveness.

may be driven by belief-based discrimination with bias (for example, biased priors or *stereotypes*). Consistent with this mechanism, Mengel et al. (2017) find that initial discrimination against females is higher in courses with math-related content, where distorted gender stereotypes are more likely to play a role (Coffman 2014).

The paper proceeds as follows. Section 2 presents the theoretical model, Section 3 presents the experiment and analysis of observational data, while Section 4 discusses the implications for policy and concludes. All proofs are in Appendix A.1.

2 A Dynamic Model of Discrimination

We develop a dynamic model of discrimination in which evaluators learn about a worker's ability from group identity and past performance, and use this information to evaluate the quality of the worker's current output. To mirror our experiment, we use gender as the group identity in our model and focus on discrimination against F(emales) compared to M(ales).

We first set-up the model and formalize the definitions of the underlying belief and preference-based sources of discrimination (Section 2.1), then briefly comment on notable features of our setting, including how we choose to model belief-based discrimination, incorrect beliefs and the subjectivity of judgement (Section 2.2). In Section Section 2.3, we characterize how beliefs and preferences impact initial evaluations, and show that varying the level of subjectivity in judgement can identify whether discrimination is due to preference- or belief-based sources (Proposition 1). In Section 2.4, we characterize how discrimination evolves across time. This yields two main results: Proposition 2 establishes the impossibility of a discrimination reversal when all evaluators have common knowledge of correct beliefs, while Proposition 3 demonstrates how one form of biased beliefs can lead to a reversal. A reader who prefers to skip the formal presentation of the theory can jump to the empirics in Section 3.

2.1 Model

Worker. Consider a worker who has observable group identity $g \in \{F, M\}$ and unobservable ability $a \sim N(\mu_g, 1/\tau_a)$, with mean $\mu_g \in \mathbb{R}$ and precision $\tau_a > 0$. The worker completes a sequence of tasks $t = 1, 2, ...$ Each task has hidden quality $q_t = a + \varepsilon_t$, where $\varepsilon_t \sim N(0, 1/\tau_\varepsilon)$ is an independent random shock with precision $\tau_\varepsilon > 1$. Ability is fixed across time, and higher ability generates higher expected quality.

Evaluators. A set of evaluators assess the worker's performance. For simplicity, assume that there is one evaluator per task, who reports evaluation $v_t \in \mathbb{R}$.

Histories and Signals. Before evaluating task t, the evaluator observes the worker's gender g, evaluations on past tasks $h_t = (v_1, ..., v_{t-1})$, where $h_1 = \emptyset$, and signal $s_t = q_t + \eta_t$ of quality of the current task, where $\eta_t \sim N(0, 1/\tau_\eta)$ is an independent random shock with precision $\tau_\eta > 0$. Lower signal precision reflects greater uncertainty in quality. This precision can be interpreted as the amount of subjectivity in judgement involved in the evaluation of quality, with lower precision implying greater subjectivity. We motivate and discuss this interpretation in further detail in Section 2.2.

Preferences and Beliefs. An evaluator's type θ_i determines her preferences and model of inference, including her subjective belief about the relationship between gender and ability and her subjective belief about other evaluators' preferences and beliefs. The evaluator receives payoff $-(v - (q - c_g^i))^2$ from reporting evaluation v on a task of quality q from a worker of gender g, where c_g^i is a type-specific taste parameter. Normalize $c_M^i = 0$. The evaluator has subjective prior belief $\hat{\mu}_g^i$ about the average ability of a worker of gender g. We allow for the possibility that the evaluator has a misspecified model of the relationship between gender and ability, in that the evaluator's subjective belief may differ from the true population average ability, $\hat{\mu}_g^i \neq \mu_g$.

An evaluator is *partial* towards men if she favors male workers, either through her subjective belief about the distribution of ability by gender, which we refer to as *belief-based partiality*, or through preferences, which we refer to as *preference-based partiality*. In the first case, an evaluator has a 'taste' favoring male workers, meaning that she has a disamenity value associated with tasks produced by female workers.

Definition 1 (Preference-Based Partiality)**.** *An evaluator of type θ_i has a* preference-based partiality *towards men if $c_F^i > 0$.*

In the second case, the evaluator believes that the average ability of male workers is higher than the average ability of female workers. Belief-based partiality can be biased or unbiased, based on whether it coincides with the true population average for each gender.

Definition 2 (Belief-Based Partiality)**.** *An evaluator of type θ_i has* belief-based partiality *towards men if $\hat{\mu}_M^i > \hat{\mu}_F^i$. This partiality is* unbiased *if $\hat{\mu}_M^i = \mu_M$ and $\hat{\mu}_F^i = \mu_F$, and otherwise is* biased.

Finally, in order to interpret the evaluation history, which consists of the assessments of other evaluators, the evaluator needs a model of other evaluators' preferences and beliefs. This is captured by her subjective belief about the distribution over types, $\hat{\pi}_i \in \Delta(\Theta)$, where Θ denotes the finite set of evaluator types. Let $\pi \in \Delta(\Theta)$ denote the true distribution over types. A misspecified model of how others evaluate workers is captured by a subjective belief about the type distribution that differs from the true distribution, $\hat{\pi}_i \neq \pi$. We discuss settings this framework can capture in Section 2.2.

Aggregate Beliefs. It is straightforward to define *aggregate* analogues of partiality with respect to the average beliefs and preferences of evaluators. There is aggregate belief-based partiality towards men if $E_\pi[\hat{\mu}^i_M] > E_\pi[\hat{\mu}^i_F]$ and aggregate preference-based partiality towards men if $E_\pi[c^i_F] > 0$, where the expectation is taken with respect to the true distribution over types. Aggregate belief-based partiality is unbiased if $E_\pi[\hat{\mu}^i_M] = \mu_M$ and $E_\pi[\hat{\mu}^i_F] = \mu_F$, and otherwise is biased. It is possible for individual types to exhibit partiality or bias, but for aggregate preferences and beliefs to be impartial or unbiased.[5]

Belief-Updating. The evaluator learns about the worker's ability from the evaluation history. Her posterior belief about ability is derived using Bayes rule, given her model of inference. She combines this updated belief about ability with the signal to learn about the quality of the current task, also using Bayes rule to form her posterior belief about quality.

Optimal Evaluations. Each evaluator chooses the evaluation that maximizes her expected payoff with respect to her posterior belief about quality. Suppose an evaluator has type θ_i and let

$$v_i(h, s, g) \equiv \arg\max_{v \in \mathbb{R}} \hat{E}_i[-(v - (q - c^i_g))^2 | h, s, g] \tag{1}$$

denote her optimal evaluation conditional on observing history h and signal s from a worker of gender g, where $\hat{E}_i$ denotes the expectation with respect to her model of inference. Then her optimal evaluation in period t is

$$v_i(h_t, s_t, g) = \hat{E}_i[q_t | h_t, s_t, g] - c^i_g. \tag{2}$$

[5] For example, suppose each type's prior belief about average ability is the true mean plus an idiosyncratic error. This would result in partiality at the individual level, in that some evaluators are partial towards men and others are partial towards women, but no aggregate partiality.

Discrimination. Discrimination is the disparate evaluation of workers based on the group to which the worker belongs, i.e. gender, rather than on individual attributes, i.e. signal and history. In our framework, gender discrimination occurs when a male and female worker with the same evaluation history and current signal receive different evaluations. Let

$$D_i(h, s) \equiv v_i(h, s, M) - v_i(h, s, F) \tag{3}$$

denote the difference between type θ_i's evaluation of a male and female worker conditional on observing history h and signal s, and let $D(h, s) \equiv E_\pi[D_i(h, s)]$ denote the expected difference in evaluations across all types.

Definition 3 (Discrimination). *A woman faces* discrimination *from type θ_i at (h, s) if $D_i(h, s) > 0$, and faces* aggregate discrimination *if $D(h, s) > 0$. A man faces (aggregate) discrimination if $D_i(h, s) < 0$ $(D(h, s) < 0)$.*

In contrast to partiality, which is a property of the primitives of the model (preferences, beliefs), discrimination is a property of behavior.

In this paper, we study whether discrimination *reverses* between histories.

Definition 4 (Discrimination Reversal). *A discrimination* reversal *occurs at history h and signal s if there exists a history $h' \subset h$ such that women face discrimination at (h', s) and men face discrimination at (h, s).*[6]

For example, a discrimination reversal occurs if women face initial discrimination at history $h_1 = \emptyset$, while men face discrimination at history $h_2 = \{v_1\}$ following some evaluation v_1. We also study whether discrimination decreases – for example, between histories or across parameters – which corresponds to a decrease in $|D(h, s)|$.

In the following sections, we explore how the different forms of partiality impact discrimination. We use these insights to illustrate how observable behavior (i.e. evaluations) can be used to identify the *source* of discrimination (i.e. preferences, beliefs).

2.2 Discussion of Model

Here, we discuss several features of the model and the types of settings it can capture.

[6] Given histories $h' = (v_1, ..., v_m)$ and $h = (v'_1, ..., v'_n)$, we say $h' \subset h$ if $m < n$ and the histories have the same first m evaluations, i.e. $v_i = v'_i$ for all $i \leq m$.

Misspecified Models of Inference. The set-up for the evaluator's model of inference builds on the framework of social learning with model misspecification developed in Bohren and Hauser (2018). This framework can capture broad classes of model misspecification, including an incorrect model of the relationship between ability and gender and an incorrect model about other evaluators' preferences or beliefs. For example, the setting where all evaluators have common knowledge that they share the same preferences and beliefs is captured by a single type θ_1, $\pi(\theta_1) = 1$, who correctly believes that all other evaluators are this type, $\hat{\pi}_1(\theta_1) = 1$. This type's subjective belief about average ability by gender may or may not be correct.

Alternatively, there may be heterogeneity in evaluators. For example, some evaluators may use a heuristic to form beliefs about the relationship between ability and gender, while other evaluators have a correct belief about average ability by gender. Evaluators who use a heuristic are likely not aware of their bias – otherwise, they would correct for it – and believe that other evaluators form beliefs in a similar manner (the bias blindspot (Pronin, Lin, and Ross 2002) or the false consensus effect (Ross, Greene, and House 1977)). Our framework can model this setting using a type θ_1 that has an incorrect subjective belief about average ability by gender, and an incorrect subjective belief that other evaluators are the same type, $\hat{\pi}_1(\theta_1) = 1$. The other type θ_2 has a correct subjective belief about the ability distribution for males and females, and can either accurately anticipate the presence of the biased type, $\hat{\pi}_2(\theta_1) = \pi(\theta_1)$, be unaware of the biased type, $\hat{\pi}_2(\theta_1) = 0$, or under- or overestimate its frequency, $0 < \hat{\pi}_2(\theta_1) < \pi(\theta_1)$ or $\hat{\pi}_2(\theta_1) > \pi(\theta_1)$. This type could also be aware that some evaluators are biased, but not understand the exact extent of the bias. Importantly, when there is heterogeneity in evaluators' subjective beliefs about the relationship between gender and ability, then at least one type has a misspecified model of inference.

Belief-Based Discrimination. Theories of belief-based discrimination have typically focused on rational, or *statistical*, discrimination, where evaluators hold correct beliefs about aggregate group differences. These models fall into two broad categories that differ primarily in how group differences in beliefs arise – whether (i) group differences are exogenous and discrimination is due to imperfect information (Phelps 1972), or (ii) group differences are "self-fulfilling" and discrimination is an equilibrium effect

(Arrow 1973).[7] In the first class of models, evaluators hold prior beliefs about workers' abilities that differ by group identity, and use these group statistics to infer individual ability (Aigner and Cain 1977; Altonji and Pierret 2001; Lundberg and Startz 1983). Our model with correctly specified evaluators falls into this class. In the second class of models, ex-ante *identical* workers decide whether to engage in costly and unobservable skill acquisition. Discrimination arises when workers from different groups coordinate on equilibria with different levels of skill acquisition (Coate and Loury 1993; Fryer 2007). In contrast to the first class of models, there are also always equilibria in which both men and women acquire the same level of skill, and evaluators treat them identically.

Belief-based discrimination can also arise from systematically incorrect, or *biased*, beliefs, where evaluators hold *misspecified* models of group differences in the distributions of ability. Several models provide microfoundations for how such biased beliefs about group differences can arise and persist. Evaluators may form biased stereotypes of ability as a result of using the representative heuristic that exaggerate empirical reality (Bordalo et al. 2016b), due to selective attention that discounts how, for example, context affects behavior (Schwartzstein 2014), or because of course categorization of experiences with a particular group (Fryer and Jackson 2008). In our setting, such stereotyping corresponds to distortions in the subjective belief about average ability, $\hat{\mu}_g$. As also noted in Schwartzstein (2014), the discrimination literature has tended to classify discrimination driven by distorted stereotypes as taste-based.[8] However, we demonstrate that biased beliefs lead to patterns of discrimination that substantially differ from those that arise in taste-based models in which evaluators have animus towards a particular group (i.e. preference-based partiality). This is one reason we clearly distinguish between discrimination due to incorrect beliefs and discrimination due to preferences.

Subjectivity of Judgment. Uncertainty over the assessment criteria – which we refer to as subjectivity in judgment – increases the variance of potential evaluations for

[7] See Fang and Moro (2011) for a more thorough review of this literature.

[8] For example, Price and Wolfers (2010) suggest that their findings of own-race partiality of basketball referees are not driven by a preference against members of a particular group, but rather by implicit associations between race and the likelihood of violence. Such discrimination is classified as taste-based, because beliefs about these associations influence behavior subconsciously (Bertrand, Chugh, and Mullainathan 2005; G. Greenwald, E. McGhee, and L. K. Schwartz 1998).

a given level of an attribute (Olson, Ellis, and Zanna 1983) and reduces the expected consensus between evaluators (Kelley 1973). The social psychology literature argues that such subjectivity is "quite vulnerable to stereotypic biases" (Fiske et al. 1991) and increases the scope for discrimination (Biernat, Manis, and Nelson 1991; Danilov and Saccardo 2017; Snyder, Kleck, Strenta, and Mentzer 1979). Indeed, researchers have documented greater reliance on beliefs about group statistics when judgment is more subjective (see Fiske and Taylor 1991, for review). As judgement becomes more objective, the available information provides more precise signals about the underlying attribute. This decreases the reliance on group statistic in forming assessments, and therefore, reduces the potential for belief-based discrimination. Target groups anticipate greater scope for discrimination when judgment is more subjective, and in response, generate output with more objective assessment criteria (Parsons et al. 2011).

We model the level of subjectivity in judgment as the precision of the signal of quality, τ_η. Factors that increase subjectivity, such as uncertainty over the evaluation criteria and noisier information sources, decrease the precision of the signal. Our theoretical results match the empirical findings on subjective judgment: we will show that a decrease in signal precision leads to greater reliance on beliefs about group statistics to assess quality, and therefore, greater scope for belief-based discrimination.

2.3 Initial Discrimination

We first compare how belief- and preference-based partiality impact initial evaluations. We show that a comparative static on how initial discrimination varies with the subjectivity of judgement (i.e. the precision of the signal) can distinguish between these two sources.

