The Dynamics of Discrimination: Theory and Evidence[*]

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

November 2017

Abstract

We model the dynamics of discrimination and show how its evolution can identify the underlying cause. 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 history of evaluations. With no prior evaluations, women face significant discrimination, while following a sequence of positive evaluations, the direction of discrimination *reverses*: posts by women are favored over those by men. According to our theoretical predictions, 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, Sally Sadoff, Lise Vesterlund, Leeat Yariv, and seminar and conference participants at Carnegie Mellon, ESSET Gerzensee, Harvard, Northwestern, Princeton, Stanford SITE, University of Chicago, 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. The project received IRB approval at Carnegie Mellon and University of Pennsylvania. The experiment was pre-registered in the AEA RCT Registry (#AEARCTR-0000950).

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

[‡] Carnegie Mellon University and University of Chicago Booth School of Business

[§] Wayfair Inc.

1 Introduction

A man and a woman with similar qualifications complete a task and produce output that generates similar signals of quality. Discrimination against women occurs when the man receives a more positive evaluation or reward for his output than the woman. A rich literature documents such discrimination in a wide range of contexts.[1] These empirical studies mostly focus 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 of performance in similar contexts. As prior work has noted, it is difficult to identify the underlying cause of discrimination from such static settings, as different causes generate the same predictions (Fang and Moro 2011).[2] In this paper, we develop a theoretical framework to show that the dynamics of discrimination can be used to identify its underlying cause, and test these theoretical predictions in a field experiment on a large online platform.

Suppose individuals repeatedly perform tasks to generate output, and in the process, produce an observable history of evaluations. For example, a man and a woman generate computer code on an online platform such as GitHub, and each has a publicly observable reputation derived from prior evaluations of his or her code – in particular, whether previous submissions were accepted or rejected. When both are starting out and lack prior evaluations, initial discrimination occurs if the woman's code is less likely to be accepted than the man's, despite the appearance of similar quality. Suppose the programmers continue producing code, and receive similar sequences of evaluations. Does discrimination persist in this dynamic setting, is it mitigated, or does it even *reverse*? We demonstrate that the answer to this question depends critically on the underlying cause of discrimination.

If the cause is belief-based – for example, the quality of code is imperfectly observable and evaluators believe that men on average have higher programming ability than women – then observing prior evaluations will reduce discrimination against women, relative to men with similar evaluations. This dynamic effect operates through two

[1] For example, discrimination has been documented in hiring (Riach and Rich 2006), housing (Ewens, Tomlin, and Wang 2014), and service markets (Gneezy, List, and Price 2012), and based on group characteristics such as race (Bertrand and Mullainathan 2004), ethnicity (Fershtman and Gneezy 2001) and gender (Milkman, Akinola, and Chugh 2012). See Bertrand and Duflo (2016) for review.

[2] One exception is Knowles, Persico, and Todd (2001). Their method is analogous to evaluating a worker based on a signal of quality, observing the true quality, and then measuring whether men and women who receive the same evaluation have the same average quality. In many settings, including the one we consider here, true quality is not observed.

channels. First, prior evaluations provide signals of a worker's ability, which reduces the impact of group statistics on how the worker's subsequent output is evaluated.[3] As a result, observing prior evaluations will mitigate discrimination against a woman's output, relative to a man who has similar prior evaluations. Second, and novel to our theoretical framework, the informational content of these signals about ability endogenously depends on the equilibrium behavior of prior evaluators. When initial beliefs favor men, a woman needs to produce *higher* quality output in order to overcome the initial disparity in beliefs and receive a similar evaluation as a man. This will speed up the mitigation of discrimination for evaluators who are aware that a woman had to meet a higher standard to receive a given evaluation. These evaluators may even come to believe that the woman is of higher ability than a man with a similar history of evaluations, and favor her future code over the man's – *reversing* the direction of discrimination in later periods. In fact, observing a reversal can disentangle whether evaluators' models are correct or biased – we show theoretically that a reversal is indicative of bias. In contrast to belief-based causes, if discrimination is caused by a taste or preference against women (Becker 1957), then a woman who receives a similar sequence of evaluations to a man will continue to face discrimination in future periods.

Motivated by these dynamic insights, we run a field experiment on a large online platform to empirically study how discrimination dynamically evolves. We find that initial discrimination against women reverses at later stages, providing evidence that discrimination is belief-based, and that evaluators have at least some level of bias. Our setting is unique in allowing us to exogenously vary the perceived gender and publicly observable evaluation history (reputation) of the poster. While output from women with no reputation was significantly less likely to be rewarded, relative to similar output from men with no reputation, output from high reputation women was *favored* over similar output from high reputation men.

These results 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

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

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. As discussed further in Section 4, our dynamic framework helps organize seemingly contradictory results on discrimination in static settings. For example, Milkman et al. (2012) documents discrimination against women in many academic settings, while Williams and Ceci (2015) finds that female academics are favored over their male colleagues. However, the studies were conducted at different stages of the academic process – students in the former case, and accomplished professors in the latter. Far from being contradictory, discrimination in favor of accomplished female professors may actually be a function of discrimination *against* women earlier in the pipeline.

Our theory formalizes the relationship between the dynamic pattern of discrimination, which is based on observable evaluations, and the underlying causes of discrimination, which are unobservable and depend on the primitives of preferences and beliefs. We define discrimination as the difference between the evaluations of output for men and women, conditional on having similar evaluation histories and current signals of quality. We focus on three potential causes of discrimination: preference-based, belief-based with correct beliefs, and belief-based with incorrect, biased beliefs. The first cause has typically been referred to as *taste-based* discrimination, where evaluators have a preference against rewarding or interacting with women (Becker 1957). Belief-based causes are driven by differences in beliefs about the average distribution of ability between men and women. The case typically considered in the literature is fully rational, *statistical* discrimination, where evaluators are partial towards men based on correct beliefs about the underlying distributions (Phelps 1972).[4] However, discrimination can also be driven by incorrect, biased beliefs.[5] As we discuss later in the paper, distinguishing between these underlying causes has significant implications for both policy and welfare.