Consider the evaluation of the first task from a worker of gender g by an evaluator who has subjective prior beliefs $(\hat{\mu}_F, \hat{\mu}_M)$ about average ability, preference parameter c_F, and observes signal s_1. Given these prior beliefs about ability, the evaluator's prior belief about quality is normally distributed with mean $\hat{\mu}_g$ and precision $\tau_q \equiv \tau_a\tau_\varepsilon/(\tau_a + \tau_\varepsilon)$, i.e. $q_1 \sim N(\hat{\mu}_g, 1/\tau_q)$. The initial signal has conditional distribution $s_1|q_1 \sim N(q_1, 1/\tau_\eta)$. Given the prior belief and signal distribution, the evaluator's posterior belief about quality conditional on observing s_1 is also normally distributed,

$q_1|s_1 \sim N\left(\frac{\tau_q\hat{\mu}_g+\tau_\eta s_1}{\tau_q+\tau_\eta}, \frac{1}{\tau_q+\tau_\eta}\right)$. From (2), the optimal evaluation is equal to

$$v(h_1, s_1, g) \;=\; \frac{\tau_q\hat{\mu}_g + \tau_\eta s_1}{\tau_q + \tau_\eta} - c_g. \tag{4}$$

Higher signals and higher expected ability result in higher evaluations – the optimal evaluation is strictly increasing in s_1 and $\hat{\mu}_g$.

Initial discrimination depends on the evaluator's preferences and prior beliefs about ability. From (4), initial discrimination is independent of the signal and equal to

$$D(h_1, s_1) = \left(\frac{\tau_q}{\tau_q + \tau_\eta}\right)(\hat{\mu}_M - \hat{\mu}_F) + c_F. \tag{5}$$

There is initial discrimination against females, i.e. $D(h_1, s_1) > 0$, if and only if the evaluator has belief-based or preference-based partiality, $\hat{\mu}_F < \hat{\mu}_M$ or $c_F > 0$. Therefore, discrimination on the first task stems from an evaluator's own partiality; it does not depend on her beliefs about the partiality of other evaluators.

It is not possible to identify the source of discrimination from observing initial evaluations at a single set of parameters. For any level of preference-based partiality, there exists a level of belief-based partiality that leads to equivalent initial evaluations and discrimination, and vice versa.[9] Therefore, we need a richer cross-section of evaluations to identify the source.

Our first result shows that varying the level of subjectivity in judgement differentially impacts discrimination depending on whether it is due to preference- or belief-based partiality. This comparative static can be used to identify the source of discrimination.

Proposition 1 (Subjectivity of Judgement)**.** *If the evaluator has belief-based partiality, initial discrimination is decreasing in the precision of the signal τ_η and otherwise, initial discrimination is constant with respect to τ_η. As the signal becomes perfectly objective, $\tau_\eta \to \infty$, there is initial discrimination if and only if the evaluator has preference-based partiality.*

[9] For any evaluator with beliefs $\hat{\mu}^1_M > \hat{\mu}^1_F$ and preference parameter $c^1_F = 0$, an evaluator with preference parameter $c^2_F = \frac{\tau_q(\hat{\mu}^1_M - \hat{\mu}^1_F)}{\tau_q+\tau_\eta} > 0$ and beliefs $\hat{\mu}^2_F = \hat{\mu}^2_M = \hat{\mu}^1_M$ chooses equivalent evaluations and exhibits an equivalent level of discrimination. This follows immediately from (4) and (5). Note that the first evaluator has belief-based partiality and the second has preference-based partiality.

As the signal provides more precise information about quality, the evaluator's belief about the worker's underlying ability has a smaller impact on the evaluation. Therefore, differences in beliefs about ability, i.e. belief-based partiality, translate into smaller differences in evaluations and less discrimination. In the limit, when quality is perfectly observable, differences in beliefs about ability do not lead to discrimination: although an evaluator with belief-based partiality expects lower quality from female workers ex-ante, male and female workers who generate the same signal receive identical evaluations. In contrast, when the evaluator has preference-based partiality, a more precise signal of quality does not mitigate the animus towards female workers. Even if quality is perfectly observable, the female workers will still face discrimination.

This analysis immediately extends to a setting where evaluators have heterogenous beliefs or preferences. Aggregate discrimination is equal to $D(h_1, s_1) = (\frac{\tau_q}{\tau_q+\tau_\eta})E_\pi[\hat{\mu}_M - \hat{\mu}_F] + E_\pi[c_F]$, and an analogue to Proposition 1 immediately follows.

2.4 Dynamics of Discrimination

We now focus our attention on belief-based partiality and study how discrimination evolves across a sequence of tasks. We show that a discrimination reversal between the initial period and a subsequent period can distinguish between belief-based partiality with a correct versus misspecified model of inference. Throughout this section, assume there is no preference-based partiality, $c_F = c_M = 0$ for all evaluators.

Beginning in the second period, evaluations from prior rounds provide information about the worker's ability. A prior evaluation reflects both the prior signal of quality and the prior evaluator's belief about the worker's ability. Therefore, interpreting prior evaluations requires a model of other evaluators' beliefs. We focus on two cases. In the first, evaluators share a common belief about the distribution of ability by gender, and this is common knowledge. In the second, evaluators have heterogeneous beliefs – some evaluators have belief-based partiality towards men and believe that all evaluators share the same beliefs, while other evaluators have no belief-based partiality but are aware that some evaluators do. Since there is only one correct distribution of ability for each gender, this heterogeneity implies that at least one type of evaluator has a misspecified model of inference.

We show that these two cases make different dynamic predictions about the pattern of discrimination. Specifically, we show that a discrimination reversal does not arise in the first case – which nests the correctly specified model – but is possible in the second. Therefore, observing a discrimination reversal suggests that some evaluators have misspecified models of inference.

Impossibility of Reversal in Correctly-Specified Model. Suppose that all evaluators share a common prior belief about the distribution of ability by gender, have belief-based partiality, and this is common knowledge – that is, evaluators have a correct model of other evaluators. In our framework, this is captured by a single type θ with beliefs $\hat{\mu}_F < \hat{\mu}_M$ and a correctly specified type distribution, $\hat{\pi}(\theta) = 1$. This case nests the correctly specified model, in which evaluators also have correct beliefs about the distribution of ability by gender.

In the first period, a female worker is subjected to stricter standards than a male. Inverting (4), let

$$s_{g,1}(v_1) \equiv \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) v_1 - \left(\frac{\tau_q}{\tau_\eta}\right) \hat{\mu}_g \tag{6}$$

denote the signal required by gender g to receive evaluation v_1. In order to receive the same evaluation as the male worker, the female must produce a higher signal of quality to offset the lower belief about her ability, $s_{F,1}(v_1) > s_{M,1}(v_1)$. Therefore, a given evaluation is a more positive signal of a female worker's ability than a male's. This decreases the difference between the posterior beliefs about the female and male workers' average abilities, thereby reducing discrimination in the next period. However, despite the stronger signal from the female worker, the higher prior belief about the male worker's average ability still maps into a higher posterior belief, and the beliefs for the male and female worker do not reverse. Hence, discrimination does not reverse. The analysis in subsequent periods is analogous: evaluators' beliefs about the average ability of male and female workers continue to move closer together following similar evaluation histories, but do not reverse. This brings us to our first dynamic result.

Proposition 2 (Impossibility of Reversal)**.** *Suppose there is a single type of evaluator with belief-based partiality and no preference-based partiality. Then fixing an evaluation history, discrimination decreases across periods but never reverses.*

An immediate implication of Proposition 2 is the impossibility of a reversal in the correctly specified model. Therefore, observing a reversal is indicative of some form of misspecification – either in evaluators' beliefs about average ability by gender, evaluators' models of other evaluators, or both.

A key feature of social learning settings, such as our model, is the endogenous informational content of evaluations. In particular, the signal required to receive a given evaluation depends on the prior belief about average ability. As shown in (6), evaluation v_1 maps into a signal $s_{g,1}(v_1)$ that is decreasing in this prior mean. Therefore, the prior mean impacts the posterior mean through two channels that move in opposite directions: (i) the prior distribution of ability has a monotone increasing likelihood ratio in its mean; and (ii) the distribution of the signal that yields a given evaluation has a monotone decreasing likelihood ratio in the prior mean. The proof of Proposition 2 lies in establishing that the first effect dominates. Therefore, the posterior mean is increasing in the prior mean, so if evaluators believe that males have a higher prior mean than females, then following a given signal, they also believe that males have a higher posterior mean.[10]

Possibility of Reversal in Misspecified Model. Next, we present one form of misspecification that leads to a discrimination reversal. The key features of the set-up that drive the reversal are (i) the existence of a type with belief-based partiality (to generate initial discrimination) and (ii) the existence of evaluators who believe that some evaluators have more extreme belief-based partiality than they do (to generate the reversal). This is a possibility result, in the sense that it demonstrates one possible way to generate a reversal. Other forms of misspecification can also lead to reversals. Theoretically or empirically distinguishing between different forms of misspecification is beyond the scope of this paper.

Suppose there are two types of evaluators. The first type θ_1 uses a heuristic to form beliefs about the relationship between ability and gender, which leads to belief-based partiality that favors men, $\hat{\mu}_F^1 < \hat{\mu}_M^1$. This type is not aware of its bias, and believes that other evaluators have the same beliefs, $\hat{\pi}_1(\theta_1) = 1$, such as in the case of the bias blind spot (Pronin et al. 2002) or false consensus effect (Ross et al. 1977). With probability $p \in (0, 1)$, an evaluator is type θ_1. We refer to this type as the *heuristic* type. The second type θ_2 has no belief-based partiality, $\hat{\mu}_F^2 = \hat{\mu}_M^2$, but is aware that some evaluators do – it has a correctly specified model of the type distribution, $\hat{\pi}_2(\theta_1) = p$.

[10] If the informational content of evaluations were exogenous – for example, if evaluators simply reported the observed signal s – then the property that beliefs do not reverse would follow immediately. This is because the monotone likelihood ratio property (MLRP) is preserved under Bayesian updating with respect to a fixed signal distribution.

We refer to this type as the *impartial* type. To close the model, assume that both types have the same belief about the average ability of male workers, $\hat{\mu}_M^1 = \hat{\mu}_M^2$, and let $\hat{\mu}_M$ denote this belief. Importantly, this heterogeneity in the subjective belief about the average ability of female workers implies that at least one type has incorrect beliefs.[11]

In the first round, a heuristic evaluator discriminates against females, while an impartial evaluator exhibits no discrimination. Aggregate initial discrimination is a weighted average of these two type's evaluations,

$$D(h_1, s_1) = \left(\frac{\tau_q}{\tau_q + \tau_\eta}\right) p(\hat{\mu}_M - \hat{\mu}_F^1) > 0. \tag{7}$$

Following evaluation v_1, let $\hat{\mu}_F^i(v_1)$ denote an evaluator of type θ_i's posterior belief about the average ability of a female worker and $\hat{\mu}_M(v_1)$ denote the posterior belief about the average ability of a male worker (which is the same for both types). A heuristic evaluator's beliefs about ability evolve in the same manner as the beliefs of an evaluator in the single-type model, since the heuristic type believes that all evaluators have the same beliefs. From Proposition 2, the heuristic type's beliefs do not reverse, $\hat{\mu}_F^1(v_1) < \hat{\mu}_M(v_1)$, and therefore, the type continues to discriminate against females in the second period. In contrast, an impartial evaluator is aware that with positive probability, a female worker was evaluated by a heuristic type and faced discrimination in the first period. Therefore, the impartial type's posterior belief about average ability immediately favors females, $\hat{\mu}_F^2(v_1) > \hat{\mu}_M(v_1)$, and this type discriminates against males in the second period. As in the first period, aggregate discrimination in the second period is a weighted average of these two type's evaluations. Whether an aggregate discrimination reversal occurs depends on whether the impartial type's posterior belief favors females enough that it reverses the aggregate posterior belief about average quality. Proposition 3 establishes that indeed, given any initial evaluation or any second period signal, aggregate discrimination reversals are possible.

[11] The literature on heuristics and biases provides a foundation for such a model. Type θ_1 can capture evaluators who use a 'representativeness' heuristic to form beliefs about the population distribution of ability, i.e. steoreotyping (Bordalo et al. 2016b), and are not aware of their cognitive bias. Type θ_2 can capture evaluators who have accurate beliefs about the population distribution of ability by gender, and are aware that a subset of evaluators stereotype. In Appendix C, we use observational data to provide a foundation for type θ_1 in our experimental setting. We show that using the 'representativeness" heuristic will magnify small performance differences in the observational data, leading to belief-based partiality with bias.

Proposition 3 (Possibility of Reversal)**.** *Suppose evaluators are the heuristic type θ_1 with probability $p \in (0,1)$ and the impartial type θ_2 with probability $1-p$.*

1. *For any initial evaluation v_1, there exist cut-offs $\overline{p} \in (0,1)$ and $\overline{s} \in \mathbb{R}$ such that for a low enough share of heuristic types $p \in (0, \overline{p})$ and a high enough second period signal, $s_2 > \overline{s}$, aggregate discrimination reverses in the second period, $D(v_1, s_2) < 0$.*

2. *For any second period signal s_2, there exist cut-offs $\overline{p}' \in (0,1)$ and $\overline{v} \in \mathbb{R}$ such that for a low enough share of heuristic types $p \in (0, \overline{p}')$ and a low enough initial evaluation, $v_1 < \overline{v}$, aggregate discrimination reverses in the second period, $D(v_1, s_2) < 0$.*

Increasing the prevalence of heuristic evaluators impacts second period discrimination through two channels. First, it increases the difference in the impartial type's second period beliefs about the average ability of a male and a female. A larger share of heuristic evaluators means that it is more likely that the female faced initial discrimination. Therefore, it is more likely that she received the higher signal that would be required to receive a given evaluation from the heuristic type, rather than the lower signal that would be required to receive this evaluation from the impartial type. Second, it increases the probability that the second period evaluator is a heuristic type with belief-based partiality. Since the heuristic type still discriminates against females in the second period, it is more likely that a female will continue to face discrimination. The first effect dominates for low p, while the latter effect dominates for high p. This leads to a non-monotonicity in how second period discrimination changes with respect to p. Further, discrimination is always zero at $p = 0$, as no evaluators have belief-based partiality, and discrimination is always positive as p approaches one, as this set-up approaches the set-up with a single type of evaluator. Figure 1 illustrates this reversal.

Proposition 3 does not rely on the assumption that the impartial evaluators exactly understand the bias of the heuristic evaluators or their prevalence in the population. It is straightforward to derive a similar result when the impartial evaluators under- or overestimate either the bias of heuristic evaluators or their frequency in the population.

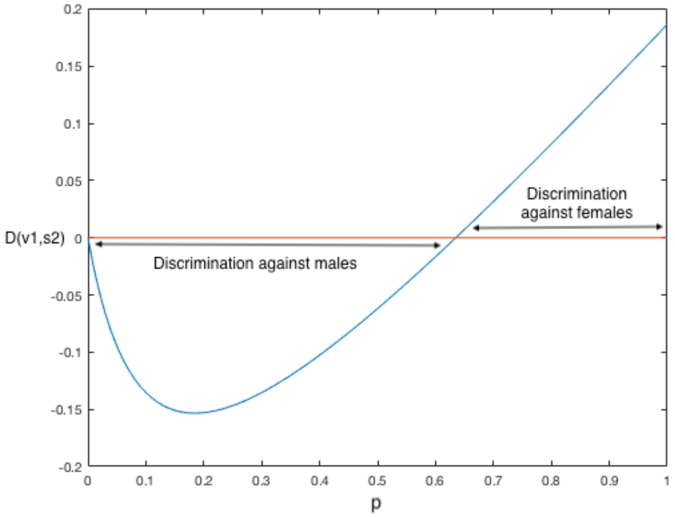


Figure 1. Second period discrimination, as a function of the proportion of heuristic evaluators.