While results from static settings establish the existence of discrimination, as dis-

[4] The theoretical literature on belief-based causes of discrimination has largely focused on the correctly specified case (Fang and Moro 2011), with theoretically-motivated empirical work following suit, providing evidence for belief-based mechanisms that are statistical in nature (Altonji and Pierret 2001; Knowles et al. 2001).

[5] Recent research has shown that incorrect beliefs can arise and persist due to systematic biases in judgment, such as individuals forming stereotypes that overweigh representative traits of a particular group (Bordalo, Coffman, Gennaioli, and Shleifer 2016b).

cussed above, it is often difficult to use such data to infer the cause (Fang and Moro 2011). In fact, we show that for every set of beliefs that lead to discrimination against women in a given period, there exist preferences that also lead to the same level of discrimination. We then demonstrate that the underlying causes make contrasting predictions across periods: depending on its cause, discrimination against women can persist, mitigate, or reverse in response to observing prior evaluations of output. We derive an impossibility result: if discrimination is statistical – based on common knowledge of correct beliefs – then observing women and men with similar evaluations will mitigate discrimination, but will never lead to a reversal. This result implies that observing a dynamic reversal is a 'smoking gun' for belief-based discrimination with *bias*, since it also rules out standard preference-based causes.

We also illustrate that the amount of subjectivity involved in judging the quality of output – modeled as the variance in signals of quality – can be used to further identify the cause of discrimination. Prior work in social psychology has shown that discrimination is exacerbated by subjectivity in judgment (Fiske, Bersoff, Borgida, Deaux, and Heilman 1991). Motivated by this insight, we show theoretically that decreasing the level of subjectivity in judgment will mitigate discrimination driven by beliefs (either correct or biased), as prior beliefs play a smaller role in assessing quality when there is less uncertainty, but will not affect discrimination driven by preferences, which will persist even when quality is perfectly observable.

We study these theoretical predictions in a field experiment on a large online Q&A forum. Users post mathematics questions or answers, which are evaluated – voted up or down – by other users on the site. A user's reputation provides a summary statistic of prior evaluations of his or her past posts: higher reputation corresponds to a greater number of positive votes on the user's posted content. Since reputation is endogenously generated by evaluations of previous posts, interpreting reputation requires a model of other users' beliefs and decision processes. Importantly, reputation is publicly observable and valuable. Both the username and the level of reputation are prominently displayed adjacent to any question or answer post. Reputation unlocks privileges and can be used as currency to *pay* other users for providing answers. The family of Q&A forums that comprise our setting has over 3 million questions asked and 4 million answers posted per year. The forum has nearly 350,000 users and is a prominent resource for students and researchers in STEM fields, which makes documenting the existence and source of gender discrimination in this setting particularly important.

In our experiment, we posted original mathematics questions on created accounts that exogenously vary in the gender of the username. Our setting is particularly well-suited for exploring the dynamics of discrimination because we are also able to exogenously vary the evaluation histories of the users, as captured by their publicly observable reputations.[6] To exogenously vary reputation, half of the questions were posted on accounts that did not have prior evaluations. We built the reputations of the other accounts by posting content until their reputations reached the top 25^{th} percentile on the forum. To avoid endogeneity issues and ensure that the underlying informational content of reputation is the same for both genders, we randomly reassigned the gender of the username after the account reached a high reputation.

Discrimination is measured as the differential number of positive votes for posts by male versus female usernames at either low or high reputation levels. We find that females face significant initial discrimination on the platform: questions posted by female usernames with no prior reputations are evaluated less favorably – they receive fewer positive votes – than questions posted by male usernames with no reputations. However, at high reputations, the direction of discrimination *reverses*: questions posted by high reputation females receive more positive votes than those posted by high reputation males. This is consistent with belief-based discrimination with bias.

Motivated by research in social psychology on stereotyping (Fiske et al. 1991), we also explored whether the amount of subjectivity involved in judging posts affects the level of 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 discrimination is preference-based, this distinction should not matter: similar levels of initial discrimination will be observed for both question and answer posts. In contrast, if discrimination is belief-based, then reducing uncertainty over the standards by which a post is judged will mitigate it. We find support for the latter prediction: answers posted by females with no prior reputations received a similar number of positive votes as answers posted by males with no reputations. Directly comparing questions and answers posted by low reputation accounts produces a significant interaction, with initial discrimination against females on questions, but

[6] Extant evidence for discrimination reversals in dynamic settings is prone to multiple explanations, such as selection and institutional factors. For example, Booth, Francesconi, and Frank (1999); Groot and van den Brink (1996) find discrimination against women at the initial hiring stage for promotable jobs, but conditional on being hired, women are more likely to be promoted. However, this reversal could be explained by unobservables, such as gender-based hiring quotas for senior positions.

no discrimination between males and females on answers. This is consistent with belief-based but not preference-based discrimination.

In addition to our experimental results, we exploit two additional data sources. First, we obtained a private dataset from the forum that contains additional information about the users evaluating the content posted in our experiment. This allows us to run additional robustness tests, and provides further evidence for some of our assumptions. As a complement to the experimental results, we also obtained a large observational dataset from the forum. We ran an algorithm to infer gender from usernames, and conducted a similar analyses as in the experiment. We found analogous patterns of discrimination in the observational data, documenting both the dynamic reversal for questions and a lack of differential evaluations by gender for answers.

We also use the observational data to explore one form of bias that may drive the documented reversal in discrimination. We calculated distributions of evaluations on answer posts by gender. Since we do not find evidence of discrimination on answers, we use these distributions as proxies of underlying ability and show that the observed distributions are quite similar for male and female users, with only a slight difference in the means. We then use the framework of Bordalo et al. (2016b) to show that biased probability judgments will generate 'stereotypes' that significantly exaggerate this small difference. If evaluators use the calculated distributions to form beliefs about underlying ability for each gender, and some evaluators are prone to biased probability judgments, this will lead to a significant divergence in their beliefs about the ability of men and women. Our theoretical analysis shows that when some individuals hold such biased stereotypes, this can lead to the type of reversal we observe in the data.