2.5 Discussion of Results

In summary, our theoretical results show that (i) it is not possible to identify the source of discrimination from a single round of evaluations with a fixed level of subjectivity; (ii) varying the subjectivity of judgment can identify whether the source of discrimination is preference-based or belief-based; (iii) a reversal of discrimination is not possible in a correctly-specified model of belief-based partiality; and (iv) a reversal of discrimination points to belief-based partiality with misspecification. Before moving to the empirical section, we briefly discuss the robustness of our theoretical framework to other specifications and relate it to several alternative models.

2.5.1 Robustness

Alternative Distributions of Ability. We combine a partial analytical derivation with numerical analysis to illustrate that the impossibility of a reversal in the correctly specified model (Proposition 2) is robust to other distributions of ability, including the beta distribution, exponential distribution, gamma distribution and a setting with binary ability and quality.[12] We present this analysis in Appendix A.2. The key feature

[12] Analytical results are possible for the normal distribution, as a normal ability distribution is the conjugate prior for a normal signal distribution. This means that the posterior distribution of ability is also normally distributed, which allows for a recursive representation of the belief-updating process and a closed-form characterization of the evolution of beliefs. A combination of analytical and numerical analysis is necessary once we lose the conjugate prior property, as is the case for the

that drives the impossibility of reversal is that the ability distribution satisfies the monotone likelihood ratio property (MLRP) with respect to the parameter that varies by gender. The MLRP is commonly assumed in information economics (Milgrom 1981) and holds for many other families of distributions. Propositions 1 and 3 are also robust to alternative distributional assumptions. It is straightforward to extend Proposition 1 analytically. By similar intuition to Proposition 3, it is possible to generate reversals in misspecified models with other ability distributions.

Coarse Evaluations. We assume that the space of possible evaluations is isomorphic to the space of beliefs about expected quality. In reality, the space of possible evaluations may be coarser than the evaluator's belief about expected quality, and it may not be possible to perfectly infer the signal she observed from the reported evaluation. For example, the evaluator may only be able to accept or reject a task, or rate it on a scale of one to five. When this is the case, information will be lost, in the sense that each observed evaluation will map back into an interval of possible signals. In Appendix A.3.1, we show that an analogue of Proposition 2 holds for coarse evaluations. In particular, a discrimination reversal does not occur between the first and second period when evaluators have common knowledge of the same beliefs about ability for men and women.

Shifting Standards. Another relevant feature for our setting is how the standard of evaluation may change with respect to reputation. Higher reputation often leads to increased responsibilities and privileges, which require greater ability to manage effectively. As such, individuals may be subject to increasingly higher benchmarks as their level of seniority increases to avoid erroneously granting responsibility to someone who is unprepared. Our framework can easily be adapted to capture shifting standards (Biernat, Vescio, and Manis 1998) with respect to reputation. We say a worker faces *shifting standards* if, conditional on receiving a positive initial evaluation, the worker faces a stricter standard in the second period – a higher signal is required to receive any evaluation, relative to the signal required for the same evaluation in the first period. We explore this extension in Appendix A.3.2.

other distributions we consider.

2.5.2 Alternative Models

Attrition. Suppose that workers exit the worker pool with positive probability after completing each task, and lower ability workers exit at a higher rate than higher ability workers. In this case, the content and the *length* of the worker's evaluation history provide information about ability. If male workers exit at a lower rate than female workers, conditional on sharing similar evaluation histories, then the length of the evaluation history has different informational content for male and female workers. If evaluators' subjective prior beliefs about ability favor males, then this differential attrition will shrink these initial differences. It can even lead to a reversal when low ability women exit at a fast enough rate that distributions following longer histories favor women. In contrast to Proposition 3, such a reversal can occur even when all evaluators have correctly specified models.

In Section 3.4 we demonstrate how one can empirically test for differential attrition by gender. We use observational data from the forum to show that males and females with similar evaluation histories leave the market at similar rates. Therefore, there is no evidence for differential attrition in our experimental setting. Evaluators may incorrectly believe that attrition differs by gender – but in this case, they have a misspecified model. Differential attrition may drive discrimination reversals in other labor market settings. It is therefore important to empirically measure attrition in order to rule out differential attrition as a potential driver of the reversal. Our empirical analysis demonstrates how one could measure attrition based on observable data from the setting of interest, and test whether it differs by gender.

Gender Differences in Variance of Ability. If the variance of ability differs for females and males, then discrimination may be non-monotonic in the signal of quality. It is straightforward to show that this can lead to a discrimination reversal when the signal of quality is imprecise. Regardless of the precision of the signal, this model also predicts higher variance in the evaluations of the group with higher variance in ability, relative to the group with lower variance in ability. In Section 3.4 we demonstrate how one can empirically test for differences in the variance of evaluations by gender. We use observational data from the forum to show that males and females have similar variances in evaluations for tasks with precise signals. Therefore, in our experimental setting, we find no evidence for gender differences in the variance of ability. Differential variance may drive discrimination reversals in other labor market settings. Our empirical analysis demonstrates how one could use variance in evaluations – which is based on observable data – to proxy for variance in ability – which is unobservable – and test whether it differs by gender.

Heterogeneous Preference-Based Partiality. Suppose that all evaluators have correct beliefs about the ability distributions for male and female workers, but vary in their preference-based partiality against females. We will show that there is no distribution over types that can simultaneously capture the following two predictions: (i) initial discrimination against females when judgement is subjective, $\tau_\eta > 0$, and (ii) no initial discrimination against females when judgement is objective, $\tau_\eta \approx 0$. Consider a type space Θ where each type $\theta_i \in \Theta$ has preference parameter c_F^i and correct beliefs, $\hat{\mu}_F^i = \mu_F$ and $\hat{\mu}_M^i = \mu_M$. From (5), when there are multiple types, initial discrimination is equal to $D(h_1, s_1) = E_\pi[c_F^i]$. Prediction (i) requires $E_\pi[c_F^i] > 0$. But $D(h_1, s_1)$ is independent of τ_η. Therefore, when $E_\pi[c_F^i] > 0$, there will also be discrimination when judgement is objective and prediction (ii) is not possible. Therefore, observing evidence for predictions (i) and (ii) rules out discrimination that is caused by preference-based partiality with heterogeneous preference parameters. In Section 3.3, we demonstrate how to empirically test these two predictions in our experimental setting.

Self-Fulfilling Beliefs. As discussed in Section 2.2, self-fulfilling beliefs are another form of belief-based discrimination. Fryer (2007) explores how discrimination dynamically evolves when it is driven by self-fulfilling beliefs. He shows that discrimination can reverse in the second period if there exist equilibria in which one group coordinates on an equilibrium with higher initial standards and looser second period standards, while the other group coordinates on looser initial standards and more stringent second period standards.[13] Thus, in Fryer (2007), the reversal depends on how coordination dynamically evolves, while in our model, the reversal stems from the endogenous informational content of prior evaluations. In Fryer's setting, multiple equilibria always exist – there are also equilibria in which either group faces discrimination in both periods and equilibria in which all workers are treated equally. Therefore, almost all

[13] The existence of an equilibrium in which beliefs flip requires fairly strict conditions. In relation to our setting, the payoff to an evaluator for accurately evaluating a product must be substantially higher than the payoff to the worker for receiving a positive evaluation. This assumption is likely not satisfied in many settings of interest, including the experimental setting we consider in Section 3 and settings with competition.

outcomes are possible, conditional on observables.

3 A Field Experiment

We conduct a field experiment on an online Q&A mathematics forum, Mathematics Stack Exchange. Mathematics Stack Exchange is part of a family of forums where, in 2017 alone, 3,517,799 questions were asked and 4,299,077 answers were provided. With over 10 million registered users, the forums are an important resource for students and researchers in STEM. We examine gender discrimination by posting content to the forum in the form of questions and answers.[14] In addition to the experiment, we exploit two additional data sources to explore the predictions of the theoretical framework. First, we collect observational data from the forum to further study potential mechanisms, including estimating distributions from publicly available statistics. Second, we use a private dataset provided by the forum on the voting behavior of users to run additional robustness tests.

3.1 Description of Forum

Organizing terms with respect to the theoretical framework, users (workers) generate content in the form of posts (tasks), the quality of which are then assessed by other users on the forum (evaluators). There are two main types of tasks – questions and answers (in response to other users' questions). See Appendix B.1 for examples of both types of posts. Users can choose to evaluate either type of post by assigning an upvote or downvote to it. Voting is anonymous – other users cannot observe any information about the identity of the user who cast a vote.[15]

The forum offers written guidelines for evaluating posts, and these guidelines are actively discussed on the forum's message boards. Voting is meant to serve a dual purpose: (i) upvoting is meant to highlight a quality post while downvoting is meant to discourage low quality posts, and (ii) upvoting rewards the *user* for a high quality post while downvoting punishes him or her for a low quality post. The second point stems from the fact that users earn publicly observable reputation points from the votes they receive for their posts. An upvote earns five reputation points on questions

[14] The experiment was pre-registered in the AEA RCT Registry, AEARCTR-0000950

[15] The anonymous setting ensured that the decisions of users interacting with our posts were not subject to experimenter demand effects.

and ten reputation points on answers, while a downvote deducts two reputation points for both questions and answers.[16] Reputation unlocks privileges, such as the ability to edit and comment on others' posts or tag questions as duplicates. It can also be used as a currency through the assignment of "bounties" – users can *spend* their reputation points to post a question with a bounty that will be awarded to the highest quality answer, as determined by the question poster, to increase the quality of answers.

The theoretical set-up in Section 2 maps onto the key features of the experimental environment. Each post on the forum is accompanied by clearly visible information summarizing its evaluation by the community – the associated net number of votes (upvotes minus downvotes) – and information about the poster – his or her username and current reputation. In judging the quality of a post, the evaluator can read the content of the post (a signal), as well as draw inference from the gender of the username (population beliefs) and the reputation (evaluation history). The number of reputation points serves as a summary statistic of past quality – greater reputation corresponds to the evaluators observing a higher sequence of signals on prior posts – while clicking on the user's profile reveals the full history of upvotes and downvotes by post. The informational content of reputation and prior evaluations endogenously depends on the voting behavior of other users on the forum. Therefore, interpreting these evaluations requires a model of how past voting behavior depends on the prior evaluators' beliefs and preferences. For example, an evaluator who is aware that female users face more exacting initial standards may take this into account when assessing a question from a high-reputation female.

Additionally, higher reputations earn users greater privileges on the forum. Reputation allows users to advance through the ranks, with each rank corresponding to a new set of privileges. This includes privileges to "supervise" other users – for example, to edit, comment on, flag, downvote and close other users' posts. In turn, the evaluation process mirrors promotion decisions in labor market contexts: the higher a user's reputation, the more influence he or she has over other users on the forum.

3.2 Experimental Design

The ability to exogenously vary the gender and reputation associated with a user makes this an ideal setting for testing the dynamic predictions of different sources of

[16] It is not possible for a user's reputation to fall below one.

discrimination. Comparing evaluations of question and answer posts allow us to test the predictions of how discrimination varies with the level of subjectivity in judgment.

Posting Questions. We generated a series of original mathematics questions and posted them under male and female usernames on accounts with low and high reputations. We opened 280 new accounts, with 140 male usernames and 140 female usernames.[17] Each account was associated with its own email address, username and password. Of these accounts, 140 (70 with female usernames and 70 with male usernames) were left as new accounts; these comprised the Novice accounts. For the other 140 accounts (70 male and 70 female), we manually built-up the reputation to the top 25^{th} percentile of reputation on the forum – at the time of the experiment, this corresponded to a reputation of at least 100. Research assistants earned reputation on each account by posting content until the accumulated reputation reached 100. Once an account reached at least 100, the research assistant stopped posting content. Because reputation was accumulated through the actions (votes) of other users on the forum, we could not control the exact number of reputation points associated with each account – the mean reputation on these accounts was $M = 155.23$. These accounts comprised the Advanced accounts.

Critically, upon achieving a high reputation, we re-randomized the gender of the username on the Advanced accounts: 35 accounts that were built-up under male usernames were switched to female, and 35 female accounts were switched to male; the remaining 70 accounts received a new username of the same gender. Importantly, when a username is switched, all past and future activity on the account became associated with the new username. That is, all *previous* posts now reflect the new username, and no public record of the name change is available. Re-randomizing the gender of the usernames avoids issues of endogeneity associated with, for example, female accounts requiring different quality posts to achieve the same level of reputation as male accounts. After reassigning usernames, the new female and male accounts had similar reputation levels ($M = 155.89$ vs. $M = 154.57$, respectively, $p = .82$).

Our goal was to write high quality questions that would be well-received on the forum. Content on the forum ranges from high school arithmetic to upper-level graduate mathematics. Questions are tagged by topic, e.g. real analysis, combinatorics. Users

[17] Names were taken from the “Top names of the 2000s” list created by the Social Security Administration, https://www.ssa.gov/oact/babynames/decades/names2000s.html.

are discouraged from posting questions directly from textbooks or duplicating content that is already posted; such posts are flagged and routinely closed by moderators. In order to minimize the chance that our content was flagged, we wrote 280 novel mathematics questions ranging in level of difficulty from upper-level undergraduate to early graduate. These questions were randomly assigned to one of the four conditions: male novice, female novice, male advanced or female advanced.

We posted questions on a pre-determined schedule to avoid altering the usual activity on the forum, i.e. flooding the forum with content. Research assistants posted one question at least twenty minutes apart between 5-10PM, Monday through Thursday. Data on the community response to the questions, e.g. upvotes, downvotes, number of answers, was collected seven days after posting for each question, both in numerical form and as screenshots. A total of 7 of the 280 questions were dropped from our analysis due to forum moderators prematurely closing the questions before the end of the seven day window or due to errors in the posting of the questions (i.e. two questions posted to the same account).

We measure discrimination as either the average change in reputation points per post ($\varDelta$Rep) or the average number of upvotes net of downvotes per post (Net Votes). The dynamic pattern of discrimination provides a test of the theoretical predictions outlined in Section 2. Conditional on observing discrimination between male and female Novice accounts, a mitigation in its intensity for Advanced accounts is consistent with belief-based partiality, including the case of statistical discrimination where beliefs are correct, while a *reversal* of discrimination for Advanced accounts is evidence for biased belief-based partiality.

We do not make a prediction on how evaluations vary by reputation within a given gender or pooled across genders, due to the potential for shifting standards (Section 2.5). Higher reputation is indicative of higher ability, which leads to a higher assessment of quality on a given post. But as previously discussed, reputation serves both the purpose of highlighting a quality post and rewarding the poster. Therefore, posts by high reputation users may be held to higher standards of quality, since reputation determines which users rise through the rungs to become moderators and receive other privileges. For example, a novice user may be rewarded with an upvote for a low-level calculus question, but an advanced user may not be. In our experiment, randomization ensures that the average quality of questions posted to novice accounts is approximately the same as that of questions posted to advanced accounts. Since the two effects point in opposite directions, the overall directional prediction regarding the effect of reputation on upvotes per question is ambiguous.

Posting Answers. We generated original answers to mathematics questions posted by other users on the forum, and posted them under male and female usernames. To examine how the subjectivity of judgment affects discrimination, we compared the evaluations of these answers to the evaluations of questions. The guidelines for determining whether a post merits an upvote or downvote are different for questions and answers. The standard of quality for answers is clear: determine whether or not the answer is correct. In contrast, there are multiple standards for judging the quality of a question, including whether it is interesting, novel, or important for the accumulation of knowledge on the forum. According to our definition of subjectivity outlined in Section 2, this difference in standards of quality should make judgment of questions more subjective than judgment of answers.

The difference in subjectivity is echoed in the meta-forums for the site. A popular post asks why the site's users upvote questions. The poster writes that for answers: "it's easy to determine what to upvote. Is it correct?" For questions, this objective criteria does not apply. What criteria do others use? This post has dozens of responses, including: is the question well-written, non-trivial or insightful, am I curious about the same question, has the poster made me curious about what they are asking, do I think it is important and should be visible to others, does it show research effort, the *combination of topic with the reputation of the poster*. One response highlights potential issues with the subjectivity in judgment for questions, noting that voting on questions may be affected by disliking the topic in general or viewing it as unimportant (this response had one of the highest number of upvotes on the forum.)

To post answers, we created a second set of 140 Novice accounts with no prior posts, split between 70 male usernames and 70 female usernames. We needed to post answers in real time, as questions on the forum are answered fairly quickly and late answers generally receive little attention. To do so, research assistants worked in pairs. One member of the pair, the 'answerer', would find a newly posted question that had not been answered yet and write an answer for it. The 'answerer' would then send the answer and a link to the question to the other research assistant, the 'poster', who would assign the answer to one of our accounts and post it. The order of accounts that the answer would be posted to was pre-determined – known to the 'poster' but not the 'answerer'. As such, the research assistant writing the answer did not know the gender of the account that the answer would be posted to, and therefore, could not be subconsciously influenced by whether the answer would be posted to a male or female account. As with the questions, answers were posted between 5-10PM, Monday through Thursday. Data was collected seven days after posting the answer, both in numerical form and as a screenshot. A total of 5 of the 140 answers were dropped due to errors, e.g. the question was closed before the seven day window concluded.

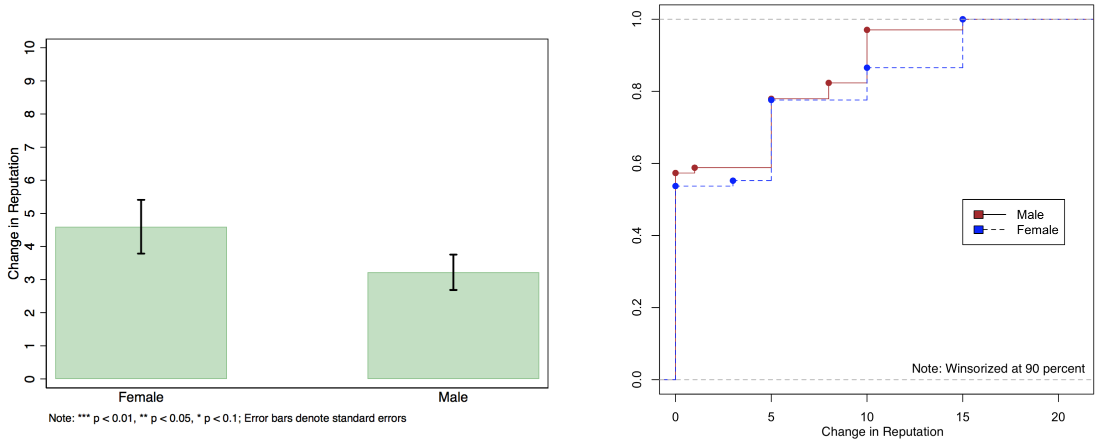
The theory in Section 2 predicts that subjectivity in judgment, modeled as the precision of the signal of quality, will affect discrimination differentially depending on its source. Conditional on observing discrimination on questions, which involve more subjectivity in judgment, a mitigation of discrimination on answers is indicative of belief-based partiality. In contrast, a similar level of discrimination for both questions and answers suggests preference-based partiality.

Site Activity. We continuously scraped the forum for activity to capture relevant metrics for the experiment and ensure that activity on the forum remained relatively similar for the duration of the experiment. The turnover in unique active users was high: the average daily turnover was 85% and the weekly turnover was 92%.

3.3 Experimental Results

We first present results comparing the evaluations of answers versus questions by gender. Examining how subjectivity of judgment affects discrimination in our setting enables us to distinguish between preference and belief-based partiality. We then present results comparing the evaluations of novice versus advanced questions by gender. This allows us to study the dynamics of discrimination and helps to distinguish between biased and unbiased belief-based partiality.

Subjectivity of Judgment. We first examine the change in reputation ($\mathit{\Delta}$Rep) for answers posted to male versus female accounts (i.e. the reputation points earned on the post). Table 1, Column (1) shows that regressing $\mathit{\Delta}$Rep per answer on gender reveals no significant difference in the evaluation of answers at conventional levels. This result is illustrated in Figure 2(a), which shows the average $\mathit{\Delta}$Rep by gender, and 2(b), which plots the distributions of $\mathit{\Delta}$Rep by gender. Table 1, Column (2) repeats the analysis



(a) Average ΔRep

(b) Distribution of ΔRep (*cdf*)

Figure 2. ΔReputation for Answers

using net votes per post as the dependent variable.[18] Together, these results suggest that there is little evidence for gender discrimination on answers.

Looking at the evaluation of questions posted to novice accounts reveals a substantially different pattern. We find significant initial discrimination against females: regressing ΔRep or net votes per question on the gender of the poster reveals that questions posted to accounts with female usernames accumulated significantly fewer reputation points (Table 1, Column (3)) and received significantly fewer net votes (Table 1, Column (4)) than questions posted to accounts with male usernames. These differences correspond to roughly 0.4 standard deviations of the average change in reputation and average number of votes. This result is illustrated in Figure 3(a), which shows the average ΔRep by gender, and Figure 4(a), which plots the distributions of ΔRep by gender. Together, these results suggest that there is significant evidence for gender discrimination on questions.

Next, we directly compare responses to answer versus question posts by gender. We first test the difference in the estimated coefficients of the male gender dummy between the question and answer regressions and find that this difference is significant for both ΔRep ($\chi^2(1) = 6.34$; $p = .01$) and net votes ($\chi^2(1) = 7.87$; $p = .005$). We then present regression results for question and answer posts within the same model. We regress ΔRep and net votes on dummies corresponding to gender, type of post (question or

[18] Downvotes were very rare in our sample. We obtain similar results when we use only upvotes as the dependent variable (Appendix B.2).

Table 1. Subjectivity: Effect of Gender on Evaluation of Novice Answers and Questions

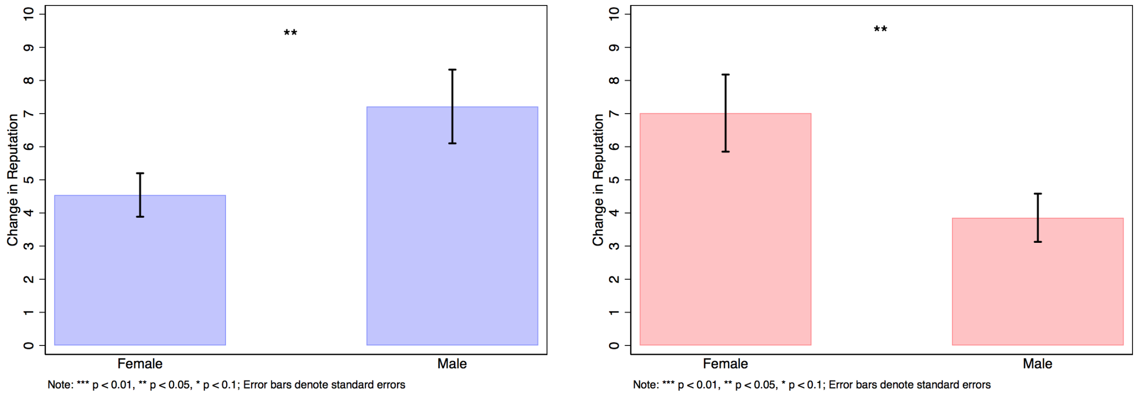
	Answers Only		Questions Only		Answers & Questions	
	ΔRep	Net Votes	ΔRep	Net Votes	ΔRep	Net Votes
	(1)	(2)	(3)	(4)	(5)	(6)
Male	−1.38	−0.31	2.86	0.58	−1.38	−0.31
	(.97)	(.17)	(1.32)	(.27)	(1.16)	(.22)
Question					0.08	0.09
					(1.16)	(.22)
Male*Question					4.24	0.89
					(1.64)	(.32)
Constant	4.60	0.79	4.68	0.88	4.60	0.79
	(.69)	(.12)	(.93)	(.19)	(.96)	(0.16)
# 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.

answer) and the interaction of gender and type of post (Table 1, Columns (5) and (6)). There is a significant mitigation of discrimination against female accounts for answers, relative to questions: the interaction effect between gender and type of post is positive and significant in both specifications. This implies that the male advantage is significantly larger for questions, compared to answers.

Taken together, these results are inconsistent with discrimination due to preference-based partiality. Rather, they support the theoretical prediction on how subjectivity affects discrimination when evaluators have *belief-based* partiality.

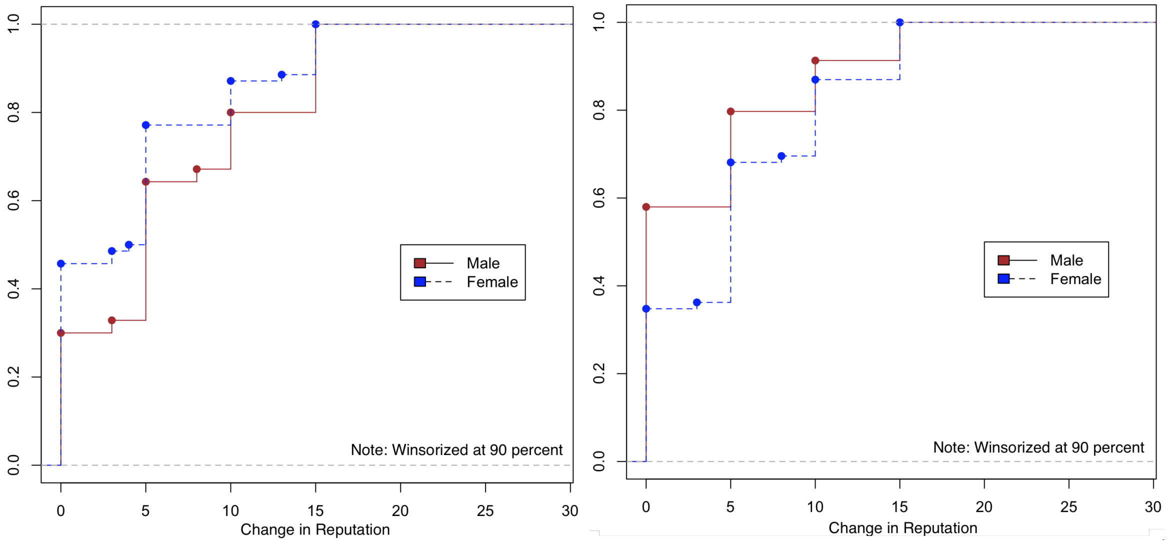
Dynamics of Discrimination. Next, we examine the dynamics of discrimination by comparing discrimination towards novice and advanced users. As shown in Figure 3(b), questions posted to advanced female accounts accumulated more reputation points, ΔRep, than those posted to advanced male accounts. This contrasts with questions posted on novice female accounts, which accumulated fewer reputation points than those posted to novice male accounts (Figure 3(a)). In other words, we observe a dynamic *reversal* of discrimination between novice and advanced accounts – questions from male users are favored at low reputations, while questions from female users are favored at high reputations. Figure 4 illustrates this reversal in the distributions of ΔRep: panel (a) shows the distribution of ΔRep on questions posted to novice



(a) Novice Accounts

(b) Advanced Accounts

Figure 3. Average *Δ*Reputation for Questions



(a) Novice Accounts

(b) Advanced Accounts

Figure 4. Distribution of *Δ*Reputation for Questions (*cdf*)

accounts, while panel (b) shows the distribution of ΔRep for advanced accounts.

For advanced accounts, regressing *Δ*Rep or net votes per question on the gender of the poster reveals that questions posted to female accounts accumulated significantly more reputation points and net votes than questions posted to male accounts (Table 2, Columns 1 and 2, respectively). These differences in evaluation correspond to roughly 0.6 standard deviations for both *Δ*Rep and net votes. This contrasts with the significantly lower evaluation of questions posted to novice female accounts relative to

Table 2. Dynamics: Effect of Gender on Evaluation Questions, Novice and Advanced

	Advanced		Novice & Advanced		
	*Δ*Rep	Net Votes	*Δ*Rep	Net Votes	Binary
	(1)	(2)	(3)	(4)	(5)
Male	−3.16 (1.37)	−0.62 (.28)	2.86 (1.36)	0.58 (.27)	0.17 (.08)
Advanced			2.33 (1.35)	0.49 (.27)	0.09 (0.08)
Male*Advanced			−6.02 (1.91)	−1.20 (.38)	−0.40 (.11)
Constant	7.01 (0.97)	1.38 (.20)	4.68 (.96)	0.88 (0.19)	0.56 (.56)
# 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.

novice male accounts, as reported in Table 1. Testing the difference in the estimated coefficients of the male gender dummy between the Novice and Advanced regressions reveals a significant difference for both *Δ*Rep ($\chi^2(1) = 9.67$; $p = .002$) and net votes ($\chi^2(1) = 9.88$; $p = .002$).

Table 2 Columns (3) and (4) present regression results for Novice and Advanced accounts within the same model. In Column (3), we regress *Δ*Rep on dummies corresponding to the gender of the poster, the reputation level of the poster (novice or advanced), and their interaction. The interaction between gender and reputation level is negative and significant, confirming the reversal of discrimination between the Novice and Advanced accounts. The same pattern of results holds for the net votes earned per question (Column 4). To ensure that these results are not driven by outliers or subsequent voters herding on the first upvote, we replicate the analysis using a binary variable that is equal to one if the question receives at least one upvote, and zero otherwise. As shown in Column (5), the results are robust to this binary specification.[19] Consistent with shifting standards, the average change in reputation and average number of net votes, pooled across both genders, does not significantly differ between Novice and Advanced accounts.

[19] Results are also robust to winsorizing the dependent variable at 5% or 10%.

In summary, we find that in our setting, not only is initial discrimination against females mitigated by reputation, but the direction of discrimination *reverses* – females are *favored* at higher reputations. Interpreting these findings through the lens of the theoretical framework, our results suggest that initial discrimination is driven by belief-based partiality with bias.

Robustness Checks. The forum provided us with a proprietary dataset that contains additional information about the evaluators in our experiment. The dataset uniquely identifies the users who evaluated our content (i.e. voted on question and answer posts in our experiment), and provides their historical activity on the forum. This data allows us to conduct further robustness checks and to explore the typical voting behavior of evaluators who interacted with our posts to determine whether the population of users who evaluated our posts is similar across groups.

We first use this data to test whether our results are robust to excluding repeat votes from evaluators who interacted with our posts more than once. We restricted the voting data to the first vote from each evaluator on a post in our experiment, and re-ran the analyses from Tables 1 and 2. Our findings are robust to excluding these repeat votes. The results are presented in Appendix B.2.

We also explored whether the users who evaluated questions in our experiment are similar to the users who evaluated answers. To determine whether users specialize in the type of content they evaluate by either evaluating mostly questions or mostly answers, or whether most users evaluate both, we tabulated each user's total number of votes by content type, and calculated the proportion of a given user's votes that were cast on questions versus answers. The proportions are very similar: on average, 48% of a user's votes were cast on questions and 52% were cast on answers, with a standard deviation of .21. This suggests that most users evaluated questions and answers in fairly equal proportions. We also examined whether the users who evaluated our content differed in their reputation levels and inferred genders, depending on the type of post.[20] Summary statistics are presented in Appendix B.3; we found no significant differences in the characteristics of voters evaluating different types of posts.

[20] Section 3.4 outlines the process of inferring gender from a username.