Our results have significant implications for policy and welfare analysis. When prior evaluations are observable, the timing of potential interventions to reduce discrimination will have significant consequences for discrimination at different stages. For example, interventions that exogenously lower the threshold for a target group to receive a given evaluation (e.g. college admittance) may exacerbate discrimination down the road. If future evaluators are aware of the lower threshold for the target group, then they will interpret prior positive evaluations as less informative for members of that group. This can result in greater subsequent discrimination against the exact group that the intervention aimed to help (e.g. lower returns to higher education).

Our findings 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 implication of discrimination caused by biased beliefs is straightforward – it is inefficient. Even if a discrimination reversal occurs, so that women eventually receive higher evaluations than men with similar *evaluation* histories, these women are still receiving lower evaluations than men with similar *signal* histories. Therefore, the reversal does not offset initial discrimination. A woman who is favored over a man with similar prior evaluations should receive an even higher evaluation than she does, relative to unbiased beliefs about her expected ability. Further, women may inefficiently select out of the process at earlier stages than men with similar abilities due to initial discrimination.

The rest of the paper proceeds as follows. Section 2 develops a theoretical model to show how the dynamics of discrimination can be used to identify its source. Section 3 describes the experiment and presents the results from both the experiment and the observational data. Section 4 discusses how our findings can organize some of the existing discrimination literature and proposes implications for policy design. All proofs not presented in the body of the paper are in Appendix A.

2 A Dynamic Model of Discrimination

We develop a dynamic model of discrimination in which evaluators learn about a worker's ability from the worker's group identity and past performance, and use this information to evaluate the quality of the worker's output. We have chosen for convenience to use gender discrimination of M(ales) against F(emales) in our model, since this is the type of discrimination we study in the experiment.

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 evaluate the worker's performance. For simplicity, assume that there is one evaluator per task, who reports evaluation $v_t \in \mathbb{R}$. Before evaluating task t, the evaluator observes the worker's gender g and publicly observable evaluations on past tasks, where $h_1 = \emptyset$ and $h_t = (v_1, ..., v_{t-1})$ for $t > 1$. The evaluator also observes a signal $s_t = q_t + \eta_t$ of the 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 precision allows for greater uncertainty in the underlying quality, conditional on the signal. The level of 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. An evaluation strategy is a mapping from the gender g, history h, and signal s to evaluation $v(h, s, g)$.

An evaluator's payoff depends on the quality of the task and the gender of the worker. She receives a payoff of $-(v - (q - c_g))^2$ from reporting evaluation v on a task of quality q from a worker of gender g, where c_g is a taste parameter. Normalize $c_M = 0$.

An evaluator is *partial* towards one gender if she favors this gender, either directly through preferences, which we refer to as *preference-based partiality*, or indirectly through her belief about the distributions of ability, which we refer to as *belief-based partiality.* In the first case, an evaluator may have a 'taste' for male workers, meaning that there is a disamenity value associated with tasks produced by female workers. Therefore, in order to receive the same evaluation as male workers, female workers have to compensate by producing higher quality output.

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

In the belief-based case, an evaluator may perceive that the population of male workers has a more favorable distribution of ability than the population of female workers. This perception can be biased or unbiased, based on whether it coincides with the true population distribution of ability. We assume that evaluators believe ability is normally distributed with mean $\hat{\mu}_g$ and precision τ_a, which coincides with the true precision.

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

Let $\hat{\mu}_g(h)$ denote the perceived mean ability of a worker with gender g following history h. A *belief-reversal* occurs at history h if an evaluator has belief-based partiality towards men, but conditional on observing history h for both a man and a woman, believes that the woman is of higher average ability than the man, $\hat{\mu}_F(h) > \hat{\mu}_M(h)$.

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 contrast to partiality, which is a property of the primitives of the model (preferences, beliefs), discrimination is a property of behavior. In our framework, gender discrimination occurs when a male and female worker with the same signal and history receive different evaluations. Let

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

denote the difference between a male's and female's evaluations at (h, s).

Definition 3 (Discrimination). *A woman (man) faces* discrimination *at* (h, s) *if* $D(h, s) > 0$ $(D(h, s) < 0)$.

We say discrimination decreases (i.e. between histories or across parameters) if the absolute value of the discrimination measure, $|D(h, s)|$, decreases. A discrimination *reversal* occurs if there exist histories $h \subset h'$ and signal s such that women face discrimination at (h, s) and men face discrimination at (h', s).

Heterogenous Evaluators. The above definitions of partiality and discrimination apply to an individual evaluator, or to a set of homogenous evaluators. Our framework can also allow for heterogeneous evaluators. Suppose that each type of evaluator is characterized by a tuple $\theta = (\hat{\mu}_F^\theta, \hat{\mu}_M^\theta, c_F^\theta, \hat{\pi}^\theta)$, which specifies initial beliefs about mean ability for males and females, a taste parameter for females, and a belief about the type distribution of other evaluators, $\hat{\pi}^\theta \in \Delta(\Theta)$, where Θ denotes the set of evaluator types, which we assume to be finite. Let $\pi \in \Delta(\Theta)$ denote the true measure over evaluator types.

This type framework can capture a variety of settings with heterogeneity. For example if there is initial uncertainty about the true population means, and evaluators are aware of this uncertainty, then the initial beliefs about population average ability $\hat{\mu}_F^\theta, \hat{\mu}_M^\theta$ differ by type, and all evaluators have a correct belief about the type distribution, $\hat{\pi}^\theta = \pi$ for all θ. Alternatively, if some evaluators are misspecified in how they believe other individuals evaluate workers, these types have a perceived belief about the type distribution that differs from the true distribution, $\hat{\pi}^\theta \neq \pi$. For example, a type θ_B who uses a heuristic to form beliefs may not be aware of his bias, and believe that other evaluators form beliefs in a similar manner, $\hat{\pi}^B(\theta_B) = 1$ (see the false consensus effect (Ross, Greene, and House 1977)). When other types of evaluators do not use this biased heuristic, the true frequency of biased types is $\pi(\theta_B) < 1$, and therefore, $\hat{\pi}^B \neq \pi$.