3.4 Observational Data

Next, we analyze an observational dataset from the forum. We estimate relevant population statistics and use these estimates to evaluate alternative potential explanations for the documented discrimination reversal, including differential attrition by gender, gender differences in the variance of the ability distributions and autocorrelation in the quality of posts. We also explore gender differences in evaluations for all users who have a reputation within the range of our experiment and compare these differences to those found in our experiment.

Description of Data. The observational dataset is compiled and made publicly available by Mathematics Stack Exchange. It contains information on the attributes (e.g. reputations, usernames, location) and posting behavior (e.g. number of question and answer posts) of 315,792 users from July 2010 to March 2017. We excluded all content posted as part of our experiment. To code gender, we ran an algorithm developed by Vasilescu, Capiluppi, and Serebrenik (2014) to classify the gender of the usernames (see Appendix B.4.1 for a description of this algorithm). Each username is classified as 'male,' 'female,' or 'x' (when gender cannot be inferred). In our sample, the gender was resolved for 55% of accounts, which we used in the analyses. Of these accounts, 19% were classified as 'female' and the remaining 81% were classified as 'male.' Of accounts that had less than 100 reputation points, 21% were classified as 'female'; of accounts that had between 100 and 240 reputation points – the Advanced range used in our experiment – 13% were classified as female.

Attrition. We studied posting behavior of users to determine whether there is differential attrition based on gender. To do so, we created a panel data set of up to the first ten posts for each user, including whether each post was a question or an answer and the amount of reputation earned on the post.[21] For each user, we observed whether their first post was followed by a second post, whether their second post was followed by a third post, and so on. Differential attrition will lead to gender differences in the likelihood of observing a subsequent post conditional on receiving similar evaluations on the prior post.

[21] The mean number of total posts per user for users in our relevant reputation range (i.e. a reputation of up to 250 at time of posting) is 4.29, with a standard devision of 4.66. Therefore, we restricted attention to a user's first 10 posts.

We first examined whether there was differential attrition by gender following the first post. We ran 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, both pooling question and answer posts and analyzing each separately. We also split the reputation earned on the first post into quartiles and ran a similar regression. Neither the gender variable nor the interaction with reputation or reputation quartile is significant in any of these specifications. This suggests that female users were no less likely than male users to generate a second post conditional on a similar first post (see Tables 8 and 9 in Appendix B.4.2). We repeated a similar analysis to study differential attrition by gender following the second through ninth posts (see Table 10 in Appendix B.4.2). These results mirror the results following the first post – neither the gender variable nor the interaction with reputation is significant in subsequent periods.

Next, we ran a pooled analysis on all posts in our panel. In order to compare attrition rates for males and females with similar evaluation histories, we created a variable corresponding to the total reputation earned on all previous posts. For example, when looking at the likelihood of a 4^{th} post, total reputation earned is the sum of the reputation earned on the 3^{rd}, 2^{nd} and 1^{st} posts. We ran a probit regression, regressing a dummy for whether a user generated a post t on the gender dummy, the log of the total reputation earned on posts one through t-1, their interaction, and a dummy of whether the previous post was a question. In two of the three specifications, we also controlled for how many posts it took to generate this total reputation. Neither the gender variable nor the interaction with total reputation is significant in any of these specifications (see Table 11 in Appendix B.4.2).

Taken together, these results suggest that attrition is similar for males and females at the post histories that are relevant for our experiment.

Variance. The observational data also allows us to examine whether there are differences in variances of the ability distributions by gender. Since we did not find evidence for discrimination on answers, we use the evaluations of answers posted to new accounts to proxy for underlying ability. We then examine whether there are differences in the variance of these evaluations by gender. Running Levene's test of equal variances on the distributions of reputation points per first answer post (ΔRep) reveals no significant differences by gender ($p = .41$ using the mean, $p = .48$ using the median, $p = .46$

using the 10% trimmed mean).

Autocorrelation. As we outline theoretically in Section B.4.3, negative autocorrelation in the error process for quality could potentially lead to a discrimination reversal in a correctly specified model. We studied the dynamic pattern of evaluations to determine whether such negative autocorrelation is present empirically. We compiled a panel dataset consisting of all answer posts by users who had a reputation between 1 and 250 at the time of posting, which is the relevant reputation range for our experiment. We used the Wooldridge test for serial correlation in panel data (Wooldridge 2010). We first ran a random effects regression, regressing the reputation earned on an answer post on a gender dummy, then tested the estimated residuals for autocorrelation. We did not find evidence for significant autocorrelation. We used a similar method for question posts and also find no evidence of significant negative autocorrelation. See Appendix B.4.3 for a more detailed description of the analysis.

Gender Differences in Evaluations. We also use the observational data to examine how evaluations of posts vary with reputation, inferred gender of the user and type of post. 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 (see Appendix B.4.4 for details).

This analysis comes with several important caveats. First, there is the obvious endogeneity problem that stems from not being able to control for the quality of question posts. Second, there may be gender-based selection between the novice and advanced accounts. Although the above analysis suggests there is little evidence of differential attrition conditional on receiving similar evaluations on prior posts, male and female users may still face different evaluation thresholds early on. In fact, our experimental results show this to be likely. Finally, 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 attempt to address these caveats by running different specifications of the regression model, e.g. controlling for number of posts required to attain advanced status.

Keeping these caveats in mind, we run regression analogous to Tables 1 and 2 using the reputation points earned per post ($\varDelta$Rep) as the dependent variable. These results are presented in Tables 12-14 in Appendix B.4.4. The evaluation patterns by gender across the different types of posts are similar to those documented in the experiment, although the effect sizes vary depending on the specification. We document three main findings: (i) no significant evidence of gender discrimination on answers, (ii) questions posted by novice accounts with female usernames tend to earn fewer reputation points than those posted by novice accounts with male usernames, and (iii) questions posted to advanced accounts with female usernames tend to earn *more* reputation points than those posted to advanced accounts with male usernames.

Stereotypes. Finally, we use the observational data to explore how the 'representativeness' heuristic can lead to biased stereotypes in our setting. We examine the distribution of users' evaluations per answer post, and show how even mild belief distortions due to 'representativeness' significantly magnify small underlying performance differences between males and females. See Appendix C for details.

4 Discussion and Conclusion

In this paper, we propose a method for identifying the source of discrimination based on (i) how it evolves dynamically, and (ii) how it responds to the degree of subjectivity in judgment. We develop a theoretical model in which evaluators learn about a worker's ability through other evaluators' assessments of previous tasks. We show that the observable patterns of discrimination along these two dimensions depend critically on the underlying source – which we term *partiality.* The theoretical analysis yields an impossibility result: discrimination does not dynamically reverse if it is driven by correctly-specified belief-based partiality. In contrast, we show that a reversal can occur if some evaluators hold biased stereotypes, while others are aware of the bias and account for it when learning from prior evaluations. We also show that discrimination driven by preference-based partiality remains constant with respect to the level of subjectivity in judgment, while discrimination driven by belief-based partiality decreases as judgment criteria becomes more objective.

We present results from a field experiment exploring discrimination along these two dimensions. We post questions and answers on an online forum (Mathematics Stack Exchange) to accounts we created that exogenously vary in the gender of the usernames and the reputation on the forum. We document three main results: (i) significant gender discrimination *exists* at the initial stage, in the form of less reputation earned per post and fewer votes per post on questions posted by low reputation female accounts relative to questions posted by low reputation male accounts; (ii) significantly *less* gender discrimination at the initial stage for answers, where judgment of quality is less subjective relative to questions; and (iii) discrimination *reverses* for questions at more advanced stages, in that more reputation is earned and more votes are received on questions posted to high reputation female accounts relative to high reputation male accounts. We complement the experimental results with an analysis of observational data from the forum. We use an algorithm to infer gender from username and run a parallel analysis of how discrimination varies with type of post and user reputation. This provides additional evidence to support the main findings outlined above. Taken together, our empirical results are consistent with discrimination driven by belief-based partiality with some form of misspecification.

The source of discrimination has important implications for policies that aim to reduce discrimination. Suppose a policymaker cares about both efficiency and 'fairness,' defined as equal treatment for equal quality of output. If discrimination is driven by belief-based partiality with incorrect beliefs, the welfare criterion is clear: incorrect initial beliefs lead to suboptimal and unfair choices, relative to correct beliefs. Therefore, campaigns that aim to correct initial beliefs will improve choices along both dimensions, as will designing more objective measures of quality.

The findings on dynamics also highlight the pernicious effects of *incorrect* beliefs about group-based differences in initial evaluation standards. Kravitz and Platania (1993) conducted a survey on beliefs about affirmative action policies. The authors found that the majority held incorrect beliefs. Respondents viewed affirmative action policies as being much more widespread (required of all organizations) and as lowering evaluation standards to a much greater extent than is actually the case. Such incorrect beliefs can perpetuate inequality in outcomes, despite members of disadvantaged groups exceeding earlier standards and earning the relevant credentials. For example, prospective employers judging the education credentials of a minority candidate may discount them, relative to the same credentials from a non-minority candidate, if they believe that the minority candidate faced a lower standard to earn them. In this case, policies that remedy incorrect beliefs about initial evaluation standards will be particularly effective in mitigating discrimination down the road. Other policies, such as oversampling from discriminated groups at the initial stages, may also lead to more equal representation without exacerbating incorrect beliefs about evaluation standards.

References

Aigner, D. and G. Cain (1977): "Statistical Theories of Discrimination in the Labor Market," *Industrial and Labor Relations Review*, XXX, 175–187.

Altonji, J. G. and C. R. Pierret (2001): "Employer Learning and Statistical Discrimination," *Quarterly Journal of Economics*, 116, 313–350.

Anwar, S. and H. Fang (2006): "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence," *American Economic Review*, 96, 127–151.

Arnold, D., W. Dobbie, and C. S. Yang (2017): "Racial Bias in Bail Decisions," *NBER Working Paper*, 23421.

Arrow, K. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter and A. Rees, Princeton University Press, 3–33.

Ayalew, S., S. Manian, and K. Sheth (2018): "Discrimination from Below: Experimental Evidence on Female Leadership in Ethiopia," *mimeo.*

Bartos, V., M. Bauer, J. Chytilova, and F. Matejka (2016): "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition," *American Economic Review*, 106, 1437–75.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova (2009): "Powerful Women: Does Exposure Reduce Bias?*," *The Quarterly Journal of Economics*, 124, 1497–1540.

Becker, G. (1957): *The Economics of Discrimination*, Oregon State monographs: Studies in economics, Univ.Pr.

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

Bertrand, M. and E. Duflo (2016): "Field Experiments on Discrimination," North-Holland, Handbook of Economic Field Experiments, –.

Bertrand, M. and S. Mullainathan (2004): "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 94, 991–1013.

Biernat, M., M. Manis, and T. Nelson (1991): "Stereotypes and Standards of Judgment," 60, 485–499.

Biernat, M., T. K. Vescio, and M. Manis (1998): "Judging and Behavior Toward Members of Stereotyped Groups: A Shifting Standards Perspective," in *Intergroup Cognition and Intergroup Behavior*, ed. by C. Sedikides, J. Schopler, and C. Insko, Lawrence Erlbaum Associates.

Bird, C., A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan (2006): "Mining email social networks," in *Proceedings of the 2006 international workshop on Mining software repositories - MSR '06*, New York, New York, USA: ACM Press, 137.

Bocart, F. Y., M. Gertsberg, and R. A. Pownall (2018): "Glass Ceilings in the Art Market," *mimeo.*

Bohren, J. A. and D. Hauser (2018): "Social Learning with Model Misspecification: A Framework and a Robustness Result," .

Booth, A., M. Francesconi, and J. Frank (1999): "Glass ceilings or Sticky Floors," *mimeo, The University of Essex.*

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016a): "Beliefs About Gender," *NBER Working Paper*, 22972.

——— (2016b): "Stereotypes," *The Quarterly Journal of Economics*, 131, 1753–1794.

Coate, S. and G. C. Loury (1993): "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review*, 83, 1220–1240.

Coffman, K. B. (2014): "Evidence on Self-Stereotyping and the Contribution of Ideas," *Quarterly Journal of Economics*, 129, 1625–1660.

Dana, J., R. A. Weber, and J. X. Kuang (2007): "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 33, 67–80.

Danilov, A. and S. Saccardo (2017): "Discrimination in Disguise," *mimeo.*

Ewens, M., B. Tomlin, and L. C. Wang (2014): "Statistical Discrimination or Prejudice? A Large Sample Field Experiment," *The Review of Economics and Statistics*, 96, 119–134.

Fang, H. and A. Moro (2011): "Theories of Statistical Discrimination and Affirmative Action: A Survey," North-Holland, vol. 1 of *Handbook of Social Economics*, 133 – 200.

Fershtman, C. and U. Gneezy (2001): "Discrimination in a segmented society: An experimental approach," *The Quarterly Journal of Economics*, 116, 351–377.

Fiske, S., D. Bersoff, E. Borgida, K. Deaux, and M. Heilman (1991): "Social Science Research on Trial: Use of Sex Stereotyping Research in Price Waterhouse v. Hopkins," *American Psychologist*, 46, 1049–1060.

Fiske, S. and S. Taylor (1991): *Social Cognition*, McGraw-Hill Series in Electrical Engineering: Networks and S, McGraw-Hill.

Fiske, S. T. (1998): "Stereotyping, prejudice, and discrimination." in *The handbook of social psychology, Vols. 1-2, 4th ed.*, New York, NY, US: McGraw-Hill, 357–411.

Fryer, R. and M. Jackson (2008): "A Categorical Model of Cognition and Biased Decision-Making," *Contributions in Theoretical Economics*, 8.

Fryer, R. G. (2007): "Belief flipping in a dynamic model of statistical discrimination," *Journal of Public Economics*, 91, 1151–1166.

G. Greenwald, A., D. E. McGhee, and J. L. K. Schwartz (1998): "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," 74, 1464–80.

Gneezy, U., J. List, and M. Price (2012): "Toward an Understanding of Why People Discriminate: Evidence from a Series of Natural Field Experiments," *NBER Working Paper Series*, 17855.

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

Groot, W. and H. M. van den Brink (1996): "Glass ceilings or dead ends: Job promotion of men and women compared," *Economics Letters*, 53, 221–226.

Keilson, J. and U. Sumita (1982): "Uniform stochastic ordering and related inequalities," *The Canadian Journal of Statistics*, 10, 181–198.

Kelley, H. H. (1973): "The Process of Causal Attribution," *American Psychologist*, February, 107–128.

Knowles, J., N. Persico, and P. Todd (2001): "Racial Bias in Motor Vehicle Searches: Theory and Evidence," *Journal of Political Economy*, 109, 203–229.

Kravitz, D. and J. Platania (1993): "Attitudes and Beliefs About Affirmative Action: Effects of Target and of Respondent Sex and Ethnicity," 78, 928–938.

Leslie, L. M., C. F. Manchester, and P. C. Dahm (2017): "Why and When Does the Gender Gap Reverse? Diversity Goals and the Pay Premium for High Potential Women," *Academy of Management Journal*, 60, 402–432.

Lewis, G. B. (1986): "Gender and Promotions: Promotion Chances of White Men and Women in Federal White-Collar Employment," *The Journal of Human Resources*, 21, 406–419.

List, J. A. (2004): "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *The Quarterly Journal of Economics*, 119, 49–89.

Lundberg, S. and R. Startz (1983): "Private Discrimination and Social Intervention in Competitive Labor Markets," *American Economic Review*, LXXIII, 340–347.

Mengel, F., J. Sauermann, and U. Zolitz (2017): "Gender Bias in Teaching Evaluations," *Journal of the European Economic Association*.