It is straightforward to define *aggregate* analogues for beliefs, partiality, and discrimination. There is *aggregate preference-based partiality* towards men if $E_\pi[c_F] > 0$, and *aggregate belief-based partiality* towards men if $E_\pi[\hat{\mu}_M] > E_\pi[\hat{\mu}_F]$, where the expectation is with respect to the true distribution over types. Aggregate belief-based partiality is *unbiased* if $E_\pi[\hat{\mu}_M] = \mu_M$ and $E_\pi[\hat{\mu}_F] = \mu_F$, and otherwise is *biased.* A woman faces *aggregate discrimination* at (h, s) if $E_\pi[D(h, s)] > 0$. It is possible for individual types to exhibit partiality, bias and/or discrimination, but for aggregate preferences, beliefs and behavior to be impartial, unbiased, and/or non-discriminatory.[7]

2.2 Discussion of Model

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 – (i) whether group differences are exogenous (Phelps 1972) and discrimination is due to imperfect information, or (ii) whether group differences are "self-fulfilling" and discrimination is an equilibrium effect (Arrow 1973).[8] 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). 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 different skill acquisition equilibria (Coate and Loury 1993; Fryer 2007). For example, women may not acquire costly skills because they believe that they will not be offered high paying jobs, while evaluators may not offer women high paying jobs because they think women do not acquire skills. In these models, there are also always equilibria in which both men and women acquire skills, and evaluators treat them identically.

In addition to these two classes of models, belief-based discrimination can also

[7] For example, suppose each type's initial belief about mean 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.

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

arise from systematically incorrect, or *biased*, beliefs. As in case (i), discrimination arises from imperfect information about ability. But this discrimination is due to evaluators' *misspecified* beliefs about group differences in the distribution of ability, rather than true group differences.[9] The discrimination literature has tended to classify such discrimination as taste-based.[10] However, we demonstrate that biased beliefs lead to discrimination patterns – how discrimination dynamically evolves, or how it changes with respect to underlying parameters – that substantially differ from those that arise in taste-based models with animus (i.e. preference-based partiality). This partly drives our motivation to distinguish between discrimination due to misperception of information versus underlying preferences.

Subjectivity of Judgment. Subjectivity in judgment – defined as uncertainty over assessment criteria – increases the variance of potential evaluations (Olson, Ellis, and Zanna 1983) and reduces the expected consensus between evaluators (Kelley 1973). A rich literature in social psychology has argued 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). In contrast, with objective judgment, the available information provides more precise information about the underlying attribute. This decreases the potential for belief-based discrimination and consensus is expected to occur.

For example, Fiske et al. (1991) discuss evaluations of counting tasks as an example of objective judgment, and assessments of competency as an example of subjective judgment. For counting tasks, the evaluation criterion is clear – did the individual report the correct number – and consensus is likely to be achieved. The latter involves uncertainty as to what information is most likely to be informative of competency: is it the individual's grades, work experience, or writing style? Even after observing several signals regarding the relevant attribute, evaluators are still likely to have significant residual uncertainty about competency. This increases the reliance on beliefs about

[9] One microfoundation for how biased beliefs about group differences may arise is a model where people form stereotypes about group differences that exaggerate empirical reality (Bordalo et al. 2016b). In our setting, stereotyping corresponds to distortions in the perceived mean ability, $\hat{\mu}_g$.

[10] 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. Discrimination is automatic, rather than deliberative (Bertrand, Chugh, and Mullainathan 2005; G. Greenwald, E. McGhee, and L. K. Schwartz 1998).

group statistics, e.g. average competency by gender, to inform judgment. Indeed, researchers have documented greater reliance on beliefs about group statistics when judgment is more subjective (see Fiske and Taylor 1991, for review).

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. In line with the literature 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.

In the following sections, we explore how the different forms of partiality impact the evaluation of workers. We use these insights to illustrate how aggregate choice behavior (i.e. evaluations), which is observable, can be used to identify the source of discrimination (i.e. the type of partiality), which is based on the primitives of the underlying model.

2.3 Discrimination with Belief-Based Partiality

First, we characterize the initial discrimination female workers face when evaluators have belief-based partiality, show how this initial discrimination varies with the precision of the signal, and characterize the dynamic evolution of discrimination. Throughout this section, assume $c_F = c_M = 0$.

2.3.1 Initial Discrimination

The evaluation of the first task depends on the evaluator's preferences and prior belief about ability, but does not depend on beliefs about other evaluators. Consider the evaluation of an initial task from a worker of gender g by an evaluator who has prior beliefs about the distribution of ability $a \sim N(\hat{\mu}_g, 1/\tau_a)$. Then the perceived distribution of quality is also normal, $q_1 \sim N(\hat{\mu}_g, 1/\tau_q)$, where $\tau_q \equiv \tau_a\tau_\varepsilon/(\tau_a + \tau_\varepsilon)$. The evaluator combines the perceived distribution of quality with the observed signal s_1, which has distribution $s_1|q_1 \sim N(q_1, 1/\tau_\eta)$, and uses Bayes rule to form the posterior belief about quality

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

The evaluator maximizes her expected payoff by choosing

$$v(h_1, s_1, g) \quad = \quad E[q_1|h_1, s_1, g] = \frac{\tau_q \hat{\mu}_g + \tau_\eta s_1}{\tau_q + \tau_\eta}. \tag{2}$$

Note that $E[q_1|h_1, s_1, g]$ is strictly increasing in s_1 and $\hat{\mu}_g$ – higher signals and higher expected ability result in higher evaluations. Recall that discrimination is measured as the difference between a male's and female's evaluations. Therefore, initial discrimination is independent of the signal and equal to

$$D(h_1, s_1) = \left(\frac{\tau_q}{\tau_q + \tau_\eta}\right)(\hat{\mu}_M - \hat{\mu}_F). \tag{3}$$

Proposition 1. *Initial discrimination against females arises if and only if the evaluator has belief-based partiality, $\hat{\mu}_F < \hat{\mu}_M$.*

Proof. This follows immediately from $D(h_1, s_1) > 0$ if and only if $\hat{\mu}_M > \hat{\mu}_F$. □

Precision of Signal. As the signal provides more precise information about quality, the evaluator's belief about the worker's underlying ability plays a smaller role in the evaluation. Therefore, belief-based partiality has a smaller impact on evaluations, and there is less discrimination, the larger the precision of the signal. In the limit, when quality is perfectly observable, differences in beliefs about ability do not translate into discriminatory evaluations of quality. Although an evaluator with belief-based partiality expects lower quality from women ex-ante, conditional on observing a signal, the evaluator has very precise information about the quality of the current task. Therefore, men and women who generate the same signal receive identical evaluations.