Milgrom, P. (1981): "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 12, 380–391.

Milkman, K. L., M. Akinola, and D. Chugh (2012): "Temporal Distance and Discrimination: An Audit Study in Academia," *Psychological Science*, 23, 710–717.

Moss-Racusin, C. A., J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman (2012): "Science faculty's subtle gender biases favor male students," *Proceedings of the National Academy of Sciences*, 109, 16474–16479.

Olson, J. M., R. J. Ellis, and M. P. Zanna (1983): "Validating Objective Versus Subjective Judgment: Interest in Social Comparisons and Consistency Information," *Personality and Social Psychology Bulletin*, 9, 427–436.

Parsons, C. A., J. Sulaeman, M. C. Yates, and D. S. Hamermesh (2011): "Strike Three: Discrimination, Incentives, and Evaluation," *American Economic Review*, 101, 1410–35.

Petersen, T. and I. Saporta (2004): "The Opportunity Structure for Discrimination," *American Journal of Sociology*, 109, 852–901.

Phelps, E. (1972): "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62, 659–61.

Price, J. and J. Wolfers (2010): "Racial Discrimination Among NBA Referees," *Quarterly Journal of Economics*, 125, 1859–1887.

Pronin, E., D. Y. Lin, and L. Ross (2002): "The Bias Blind Spot: Perceptions of Bias in Self Versus Others," *Personality and Social Psychology Bulletin*, 28, 369–381.

Reuben, E., P. Sapienza, and L. Zingales (2014): "How stereotypes impair women's careers in science." *Proceedings of the National Academy of Sciences of the United States of America*, 111, 4403–8.

Riach, P. a. and J. Rich (2006): “An Experimental Investigation of Sexual Discrimination in Hiring in the English Labor Market,” *Advances in Economic Analysis & Policy*, 5, 1–20.

Rosette, A. S. and L. P. Tost (2010): “Agentic women and communal leadership: How role prescriptions confer advantage to top women leaders.” *Journal of Applied Psychology*, 95, 221–235.

Ross, L., D. Greene, and P. House (1977): “The “false consensus effect”: An egocentric bias in social perception and attribution processes,” *Journal of Experimental Social Psychology*, 13, 279 – 301.

Sarsons, H. (2017): “Interpreting Signals in the Labor Market: Evidence from Medical Referrals [Job Market Paper],” .

Schwartzstein, J. (2014): “Selective Attention and Learning,” *Journal of the European Economic Association*, 12, 1423–1452.

Snyder, M. L., R. E. Kleck, A. Strenta, and S. J. Mentzer (1979): “Avoidance of the Handicapped: An Attributional Ambiguity Analysis,” *Journal of Personality and Social Psychology*, 37, 2297–2306.

Tversky, A. and D. Kahneman (1983): “Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment,” *Psychological Review*, 90, 293–315.

Vasilescu, B., A. Capiluppi, and A. Serebrenik (2014): “Gender, representation and online participation: A quantitative study,” *Interacting with Computers*, 26, 488–511.

Williams, W. M. and S. J. Ceci (2015): “National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track,” *Proceedings of the National Academy of Sciences*, 112, 201418878.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT press.

A Additional Analysis from Section 2

A.1 Proofs of Propositions 1, 2 and 3

Proof of Proposition 1. From (5), it is clear that $|D(h_1, s_1)|$ is decreasing in τ_η iff $\hat{\mu}_M \neq \hat{\mu}_F$. From (4), if $c_F > 0$ and $\hat{\mu}_F = \hat{\mu}_M$, initial discrimination is equal to $D(h_1, s_1) = c_F$ for all $s_1 \in \mathbb{R}$, which is constant with respect to τ_η. In a model with both preference-based and belief-based partiality, initial discrimination is equal to

$$D(h_1, s_1) = \frac{\tau_q}{\tau_q + \tau_\eta}(\hat{\mu}_M - \hat{\mu}_F) + c_F.$$

Taking the limit, $\lim_{\tau_\eta \to \infty} D(h_1, s_1) = c_F$, which is nonzero iff $c_F \neq 0$. □

The following lemma is used in the proofs of Propositions 2 and 3.

Lemma 1. *Suppose an evaluator has subjective belief $\hat{\mu}$ and taste parameter c, and believes that all other evaluators are also this type. Then following any history h_t, the subjective posterior distribution of ability $f_{\hat{\mu}}(a|h_t)$ is normally distributed with mean*

$$\hat{\mu}(h_t) = \frac{\tau_a \hat{\mu} + \tau_{\varepsilon\eta} \sum_{n=1}^{t-1} s_n}{\tau_a + (t-1)\tau_{\varepsilon\eta}}$$

and precision $\tau_a(t) \equiv \tau_a + (t-1)\tau_{\varepsilon\eta}$, where

$$s_n = \left(\frac{\tau_q(n) + \tau_\eta}{\tau_\eta}\right)(v_n + c) - \left(\frac{\tau_q(n)}{\tau_\eta}\right)\hat{\mu}(h_n)$$

for all $n < t$ and

$$\tau_q(t) \equiv \tau_a(t)\tau_\varepsilon/(\tau_a(t) + \tau_\varepsilon). \tag{8}$$

Proof. Suppose $f_{\hat{\mu}}(a|h_1) \sim N(\hat{\mu}, 1/\tau_a)$. From (4), conditional on observing signal s_1, the first evaluation is

$$v_1 = \frac{\tau_q \hat{\mu} + \tau_\eta s_1}{\tau_q + \tau_\eta} - c.$$

It is possible to back out s_1 from observing v_1,

$$s_1 = s(v_1, \hat{\mu}) \equiv \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right)(v_1 + c) - \frac{\tau_q}{\tau_\eta}\hat{\mu}.$$

Recall $s_1 = a + \varepsilon_1 + \eta_1$. Therefore, the signal distribution, conditional on ability, is normally distributed and independent of $\hat{\mu}$, $f_s(s_1|a) \sim N(a, 1/\tau_{\varepsilon\eta})$, where $\tau_{\varepsilon\eta} \equiv$

$\tau_\eta \tau_\varepsilon/(\tau_\eta + \tau_\varepsilon)$. Consider the posterior distribution of ability, following evaluation v_1. From Bayes rule,

$$f_{\hat{\mu}}(a|v_1, h_1) = \frac{P_{\hat{\mu}}(v_1|a, h_1) f_{\hat{\mu}}(a|h_1)}{\int P_{\hat{\mu}}(v_1|a', h_1) f_{\hat{\mu}}(a'|h_1) da'} = \frac{f_s(s(v_1, \hat{\mu})|a, h_1) f_{\hat{\mu}}(a|h_1)}{\int f_s(s(v_1, \hat{\mu})|a', h_1) f_{\hat{\mu}}(a'|h_1) da'} \ ,$$

where the second equality follows from $P_{\hat{\mu}}(v_1|a, h_1) = \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) f_s(s(v_1, \hat{\mu})|a)$. The normal distribution is conjugate to itself for a normal likelihood function. Since the prior belief about ability is normal, and the signal distribution conditional on ability is normal, the posterior belief about ability $f_{\hat{\mu}}(a|v_1, h_1)$ is also normal,

$$f_{\hat{\mu}}(a|v_1, h_1) \sim N\left(\frac{\tau_a \hat{\mu} + \tau_{\varepsilon\eta} s(v_1, \hat{\mu})}{\tau_a + \tau_{\varepsilon\eta}}, \frac{1}{\tau_a + \tau_{\varepsilon\eta}}\right).$$

Given the normality of the posterior belief about ability, we can define the evaluation and belief-updating processes recursively. Let $\hat{\mu}(h_t)$ and $\tau_a(t)$ denote the mean and precision of the distribution of ability at the beginning of period t, following history h_t, i.e. $f_{\hat{\mu}}(a|h_t) \sim N(\hat{\mu}(h_t), 1/\tau_a(t))$. The evaluation process in period $t > 1$ is analogous to $t = 1$. The posterior distribution of quality q_t, conditional on observing signal s_t, is normal,

$$q_t|s_t, h_t \sim N\left(\frac{\tau_q(t)\hat{\mu}(h_t) + \tau_\eta s_t}{\tau_q(t) + \tau_\eta}, \frac{1}{\tau_q(t) + \tau_\eta}\right),$$

where $\tau_q(t)$ is as defined in (8). The evaluator maximizes her expected payoff by choosing

$$v_t = \frac{\tau_q(t)\hat{\mu}(h_t) + \tau_\eta s_t}{\tau_q(t) + \tau_\eta} - c_g \tag{9}$$

Therefore, it is possible to back out s_t from v_t,

$$s_t = s(v_t, \hat{\mu}(h_t), t) \equiv \left(\frac{\tau_q(t) + \tau_\eta}{\tau_\eta}\right)(v_t + c) - \left(\frac{\tau_q(t)}{\tau_\eta}\right)\hat{\mu}(h_t).$$

The posterior update is also analogous to $t = 1$. For $t > 1$, the posterior belief about ability, conditional on observing evaluation v_t, is normally distributed with mean

$$\hat{\mu}(h_{t+1}) = \frac{\tau_a(t)\hat{\mu}(h_t) + \tau_{\varepsilon\eta} s(v_t, \hat{\mu}(h_t), t)}{\tau_a(t) + \tau_{\varepsilon\eta}}$$

and precision

$$\tau_a(t+1) = \tau_a(t) + \tau_{\varepsilon\eta}.$$

Initialize $\hat{\mu}(h_1) = \hat{\mu}$ and $\tau_a(1) = \tau_a$. Solving the recursive expressions for $\hat{\mu}(h_t)$ and $\tau_a(t)$ yields solution

$$\begin{aligned} \hat{\mu}(h_t) &= \frac{\tau_a \hat{\mu} + \tau_{\varepsilon\eta} \sum_{n=1}^{t-1} s(v_n, \hat{\mu}(h_n), n)}{\tau_a + (t-1)\tau_{\varepsilon\eta}} && (10) \\ \tau_a(t) &= \tau_a + (t-1)\tau_{\varepsilon\eta}. && (11) \end{aligned}$$

Therefore, when the prior belief about ability is normal, the posterior belief about ability $f_{\hat{\mu}}(a|v_t, h_t)$ is also normal with mean $\hat{\mu}(h_{t+1})$ and precision $\tau_a(t+1)$ defined in (10) and (11). □

Proof of Proposition 2. We proceed by a series of lemmas.

Lemma 2. *Suppose $c_F = 0$. If $\hat{\mu}_F(h_t) < \hat{\mu}_M(h_t)$, then for all v_t,*

1. *$\hat{\mu}_F(h_{t+1}) < \hat{\mu}_M(h_{t+1})$ i.e. there is no belief reversal between periods t and $t+1$;*
2. *$\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) < \hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)$ i.e. the difference in means decreases between periods t and $t+1$.*

Proof. Suppose $\hat{\mu}_F(h_t) < \hat{\mu}_M(h_t)$ and $c_F = 0$. Then

$$\hat{\mu}_g(h_{t+1}) = \frac{\tau_a(t)\hat{\mu}_g(h_t) + \tau_{\varepsilon\eta} s_g(v_t, \hat{\mu}_g(h_t), t)}{\tau_a(t) + \tau_{\varepsilon\eta}},$$

where

$$s_g(v_t, \hat{\mu}_g(h_t), t) \equiv \left(\frac{\tau_q(t) + \tau_\eta}{\tau_\eta}\right) v_t - \left(\frac{\tau_q(t)}{\tau_\eta}\right) \hat{\mu}_g(h_t).$$

Following evaluation v_t,

$$s_M(v_t, \hat{\mu}_M(h_t), t) - s_F(v_t, \hat{\mu}_F(h_t), t) = -\left(\frac{\tau_q(t)}{\tau_\eta}\right)(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t))$$

Therefore,

$$\begin{aligned}\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) &= \left(\frac{\tau_a(t)}{\tau_a(t)+\tau_{\varepsilon\eta}}\right)(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)) \\ &\quad + \left(\frac{\tau_{\varepsilon\eta}}{\tau_a(t)+\tau_{\varepsilon\eta}}\right)(s_M(v_t, \hat{\mu}_M(h_t), t) - s_F(v_t, \hat{\mu}_F(h_t), t)) \\ &= \left(\frac{\tau_a(t)}{\tau_a(t)+\tau_{\varepsilon\eta}} - \frac{\tau_{\varepsilon\eta}\tau_q(t)}{(\tau_a(t)+\tau_{\varepsilon\eta})\tau_\eta}\right)(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)), \quad (12)\end{aligned}$$

which is positive if

$$\begin{aligned} & \frac{\tau_a(t)}{\tau_a(t)+\tau_{\varepsilon\eta}} - \frac{\tau_{\varepsilon\eta}\tau_q(t)}{(\tau_a(t)+\tau_{\varepsilon\eta})\tau_\eta} > 0 \\ \Leftrightarrow \quad & \frac{\tau_a(t)\tau_\eta - \frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon+\tau_\eta} \times \frac{\tau_a(t)\tau_\varepsilon}{\tau_a(t)+\tau_\varepsilon}}{(\tau_a(t) + \frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon+\tau_\eta})\tau_\eta} > 0. \qquad (13)\end{aligned}$$

This will be the case if the numerator of (13) is positive,

$$\begin{aligned} & \tau_a(t)\tau_\eta - \frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon+\tau_\eta} \times \frac{\tau_a(t)\tau_\varepsilon}{\tau_a(t)+\tau_\varepsilon} > 0 \\ \Leftrightarrow \quad & (\tau_\varepsilon+\tau_\eta)(\tau_a(t)+\tau_\varepsilon) > \tau_\varepsilon^2 \\ \Leftrightarrow \quad & \tau_\varepsilon^2 + \tau_\varepsilon\tau_\eta + \tau_a(t)(\tau_\varepsilon+\tau_\eta) > \tau_\varepsilon^2 \\ \Leftrightarrow \quad & \tau_\varepsilon\tau_\eta + \tau_a(t)(\tau_\varepsilon+\tau_\eta) > 0,\end{aligned}$$

which always holds since all precisions are positive. Therefore, $\hat{\mu}_M(h_{t+1}) > \hat{\mu}_F(h_{t+1})$.

From (12), $\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) < \hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)$ iff (13) is less than one, which always holds since

$$\frac{\tau_a(t)\tau_\eta - \frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon+\tau_\eta} \times \frac{\tau_a(t)\tau_\varepsilon}{\tau_a(t)+\tau_\varepsilon}}{\tau_a(t)\tau_\eta + \frac{\tau_\varepsilon\tau_\eta^2}{\tau_\varepsilon+\tau_\eta}} = \frac{\tau_a(t)\tau_\eta}{\tau_a(t)\tau_\eta + \frac{\tau_\varepsilon\tau_\eta^2}{\tau_\varepsilon+\tau_\eta}} - \frac{\frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon+\tau_\eta} \times \frac{\tau_a(t)\tau_\varepsilon}{\tau_a(t)+\tau_\varepsilon}}{\tau_a(t)\tau_\eta + \frac{\tau_\varepsilon\tau_\eta^2}{\tau_\varepsilon+\tau_\eta}},$$

where the first term on the right hand side is less than one, and the second term is negative. □

Lemma 3. *Suppose $c_F = 0$. A discrimination reversal occurs between periods t and $t+1$ iff there is a belief reversal between periods t and $t+1$.*

Proof. Suppose $c_F = 0$. From (9), discrimination in period t is equal to

$$D(h_t, s_t) = \frac{\tau_q(t)}{\tau_q(t) + \tau_\eta}(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)). \tag{14}$$

Therefore, discrimination reverses between periods t and $t+1$ if and only if $\hat{\mu}_M(h_t) > \hat{\mu}_F(h_t)$ and $\hat{\mu}_M(h_{t+1}) < \hat{\mu}_F(h_{t+1})$, or vice versa. □

We can now complete the proof of Proposition 2.

Proof. Suppose $\hat{\mu}_F < \hat{\mu}_M$ and $c_F = 0$. From Lemma 2, for all v_1, $\hat{\mu}_F(h_2) < \hat{\mu}_M(h_2)$ and $\hat{\mu}_M(h_2) - \hat{\mu}_F(h_2) < \hat{\mu}_M - \hat{\mu}_F$. By induction, $\hat{\mu}_F(h_t) < \hat{\mu}_M(h_t)$ and

$$\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) < \hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)$$

for all t and h_{t+1}. From Lemma 3, there is no discrimination reversal between any periods t and $t+1$, since $\hat{\mu}_F(h_t) < \hat{\mu}_M(h_t)$ for all t and h_t.

It remains to show that discrimination decreases. Discrimination in period t is equal to

$$D(h_t, s_t) = \left(\frac{\tau_q(t)}{\tau_q(t) + \tau_\eta}\right)(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)),$$

and in period $t+1$ is equal to

$$\begin{aligned} D(h_{t+1}, s_t) &= \left(\frac{\tau_q(t+1)}{\tau_q(t+1) + \tau_\eta}\right)(\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1})) \\ &= \left(\frac{\tau_q(t+1)}{\tau_q(t+1) + \tau_\eta}\right)\left(\frac{\tau_a(t)\tau_\eta - \tau_{\varepsilon\eta}\tau_q(t)}{(\tau_a(t) + \tau_{\varepsilon\eta})\tau_\eta}\right)(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)). \end{aligned}$$

□

Proof of Proposition 3. Type θ_1's belief about male and female ability evolve as in Lemma 1, since this type believes that all other evaluators have the same beliefs as it. Type θ_2's belief about male ability also evolve as in Lemma 1, since both types have the same prior belief about male ability. Thus, the novelty stems from characterizing how type θ_2's belief about female ability evolves.

When type θ_2 observes evaluation v_1, she believes that with probability p, it is from a heuristic type who observed signal $s_1^1(v_1) = \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) v_1 - \left(\frac{\tau_q}{\tau_\eta}\right)\hat{\mu}_F^1$, and with probability $1-p$, it is from an impartial type who observed signal $s_1^2(v_1) = \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) v_1 - \left(\frac{\tau_q}{\tau_\eta}\right)\hat{\mu}_M$.

Note $s_1^1(v_1) > s_1^2(v_1)$. Therefore, the likelihood function for evaluation v_1 is a mixture of two normal distributions,

$$f_v(v_1|a) = (pf_s(s_1^1(v_1)|a) + (1-p)f_s(s_1^2(v_1)|a)) \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right).$$

Since the prior belief $f_a(a) \sim N(\hat{\mu}_F^2, 1/\tau_a)$ is normal, the posterior belief will be a mixture of two normal distributions,

$$f_a(a|v_1) = pf_1(a|v_1)\frac{C_1}{C} + (1-p)f_2(a|v_1)\frac{C_2}{C},$$

where

$$\begin{aligned} f_1(a|v_1) &\sim N\left(\frac{\tau_a\hat{\mu}_F + \tau_{\varepsilon\eta}s_1^1(v_1)}{\tau_a + \tau_{\varepsilon\eta}}, \frac{1}{\tau_a + \tau_{\varepsilon\eta}}\right) \\ f_2(a|v_1) &\sim N\left(\frac{\tau_a\hat{\mu}_F + \tau_{\varepsilon\eta}s_1^2(v_1)}{\tau_a + \tau_{\varepsilon\eta}}, \frac{1}{\tau_a + \tau_{\varepsilon\eta}}\right) \end{aligned}$$

are the posterior distributions of ability, conditional on observing signals $s_1^1(v_1)$ and $s_1^2(v_1)$, respectively, and

$$\begin{aligned} C_1 &= \int f_s(s_1^1(v_1)|a)f_a(a)da \\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\tau_a\tau_{\varepsilon\eta}}{\tau_a + \tau_{\varepsilon\eta}}}\exp(-0.5(\tau_a(\hat{\mu}_F)^2 + s_1^1(v_1)^2\tau_{\varepsilon\eta} - (\tau_a + \tau_{\varepsilon\eta})\hat{\mu}_1(v_1)^2) \\ C_2 &= \int f_s(s_1^2(v_1)|a)f_a(a)da \\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\tau_a\tau_{\varepsilon\eta}}{\tau_a + \tau_{\varepsilon\eta}}}\exp(-0.5(\tau_a(\hat{\mu}_F)^2 + s_1^2(v_1)^2\tau_{\varepsilon\eta} - (\tau_a + \tau_{\varepsilon\eta})\hat{\mu}_2(v_1)^2) \\ C &= pC_1 + (1-p)C_2 \end{aligned}$$

are the normalization coefficients. The convolution of a normal distribution with a mixture of two normal distributions is a mixture of two normal distributions. Therefore, the prior belief about quality in the second period, $g(q_2|v_1)$, is a mixture of two normal distributions. Therefore, the posterior belief about quality in the second period, conditional on observing signal s_2, $g(q_2|v_1, s_2)$, is also a mixture of two normal distributions,

$$g(q_2|s_2, v_1) = p\frac{C_1 D_1}{CD} g_1(q_2|s_2, v_1) + (1-p)\frac{C_2 D_2}{CD} g_2(q_2|s_2, v_1)$$

where, given $\hat{\mu}_1(v_1)$ and $\hat{\mu}_2(v_1)$ are the means of $f_1(a|v_1)$ and $f_2(a|v_2)$, respectively, and $\tau_{q,2} \equiv \frac{(\tau_a+\tau_{\varepsilon\eta})\tau_\varepsilon}{\tau_a+\tau_{\varepsilon\eta}+\tau_\varepsilon}$,

$$\begin{aligned} g_1(q_2|s_2, v_1) &\sim N\left(\frac{\tau_{q,2}\hat{\mu}_1(v_1) + \tau_\eta s_2}{\tau_{q,2}+\tau_\eta}, \frac{1}{\tau_{q,2}+\tau_\eta}\right) \\ g_2(q_2|s_2, v_1) &\sim N\left(\frac{\tau_{q,2}\hat{\mu}_2(v_1) + \tau_\eta s_2}{\tau_{q,2}+\tau_\eta}, \frac{1}{\tau_{q,2}+\tau_\eta}\right) \end{aligned}$$

and, given $\hat{\mu}_1(v_1, s_2)$ and $\hat{\mu}_2(v_1, s_2)$ are the means of g_1 and g_2, respectively,

$$\begin{aligned} D_1 &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\tau_{q,2}\tau_\eta}{\tau_{q,2}+\tau_\eta}} \exp(-0.5(\tau_{q,2}\hat{\mu}_1(v_1)^2 + s_2^2\tau_\eta - (\tau_{q,2}+\tau_\eta)\hat{\mu}_1(v_1, s_2)^2) \\ D_2 &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\tau_{q,2}\tau_\eta}{\tau_{q,2}+\tau_\eta}} \exp(-0.5(\tau_{q,2}\hat{\mu}_2(v_1)^2 + s_2^2\tau_\eta - (\tau_{q,2}+\tau_\eta)\hat{\mu}_2(v_1, s_2)^2) \\ D &= p\frac{C_1}{C}D_1 + (1-p)\frac{C_2}{C}D_2 \end{aligned}$$

are the normalizing coefficients. Define $\gamma(v_1, s_2) \equiv \frac{pC_1D_1}{CD}\hat{\mu}_1(v_1) + \frac{(1-p)C_2D_2}{CD}\hat{\mu}_2(v_1)$. Then in the second period, the impartial type gives females evaluation

$$v_F^2(s_2, v_1) = \left(\frac{\tau_\eta}{\tau_{q,2}+\tau_\eta}\right)s_2 + \left(\frac{\tau_{q,2}}{\tau_{q,2}+\tau_\eta}\right)\gamma(v_1, s_2).$$

In the second period, following initial evaluation v_1 and signal s_2, the heuristic type gives females evaluation

$$v_F^1(s_2, v_1) = \left(\frac{\tau_\eta}{\tau_{q,2}+\tau_\eta}\right)s_2 + \left(\frac{\tau_{q,2}}{\tau_{q,2}+\tau_\eta}\right)\hat{\mu}_F^1(v_1),$$

where from Lemma 1, $\hat{\mu}_F^1(v_1) = \frac{\tau_a\hat{\mu}_F^1 + \tau_{\varepsilon\eta}s_1^1(v_1)}{\tau_a+\tau_{\varepsilon\eta}}$, and both types give males evaluation

$$v_M(s_2, v_1) = \left(\frac{\tau_\eta}{\tau_{q,2}+\tau_\eta}\right)s_2 + \left(\frac{\tau_{q,2}}{\tau_{q,2}+\tau_\eta}\right)\hat{\mu}_M(v_1),$$

where from Lemma 1, $\hat{\mu}_M(v_1) = \frac{\tau_a\hat{\mu}_M + \tau_{\varepsilon\eta}s_1^2(v_1)}{\tau_a+\tau_{\varepsilon\eta}}$. Therefore, aggregate discrimination in the second period is equal to

$$\begin{aligned} D(v_1, s_2) &= v_M(s_2, v_1) - p v_F^1(s_2, v_1) - (1-p) v_F^2(s_2, v_1) \\ &= \left(\frac{\tau_{q,2}}{\tau_{q,2} + \tau_\eta} \right) (\hat{\mu}_M(v_1) - p \hat{\mu}_F^1(v_1) - (1-p)\gamma(v_1, s_2)). \end{aligned}$$

Aggregate discrimination reverses at (v_1, s_2) if $D(v_1, s_2) < 0$. We know that at $p = 0$, $D(v_1, s_2) = 0$, as this is the case with no partiality, and at $p = 1$, $D(v_1, s_2) > 0$, as this is the case with a single type of evaluator with belief-based partiality from Proposition 2. Therefore, if the derivative of $D(v_1, s_2)$ with respect to p is negative at $p = 0$, discrimination will become negative for an interval $(0, \overline{p})$ before becoming positive. This derivative simplifies to showing that

$$1 < \left(\frac{\tau_\varepsilon^2}{(\tau_\varepsilon + \tau_\eta)(\tau_a + \tau_\varepsilon)} \right) \left(1 + \frac{C_1 D_1}{C_2 D_2} \right).$$

From the expressions above,

$$\begin{aligned} \frac{C_1 D_1}{C_2 D_2} &= \exp(-0.5\tau_{\varepsilon\eta}(s_1^1(v_1)^2 - s_1^2(v_1)^2) + 0.5(\tau_a + \tau_{\varepsilon\eta} - \tau_{q,2})(\hat{\mu}_1(v_1)^2 - \hat{\mu}_2(v_1)^2) \\ &\quad +0.5(\tau_{q,2} + \tau_\eta(\hat{\mu}_1(v_1, s_2)^2 - \hat{\mu}_2(v_1, s_2)^2)), \end{aligned}$$

which is increasing in v_1 and decreasing in s_2, and becomes arbitrarily large as v_1 approaches negative infinity or s_2 approaches infinity. Therefore, for any prior beliefs about ability for each type, it is possible for discrimination to reverse in the second period. □

A.2 Robustness to 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.

A.2.1 Continuous Ability Distributions

Suppose a worker has ability distributed according to $a \sim f_\mu(a)$ with parameter μ, and assume that the family of distributions $\{f_\mu(a)\}_{\mu \in \mathbb{R}}$ 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 and no discrimination reversal occurs.

Each task has hidden quality $q_t = a + \varepsilon_t$, where $\varepsilon_t \sim N(0, 1/\tau_\varepsilon)$. As before, 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 $v_t = E_\mu[q_t|s_t, h_t]$.

Consider the first period. The prior distribution about quality

$$f_\mu(q_1) = \int_A f(q_1|a) f_\mu(a) da$$

is the convolution of the prior distribution about 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 s_1 is

$$f_\mu(q_1|s_1) = \frac{f(s_1|q_1) f_\mu(q_1)}{\int_Q f(s_1|q_1) f_\mu(q_1) dq_1}.$$

The MLRP is preserved under Bayesian updating when the likelihood function is independent of μ. Therefore, since $f(s_1|q_1)$ is independent of μ, the posterior $f_\mu(q_1|s_1)$ satisfies the MLRP in μ. By FOSD, $E_\mu[q_1|s_1]$ is increasing in μ. After observing s_1, the evaluator reports initial evaluation

$$\begin{aligned} v(s_1, \mu) &= E_\mu[q_1|s_1] \\ &= \int_Q q_1 f(q_1|s_1) dq \\ &= \frac{\int_Q \int_A q_1 f(s_1|q_1) f(q_1|a) f_\mu(a) da dq_1}{\int_Q f(s_1|q_1) f_\mu(q_1) dq_1}. \end{aligned} \tag{15}$$

The initial evaluation is strictly increasing in s_1, and therefore, each signal s_1 maps to a unique evaluation $v(s_1, \mu)$. Let $s(v_1, \mu)$ be the signal required to receive initial evaluation v_1, i.e. the solution to

$$v_1 = \frac{\int_Q \int_A q_1 f(s(v_1, \mu)|q_1) f(q_1|a) f_\mu(a) da dq_1}{\int_Q f(s(v_1, \mu)|q_1) f_\mu(q_1) dq_1}. \tag{16}$$

After observing evaluation v_1, the posterior public belief about ability is

$$f_\mu(a|v_1) = \frac{f(s(v_1,\mu)|a)f_\mu(a)}{f_\mu(s(v_1,\mu))}. \tag{17}$$

Suppose that the posterior distribution of ability $f_\mu(a|v_1)$ satisfies the MLRP in μ. Then by the same reasoning as above, $f(q_2|v_1,s_2)$ satisfies the MLRP in μ and $E_\mu[q_2|v_1,s_2]$ is increasing in μ. For any v_1 and s_2, the second period evaluation $v_2(s_2,v_1,\mu)$ is also increasing in μ. Therefore, if all evaluators believe that males have a higher parameter μ than females, i.e. $\hat{\mu}_M > \hat{\mu}_F$, both the first and the second period evaluation is higher for males, $v_1(s_1,\hat{\mu}_M) > v_1(s_1,\hat{\mu}_F)$ and $v_2(s_2,v_1,\hat{\mu}_M) > v_2(s_2,v_1,\hat{\mu}_F)$, and no discrimination reversal occurs.

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

$$\begin{aligned}
& \frac{\partial^2}{\partial\mu\partial a}\left(\log f(s(v_1,\mu)|a) + \log f_\mu(a)\right) > 0 \\
i.e. \quad & \frac{\partial^2}{\partial\mu\partial a}\left(-\frac{\tau_\varepsilon\tau_\eta}{2(\tau_\varepsilon+\tau_\eta)}((s(v_1,\mu)-a)^2 + \log f_\mu(a)\right) > 0 \\
i.e. \quad & \frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon+\tau_\eta}\frac{\partial s(v_1,\mu)}{\partial\mu} + \frac{\partial^2}{\partial\mu\partial a}\log f_\mu(a) > 0 \qquad (18)
\end{aligned}$$

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 $\frac{\partial^2}{\partial\mu\partial a}\log f_\mu(a)$ is positive, since by assumption $f_\mu(a)$ satisfies the MLRP.

We numerically show that (18) holds for several classes of distributions by (i) numerically solving (16) for $s(v_1,\mu)$, and (ii) numerically calculating $\frac{\partial s(v_1,\mu)}{\partial\mu}$.