Proposition 2. *If the evaluator has belief-based partiality, then discrimination is decreasing in the precision τ_η of the signal. If quality is observable ($\tau_\eta = \infty$), there is no discrimination.*

Proof. From (3), it is clear that $|D(h_1, s_1)|$ is decreasing in τ_η iff $\hat{\mu}_M \neq \hat{\mu}_F$, and $\lim_{\tau_\eta \to \infty} D(h_1, s_1) = 0$. □

When evaluators are heterogenous, analogues to Propositions 1 and 2 immediately follow for aggregate discrimination, where

$$E_\pi[D(h_1, s_1)] = \left(\frac{\tau_q}{\tau_q + \tau_\eta}\right) E_\pi[\hat{\mu}_M - \hat{\mu}_F].$$

2.3.2 Dynamics of Discrimination

Next, we study how discrimination evolves across subsequent rounds of evaluation. After the first round, the evaluator has access to an additional source of information about the worker: past evaluations. These past evaluations provide information about the worker's ability, which improves the estimate of current quality. In order to interpret prior evaluations, an evaluator needs a model of other evaluators' beliefs about the distribution of ability. We consider two cases: (i) a model in which all evaluators have the same beliefs about the distributions of ability, and (correctly) believe that all other evaluators have the same beliefs as they do; (ii) a model with misspecification, in which some evaluators hold biased stereotype beliefs about the distributions of ability, and others hold correct beliefs but are aware of the presence of the biased evaluators. Note that the *correctly specified model*, in which evaluators have unbiased beliefs about the distributions of ability and a correct model of how other evaluators behave, is a special case of (i). We show that the dynamic predictions of these two cases – in particular, whether the direction of discrimination can reverse – separates whether evaluators have biased or correctly specified beliefs.

Case (i): Impossibility of Reversal with Correct Beliefs. Suppose that all evaluators have the same prior beliefs about the distributions of ability, a correct model of the beliefs of other evaluators, and belief-based partiality. Formally, there is a single type $\theta = \{\hat{\mu}_F, \hat{\mu}_M, 0, \pi\}$, where $\hat{\mu}_F < \hat{\mu}_M$ and $\pi(\theta) = 1$ is the true type distribution.

In the first period, a female is subjected to stricter standards than a male. In order to receive the same evaluation as a male, she must produce a higher signal to offset the lower belief about her ability. From (2), let

$$s_1^g(v_1) \equiv \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) v_1 - \left(\frac{\tau_q}{\tau_\eta}\right) \hat{\mu}_g$$

denote the signal required by gender g to receive evaluation v_1. Then $s_1^F(v_1) > s_1^M(v_1)$, i.e. a given evaluation is indicative of a higher signal of a female's ability than a male's. This moves the posterior distribution of the female's ability closer to that of a male who receives the *same* evaluation, reducing discrimination in the next period. However, the higher prior belief about average ability for the male still maps into a higher posterior belief about average ability, despite the more informative signal from the female. Therefore, although discrimination is mitigated, the beliefs about average ability do not reverse, and hence, discrimination does not reverse. The analysis in subsequent periods is analogous: the perceived average ability of men and women continues to move closer together following similar evaluation histories, but does not reverse. Therefore, discrimination continues to decrease, but does not reverse.

Proposition 3. *Suppose 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 discrimination decreases across periods, following similar evaluation histories, but never reverses.*

Proposition 3 establishes the impossibility of a reversal when there is a single type of evaluator who has a correct model of other evaluators. The correctly specified model, in which evaluators also have correct beliefs about the population distributions of ability for men and women, is a special case of this model. Therefore, reversals do not occur in the correctly specified model, and observing a reversal is a smoking-gun for some form of misspecification – either in perceived average ability, perceived behavior of other evaluators, or both.

The intuition for Proposition 3 is as follows. The family of normal distributions satisfies the monotone likelihood ratio property (MLRP) in the mean. For a fixed signal, the MLRP is preserved under Bayesian updating, and the mean of the posterior distribution of ability is increasing in the mean of the prior distribution of ability $\hat{\mu}$. But our comparison is between the posterior distributions of ability for a male and a female who receive the same *evaluation*, not the same *signal.* An evaluation v_1 implies a higher signal for the female. Therefore, the informativeness of v_1 is endogenously determined by $\hat{\mu}$ – in fact, it moves in exactly the opposite direction of the prior belief, as the family of distributions of v_1 indexed by $\hat{\mu}$ are monotone decreasing in the likelihood ratio order. Therefore, $\hat{\mu}$ impacts the mean of the posterior distribution of ability through two channels: (i) the prior distribution of ability is MLRP increasing in $\hat{\mu}$; and (ii) the informativeness of an evaluation is MLRP decreasing in $\hat{\mu}$. The proof of Proposition 3 lies in establishing that the first effect dominates, and therefore, the posterior distribution of ability is also MLRP increasing in $\hat{\mu}$.

Case (ii): Possibility of Reversal with Misspecification. In order to demonstrate that discrimination reversals can arise with misspecified beliefs, we explore one potential model that leads to a reversal. This is a possibility result, in the sense that it demonstrates a reversal is possible with misspecification; we do not claim that this is the only type of misspecification that can produce a reversal.