Exponential Distribution. The exponential distribution has density $f_\mu(a) = \frac{1}{\mu}e^{-a/\mu}$, where $a \in [0,\infty)$ and $E[a] = \mu$. Therefore, $\frac{\partial^2}{\partial\mu\partial a}\log f_\mu(a) = 1/\mu^2 > 0$ and the prior distribution satisfies the MLRP in μ. We show that (18) holds numerically for all parameters $\mu \in \{.01, .02, ..., 2.99, 3\}$ and $v \in \{-2, -1.99, ..., 5.99, 6\}$. Given that $\frac{\partial^2}{\partial\mu\partial a}\log f_\mu(a)$ is independent of a, (18) 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 $f_{\alpha,\beta}(a) = \frac{1}{B(\alpha,\beta)} a^{\alpha-1}(1-a)^{\beta-1}$, where $a \in [0,1]$ and $E[a] = \alpha/(\alpha+\beta)$. Therefore, $\frac{\partial^2}{\partial\alpha\partial a} \log f_{\alpha,\beta}(a) = 1/a > 0$ 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 (18) 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 $f_{k,\theta}(a) = \frac{1}{\Gamma(k)\theta^k} a^{k-1} e^{-a/\theta}$, where $a \in (0,\infty)$ and $E[a] = k\theta$. Therefore, $\frac{\partial^2}{\partial k\partial a} \log f_{k,\theta}(a) = 1/a > 0$ 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 (18) does not hold for all a, k, θ and v. Since (18) is sufficient, but not necessary, for a reversal, we can also show that $E_k[a|v_1]$ 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.2.2 Binary Ability and Quality

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

$$\frac{v(s,p)}{1-v(s,p)} = \frac{f^h(s)}{f^l(s)} * \frac{\phi(p)}{1-\phi(p)}. \tag{19}$$

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 (19),

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

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

$$\frac{B(v,p)}{1-B(v,p)} = \frac{f^h(s(v,p))\rho_H + f^l(s(v,p))(1-\rho_H)}{f^h(s(v,p))\rho_L + f^l(s(v,p))(1-\rho_L)} * \frac{p}{1-p}. \tag{20}$$

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 (20) is equivalent to

$$\frac{d}{dp}\left[\log Pr(v|H) - \log Pr(v|L) + \log p + \log(1-p)\right] > 0 \tag{21}$$

where

$$Pr(v|a) = f^h(s(v,p))\rho_a + f^l(s(v,p))(1-\rho_a).$$

This is equivalent to showing

$$\frac{\frac{d}{dp}Pr(v|H)}{Pr(v|H)} - \frac{\frac{d}{dp}Pr(v|L)}{Pr(v|L)} + \frac{1}{p} + \frac{1}{1-p} > 0, \tag{22}$$

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 (22) 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.

A.3 Robustness to Alternative Models

In this subsection, we explore two alternative models: (i) coarse evaluations and (ii) shifting standards. We show that our main results from Section 2 extend to these settings.

A.3.1 Coarse Evaluations

Set-up. Suppose that the set-up is identical to Section 2.1, 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 5 (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

$$E[q_t|h_t, s_t, g] \geq c_g, \tag{23}$$

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 2, 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 4, 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 5, 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 4. *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 4. 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 4. *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

$$\begin{aligned}
\frac{f_\mu(a|v_1 = 1)}{f_\mu(a'|v_1 = 1)} &= \frac{P_\mu(v_1 = 1|a)}{P_\mu(v_1 = 1|a')} \cdot \frac{f_\mu(a)}{f_\mu(a')} \\
&= \frac{\Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a - \overline{s}(\mu, c))\right)}{\Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a' - \overline{s}(\mu, c))\right)} \cdot \frac{\phi(\sqrt{\tau_a}(a - \mu))}{\phi(\sqrt{\tau_a}(a' - \mu))}. \qquad (24)
\end{aligned}$$

The goal is to show that (24) 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,

$$\frac{\partial^2}{\partial a \partial \mu} \log P_\mu(v_1 = 1|a) + \log f_\mu(a) \geq 0. \tag{25}$$

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 (25),

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

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 (25)

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

Therefore, need to show that for all x,

$$\begin{aligned}
\tau_a - \left(\frac{\phi(x)x}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2}\right)\left(\frac{\tau_q \tau_{\varepsilon\eta}}{\tau_\eta}\right) &\geq 0 \\
\Leftrightarrow \tau_x - \left(\frac{\phi(x)x}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2}\right) &\geq 0, \qquad (26)
\end{aligned}$$

where $\tau_x \equiv \frac{\tau_a \tau_\eta}{\tau_q \tau_{\varepsilon\eta}}$. From Stack Exchange[22], we know that

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

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

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, (26) holds for all x. Therefore, for all $a > a'$, (24) is increasing in μ and $\{f_\mu(a|v = 1)\}_{\mu\in\mathbb{R}}$ satisfies the MLRP. □

Given Lemma 4, 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 5 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 5. *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 4, $\{f_\mu(a|v_1 = 1)\}_{\mu\in\mathbb{R}}$ 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 Keilson and Sumita (1982), 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,

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

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.

A.3.2 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

$$v(h_t, s_t, g) = \frac{\tau_{q,t}\hat{\mu}_g(h_t) + \tau_\eta s_t}{\tau_{q,t} + \tau_\eta} - c(r(h_t)) - c_g, \tag{27}$$

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 6. *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 $v \in \mathbb{R}$.*

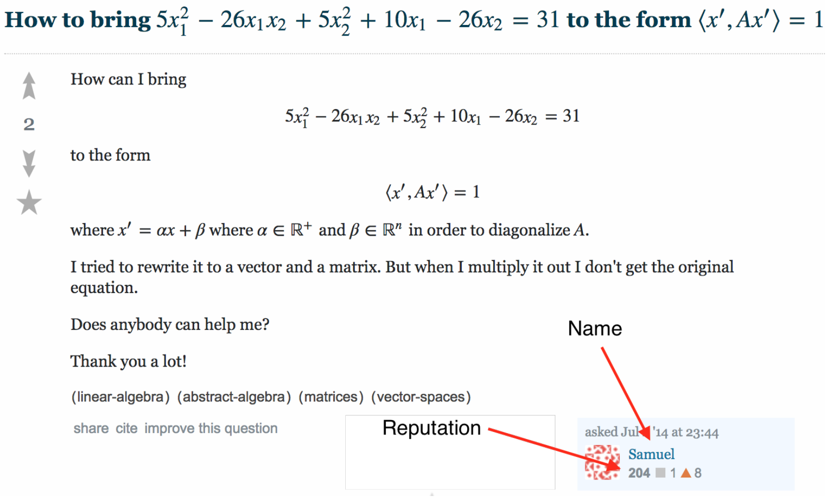
Shifting standards implies that the positive evaluation's negative impact on the benchmark of evaluation outweighs the positive impact on the belief about the worker's expected quality. Note that the definition is required to hold at all evaluations $v \in \mathbb{R}$, but this is not restrictive, as 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 $\bar{c}$ such that if the new benchmark of evaluation exceeds this cut-off, $c(v_1) > \bar{c}$, 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.

B Additional Empirical Analysis

B.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.



How to bring $5x_1^2 - 26x_1x_2 + 5x_2^2 + 10x_1 - 26x_2 = 31$ to the form $\langle x', Ax' \rangle = 1$

How can I bring

$$5x_1^2 - 26x_1x_2 + 5x_2^2 + 10x_1 - 26x_2 = 31$$

to the form

$$\langle x', Ax' \rangle = 1$$

where $x' = \alpha x + \beta$ where $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^n$ in order to diagonalize A.

I tried to rewrite it to a vector and a matrix. But when I multiply it out I don't get the original equation.

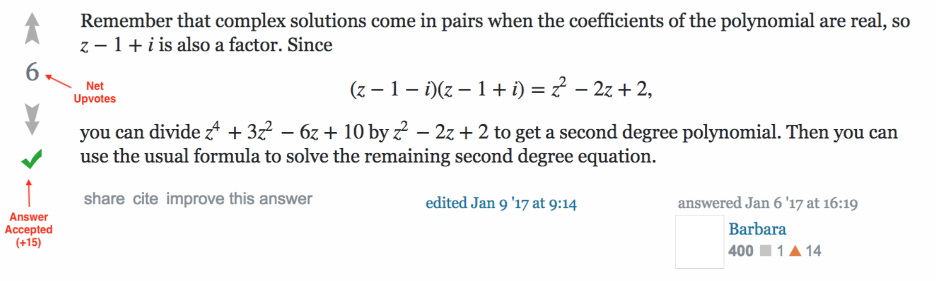
Does anybody can help me?

Thank you a lot!

(linear-algebra) (abstract-algebra) (matrices) (vector-spaces)

asked Jul '14 at 23:44
Samuel
204 ■ 1 ▲ 8

Figure 5. Question Post



Remember that complex solutions come in pairs when the coefficients of the polynomial are real, so $z - 1 + i$ is also a factor. Since

$$(z - 1 - i)(z - 1 + i) = z^2 - 2z + 2,$$

you can divide $z^4 + 3z^2 - 6z + 10$ by $z^2 - 2z + 2$ to get a second degree polynomial. Then you can use the usual formula to solve the remaining second degree equation.

edited Jan 9 '17 at 9:14

answered Jan 6 '17 at 16:19
Barbara
400 ■ 1 ▲ 14

Figure 6. Answer Post

B.2 Robustness

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

Table 3. 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 4. Dynamics: Effect of Gender on Evaluation Questions, Novice and Advanced (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, using only the first vote on a post in our experiment from each evaluator.

Table 5. Subjectivity: Effect of Gender on Evaluation of Novice Answers and Questions

	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 (.82)	−0.28 (.16)	2.17 (1.07)	0.44 (.22)	−1.15 (.96)	−0.28 (.19)
Question					−0.42 (.96)	−0.13 (.19)
Male*Question					3.32 (1.35)	0.72 (.27)
Constant	3.55 (.58)	0.70 (.12)	3.13 (.76)	0.57 (.15)	3.55 (.68)	0.70 (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 6. Dynamics: Effect of Gender on Evaluation Questions, Novice and Advanced

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

B.3 Voter Characteristics

Table 7. Voter Characteristics by Post Type

	Voter Reputation	Voter Gender: % Female
	(1)	(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.

B.4 Observational Data Analysis

We next present the analysis of the observational data described in Section 3.4.

B.4.1 Description of Algorithm to Code Gender.

Vasilescu et al. (2014) 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 (Bird, Gourley, Devanbu, Gertz, and Swaminathan 2006). Vasilescu et al. (2014) 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.

B.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 8, 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 8. 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 9, 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.[23]

Table 9. Likelihood of Generating a Second Post (Quartiles)

	Pooled	Questions	Answers
	(1)	(2)	(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 10, we repeat the analysis from Table 8 for the likelihood of generating

[23] 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 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.

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 $t = 3, ..., 10$.

Table 10. 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 11, 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 11, neither the coefficient on the gender dummy nor on the interaction with total reputation is significant in any

Table 11. Likelihood of Generating Next Post

	(1)	(2)	(3)
Male	0.01	0.02	0.01
	(.01)	(.01)	(.17)
Total Reputation	0.05	0.22	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	-0.19
	(.01)	(.01)	(.01)
Constant	0.08	-0.91	-2.48
	(.01)	(.03)	(.16)
Post Number Dummies	No	Yes	Yes
# Obs	244,854	244,854	244,854

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.

of the specifications. In Column (3), the coefficient on the interaction of gender and reputation earned on previous post also is not significant.

B.4.3 Autocorrelation in Error Process for Quality

Suppose individual i has ability a_i. From Section 2.1, 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:

$$v_{i,t} = \beta_0 + \beta_1 * \mathbf{1}_{g_i=M} + u_i + e_{i,t}, \tag{28}$$

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 $\varepsilon_{i,t}$ (Wooldridge 2010). 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 (28), 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 (9),

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

where $\tau_q(t)$ is defined in (8), and the second line follows from signal $s_{i,t} = a_i + \varepsilon_{i,t} + \eta_{i,t}$. From (10), 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:

$$\begin{aligned} v_{i,t} = \quad & \beta_0 + \beta_1 * \mathbf{1}_{g_i=M} + \beta_2 * R_{i,t} + \beta_3 * R_{i,t} * \mathbf{1}_{g_i=M} \\ & + \beta_4 * NumPosts_{i,t} + \beta_5 * NumPosts_{i,t} * \mathbf{1}_{g_i=M} + u_i + e_{i,t}, \end{aligned} \tag{29}$$

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 (29), 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.

B.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.[24]

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 12), we regress reputation points earned per answer post ($\mathit{\Delta}$Rep) 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.

[24] 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 12. 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.

For the evaluation of novice questions (Table 13), 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 content, given the reputation they eventually earn. 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 13. Evaluation of Questions Posted by Novice Users: ΔRep

	All	Experiment Window	Reach 100
	(1)	(2)	(3)
Male	0.58	0.30	2.52
	(.11)	(.14)	(1.37)
Constant	7.92	5.05	17.35
	(.10)	(.12)	(1.20)
# Obs	72,896	26,092	2,123

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 14). 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 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 14. 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.

C Stereotyping

In Section 2, 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.

Bordalo et al. (2016b) develop a framework in which biased stereotypes arise and persist due to 'representativeness', a well-documented cognitive heuristic used to simplify complex probability judgments (Tversky and Kahneman 1983). 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.[25]

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, Bordalo et al. (2016b) define the stereotyped belief as

$$\pi^{st}_{t,g} \equiv \pi_{t,g} \frac{\left(\frac{\pi_{t,g}}{\pi_{t,-g}}\right)^{\theta}}{\sum_{s \in T} \pi_{s,g} \left(\frac{\pi_{s,g}}{\pi_{s,-g}}\right)^{\theta}}, \tag{30}$$

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

[25] 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.

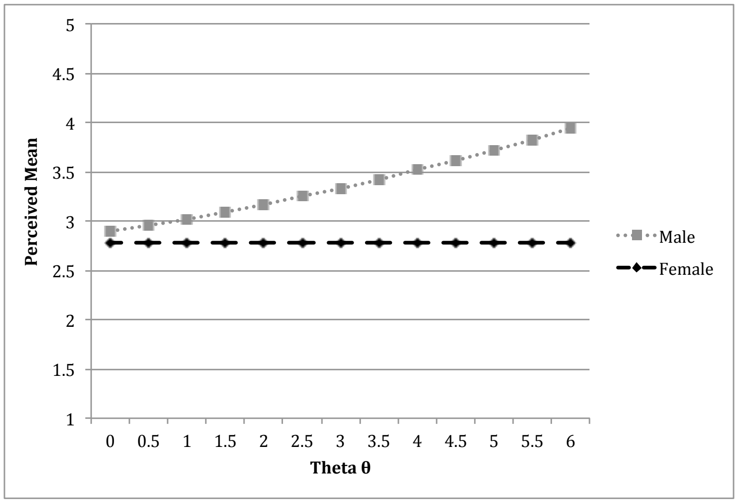


Figure 7. Subjective average ability by gender $\hat{\mu}_g$ as function of θ.

type, but the overall type distributions are largely the same. This is consistent with recent empirical work that finds support for the model (Arnold et al. 2017; Bordalo, Coffman, Gennaioli, and Shleifer 2016a; Coffman 2014).

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 7 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 (Arnold et al. 2017)), the difference in perceived means triples to nearly half a quintile. As shown in Section 2, if even a small proportion of individuals hold such distorted beliefs, this can lead to a dynamic reversal of discrimination.