Suppose there are two types of evaluators, one of whom has misspecified beliefs in the form of biased stereotypes. With probability $p \in (0,1)$, an evaluator is a type θ^B with belief-based partiality, $\hat{\mu}_F^B < \hat{\mu}_M^B \equiv \hat{\mu}$ for some $\hat{\mu} \in \mathbb{R}$, and with probability $1-p$, an evaluator is an impartial type θ^I with beliefs $\hat{\mu}_F^I = \hat{\mu}_M^I = \hat{\mu}$. In other words, both types have the same prior belief about male ability, and type θ^B believes that females have lower average ability than males. Type θ^B is naive, in the sense that she believes that all other evaluators have the same belief about the distributions of ability as herself, $\hat{\pi}^B(\theta^B) = 1$. Type θ^I is aware that some evaluators have different beliefs about the distributions of ability – she has a correctly specified model of evaluator types, $\hat{\pi}^I(\theta^B) = p$ and $\hat{\pi}^I(\theta^I) = 1-p$. The literature on heuristics and biases provides a foundation for such a model. Suppose some evaluators use the 'representativeness" heuristic to form beliefs about the population distribution of ability, i.e. steoreotyping as in the framework of (Bordalo et al. 2016b), and is not aware of this cognitive bias, while the impartial types have accurate beliefs about the population distribution of ability, and are aware that a subset of evaluators stereotype.

In the first round, the stereotype evaluators discriminate against females, while the impartial types evaluate females and males in exactly the same manner. Therefore, initial aggregate discrimination is positive and equal to

$$D(h_1, s_1) = \left(\frac{\tau_q}{\tau_q + \tau_\eta}\right) p(\hat{\mu} - \hat{\mu}_F^B) \tag{4}$$

for all s_1. In subsequent rounds, the biased evaluators behave in the same way as evaluators in the single-type model with the same beliefs. Their discrimination decreases across periods, but does not reverse. In contrast, the impartial evaluators' beliefs immediately favor females: they are aware that with some probability, females faced discrimination in the first round. Therefore, conditional on receiving the same evaluation as a male, these females received a higher signal in expectation. Since the impartial types' prior beliefs about ability are identical for males and females, this pushes their posterior belief about the average ability of females above that of males. Following evaluation v_1, let $\hat{\mu}_F^\theta(v_1)$ denote an evaluator of type θ's belief about average ability for a female worker, and $\hat{\mu}(v_1)$ denote the belief about the average ability for a male worker (which is the same for both types). Then discrimination in period 2, given $h_2 = \{v_1\}$, is equal to

$$D(v_1, s_2) = \left(\frac{\tau_{q,2}}{\tau_{q,2} + \tau_\eta}\right) [\hat{\mu}(v_1) - p\hat{\mu}_F^B(v_1) + (1-p)\hat{\mu}_F^I(v_1)], \tag{5}$$

where $\hat{\mu}_F^B(v_1) < \hat{\mu}(v_1) < \hat{\mu}_F^I(v_1)$ and $\tau_{q,2} \equiv (\tau_a + \tau_{\varepsilon\eta})\tau_\varepsilon/(\tau_a + \tau_{\varepsilon\eta} + \tau_\varepsilon)$ is the precision of quality in period 2, given $\tau_{\varepsilon\eta} \equiv \tau_\eta\tau_\varepsilon/(\tau_\eta + \tau_\varepsilon)$.[11] A discrimination reversal will occur iff the impartial type's favorable beliefs towards females reverses the aggregate belief in favor of females, i.e. $p\hat{\mu}_F^B(v_1) - (1-p)\hat{\mu}_F^I(v_1) > \hat{\mu}(v_1)$. Proposition 4 establishes that indeed, given any initial beliefs and any initial evaluation, reversals are possible when some evaluators have a misspecified model.

Proposition 4. *Suppose evaluators are type $\theta^B = \{\hat{\mu}_F^B, \hat{\mu}, 0, \delta_{\theta^B}\}$ with probability $p \in (0,1)$, and type $\theta^I = \{\hat{\mu}, \hat{\mu}, 0, \pi\}$ with probability $1-p$, where $\hat{\mu}_F^B < \hat{\mu}$, δ_θ is the dirac measure on type θ, and π is the true measure over types.*

1. *For any initial evaluation v_1, there exists a $\overline{p} \in (0,1)$ and $\overline{s} \in \mathbb{R}$ such that for frequency of stereotype evaluators $p \in (0, \overline{p})$ and signals $s_2 > \overline{s}$, aggregate discrimination reverses in period 2.*

2. *For any signal s_2, there exists a $\overline{p}' \in (0,1)$ and $\overline{v} \in \mathbb{R}$ such that for frequency of stereotype evaluators $p \in (0, \overline{p}')$ and initial evaluations $v_1 < \overline{v}$, aggregate discrimination reverses in period 2.*

The intuition is as follows. Increasing the prevalence of biased evaluators has two effects on discrimination in the second period. First, conditional on the same signal histories, it increases the magnitude of the difference in the posterior belief of average ability between a male and a female for the impartial type. More biased evaluators means that it is more likely the female faced initial discrimination, and therefore, for any initial evaluation, her expected signal increases with p. Second, increasing p increases the probability that the second period evaluator has belief-based partiality. Since biased evaluators still discriminate against females in $t = 2$, it is more likely that this 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

[11] Given a normal prior belief about ability and a normal likelihood function, from Bayes rule, the precision of the posterior belief about ability in period 2 is $\tau_{a,2} \equiv \tau_a + \tau_{\varepsilon\eta}$. Therefore, the precision of the posterior belief about quality in period 2 is $\tau_{q,2} \equiv (\tau_{a,2} + \tau_{\varepsilon\eta})\tau_\varepsilon/(\tau_{a,2} + \tau_{\varepsilon\eta} + \tau_\varepsilon)$. See Lemma 1 in Appendix A for the complete derivation.

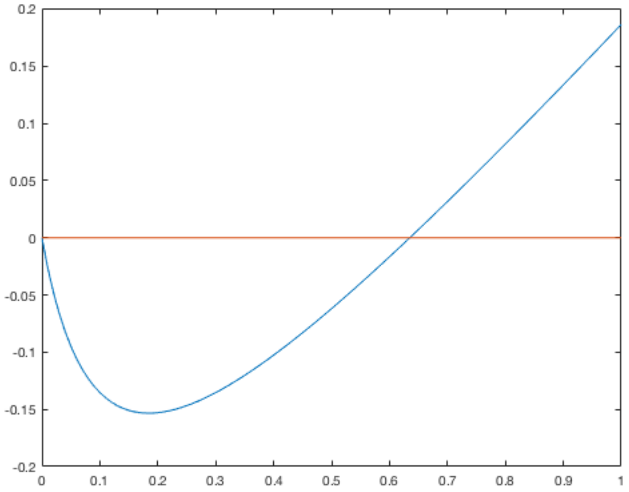

Figure 1. Discrimination in second period, as a function of p (positive = discrimination against females; negative = against males)

second period discrimination changes with respect to p – discrimination starts at zero for $p = 0$, decreases in p following some signals, which leads to a reversal in period two following these signals, and finally increases in p. Discrimination is always positive as p approaches one, as the stereotype model approaches the model with common knowledge of a single type of evaluator. Figure 1 illustrates this reversal: for p between 0 and 0.62, discrimination reverses in the second period.

2.4 Discrimination with Preference-Based Partiality

Next, we consider the implications of preference-based partiality, with the goal of determining what patterns of discrimination can be used to distinguish it from belief-based partiality.

2.4.1 Initial Discrimination

Consider an evaluator with preference-based partiality and no belief-based partiality, $c_F > 0$ and $\hat{\mu}_F = \hat{\mu}_M \equiv \hat{\mu}$ for some $\hat{\mu} \in \mathbb{R}$. As in (1), the posterior distribution of quality, conditional on s_1, is normal. The evaluator maximizes her expected payoff by choosing

$$v(h_1, s_1, g) = \frac{\tau_q \hat{\mu} + \tau_\eta s_1}{\tau_q + \tau_\eta} - c_g. \tag{6}$$

Therefore, initial discrimination is equal to the preference parameter, $D(h_1, s_1) = c_F$, and initial discrimination occurs if and only if $c_F > 0$.

It is not possible to identify the source of discrimination from a single round of evaluations. Both preference-based and belief-based partiality lead to discrimination in the first period. Further, for any level of preference-based partiality, there exist prior beliefs about ability that lead to equivalent evaluation behavior and discrimination.

Proposition 5. *For any level of preference-based partiality, there exists a level of belief-based partiality that yields an equivalent initial evaluation and initial discrimination, and vice versa.*

Proof. Suppose the evaluator has belief-based partiality with beliefs $\hat{\mu}_M > \hat{\mu}_F$, but no preference-based partiality, $c_F = 0$. Then $v(h_1, s_1, F) = \frac{\tau_q \hat{\mu}_F + \tau_\eta s_1}{\tau_q + \tau_\eta}$ and $v(h_1, s_1, M) = \frac{\tau_q \hat{\mu}_M + \tau_\eta s_1}{\tau_q + \tau_\eta}$, yielding discrimination $D(h_1, s_1) = \frac{\tau_q}{\tau_q + \tau_\eta}(\hat{\mu}_M - \hat{\mu}_F)$. Setting $c'_F = \frac{\tau_q(\hat{\mu}_M - \hat{\mu}_F)}{\tau_q + \tau_\eta}$, $\hat{\mu}'_F = \hat{\mu}_M$ and $\hat{\mu}'_M = \hat{\mu}_M$ will yield equivalent evaluations and discrimination. The proof in the other direction is analogous. □

This proposition also holds for observing evaluations in a single round $t > 1$. For any level of discrimination in period t, there exist a type of evaluator with only preference-based partiality and a type of evaluator with only belief-based partiality that yield the same level of discrimination. Therefore, in order to identify the source of discrimination, we need to observe a richer cross-section of evaluations.

Precision of Signal. Next, we show that varying the level of subjectivity in judgement can identify whether discrimination is due to preference-based or belief-based partiality. In contrast to belief-based partiality, when the evaluator has preference-based partiality, less subjectivity in the judgment of quality, i.e. a more precise signal of quality, does not mitigate the animus towards women. Even if judgment is perfectly objective – signals are very precise – the female workers will still face discriminatory evaluations.

Proposition 6. *If the evaluator has preference-based partiality and no belief-based partiality, then discrimination is constant with respect to the precision τ_η of the signal.*

As judgement becomes perfectly objective ($\tau_\eta \to \infty$), discrimination persists if and only if an evaluator has preference-based partiality.

Proof. From (6), 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 taste-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 $c_F \neq 0$. □

Therefore, the comparative static with respect to the signal precision can distinguish between belief-based and preference-based partiality.[12] Observing either (i) no discrimination when the signal of quality is precise and discrimination when the signal of quality is imprecise; or (ii) a decrease in the level of discrimination with respect to the precision of the signal, provides evidence that belief-based partiality is the source of discrimination. In contrast, observing discrimination when the signal of quality is precise, or a constant level of discrimination with respect to the precision of the signal, provides evidence that preference-based partiality is the source of discrimination.

2.4.2 Dynamics of Discrimination

With preference-based partiality, evaluators believe that males and females have the same prior distribution of ability, but they subject the females to stricter standards. Similar to the belief-based case, a female must produce a higher signal than a male to receive the same evaluation. However, these stricter standards are required to offset the evaluator's distaste for females, rather than to offset lower beliefs about ability. Therefore, after the initial period, a female is perceived to be of *higher* average ability than a male who receives the same evaluation.[13] In subsequent periods, a female produces higher expected quality than a male with the same evaluation history. This reduces discrimination, but does not reverse it. Despite the higher expected quality, females are still subjected to stricter standards in subsequent rounds, due to the

[12] Note that for any level of belief-based partiality, the level of preference-based partiality that leads to equivalent evaluation behavior (Proposition 5) varies with τ_η.

[13] The intuition is similar to the reason that the impartial belief type has favorable beliefs towards females in periods $t > 1$ (Section 2.3.2).

preference-based partiality. Therefore, observing a discrimination reversal also rules out a preference-based model where evaluators have the same animus against females.

2.5 Discussion of Results

In summary, these 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 information; (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 either a correctly-specified model of belief-based partiality or a model of preference-based partiality in which all evaluators have common knowledge of animus c_F against females; and (iv) a reversal of discrimination points to belief-based partiality with misspecification. Before moving to the empirical section, a few aspects of our theoretical framework warrant further discussion.

Coarse Evaluations. Our set-up assumes 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 1-5. When this is the case, information will be lost, in the sense that each observed evaluation will correspond to an interval of possible signals. In Appendix B, we show that Proposition 3 generalizes to coarse evaluations, in that 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. This establishes that allowing evaluations to perfectly reveal signals is not the driving feature of the impossibility result.

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 B.1. Note that a shifting standard has no effect on discrimination, as it affects the standards faced by a worker across periods, but not the comparison between two workers of different genders in any given period.

Relation to Self-fulfilling Beliefs. Self-fulfilling beliefs are another form of unbiased belief-based partiality that can lead to discrimination. In these models, members of different groups are ex-ante identical and discrimination is an *equilibrium effect* – group members of a given type face a more exacting standard because they chose a lower investment level which, given the more exacting standard, is a best response. Fryer (2007) explores how self-fulfilling beliefs dynamically evolve. In his framework, an employer is "pessimistic about a group in general, but optimistic about the successful members of that group." He shows that beliefs can flip in the promotion (second) round if there exist equilibria in which the employer and employee of one group coordinate on an equilibrium with higher hiring standards and looser promotion standards, while employees of the other group coordinate on the reverse ordering – looser hiring standards and more stringent promotion standards. Thus, in Fryer (2007), belief-flipping depends on how this self-fulfilling equilibrium dynamically evolves, while in our model, discrimination reversals are a property of the endogenous informativeness of prior evaluations.

The existence of an equilibrium in which beliefs flip requires fairly strict conditions. For example, the payoff to employers who hire a qualified applicant must be significantly higher than the payoff to the applicant. In relation to our setting, this implies that 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. Additionally, multiple equilibria always exist – there are also equilibria in which beliefs do not flip and discrimination persists, equilibria in which all workers are treated equally, and equilibria in which the opposite group is initially discriminated against – so almost all outcomes are possible, conditional on observables.

3 A Field Experiment

We conduct a field experiment on an online Q&A mathematics forum. 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 calculating distributions from publicly available statistics. Second, the forum provided us with a private dataset on the voting behavior of users, which allows us 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). 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.[16] Reputation unlocks privileges, such as the ability to edit and comment on others' posts or tag questions as duplicates, and can be used as a currency through the assignment of "bounties." Users can increase the chances of getting a quality answer to their own questions by *spending* part of their reputation points on posting the question with a bounty. The reputation points associated with the bounty are transferred to the user who provides the highest quality answer, as determined by the question poster.

[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.

[16] Upvotes add 5 points to the poster's reputation for questions and 10 points for answers. Downvotes deduct 2 points from the poster's reputation for both questions and answers. It is not possible for a user's reputation to fall below 1.

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 an informative 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 informativeness 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 decision-processes. 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.

Related to the shifting standards scenario referenced in Section 2.5 and derived in Appendix B, the standard of quality used to determine whether to upvote a post may increase with reputation. Reputation determines which users rise through the rungs to become editors and moderators. Every upvote brings the user closer to positions on the forum where certain levels of mastery are expected. Hence, posts by high reputation users may be held to higher standards. For example, a new user may be rewarded with an upvote for a low-level calculus question, but a high-reputation user may not be.

3.2 Experimental Design

Varying gender and reputation permits us to test the dynamic predictions of different sources of discrimination, while the guidelines for assessing questions and answers allow us to study 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. The ability to exogenously vary the gender and reputation associated with the question poster made this an ideal setting for testing the dynamics of discrimination.

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 the accounts, 70 with female usernames and 70 with male usernames were left as-is; these comprised the Low Reputation accounts. For the other half of the accounts, we 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 by posting content on each account until the accumulated reputation reached 100. 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 ($M = 155.23$). Once an account reached at least 100 reputation points, the research assistant stopped posting content. These accounts comprised the High Reputation accounts. Critically, upon achieving a high reputation, we re-randomized the gender of the username: 35 of the accounts that were built-up under male usernames were switched to female, and 35 of the female accounts were switched to male; the remaining 70 accounts received a new username of the same gender. After reassigning usernames, the new female and male accounts had similar reputation levels ($M = 155.89$ vs. $M = 154.57$, respectively, $p = .82$). Importantly, when a username is switched, all past and future activity on the account became associated with the new username – 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.

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 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 chances of our content being 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: varying the gender of the username (Male vs. Female) and reputation level (Low vs. High).

In order to avoid unusual activity, i.e. flooding the forum with content, we posted questions on a pre-determined schedule. Research assistants posted one question at

[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.

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 7 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 or errors in the posting of the questions (i.e. two questions posted to the same account).

This set-up allows us to test the theoretical predictions outlined in Section 2. We measure discrimination as either the average number of upvotes per post, or the average change in reputation points per post, which is an aggregate measure of upvotes and downvotes. Conditional on observing discrimination between Low Reputation male and female accounts, a mitigation in its intensity for High Reputation accounts is consistent with belief-based partiality, including the case of statistical discrimination where beliefs are correct. Observing a *reversal* of discrimination for High Reputation accounts is evidence for biased belief-based partiality.

Note that we do not make a prediction on how reputation affects the overall level of upvotes between high and low reputation accounts (i.e. pooling genders), due to potential shifting standards (Section 2.5 and Appendix B). As previously discussed, reputation serves both the purpose of highlighting a quality post and rewarding the poster. Higher reputation should increase voters' beliefs about the quality of a question. At the same time, the same question asked by a low reputation user may not be rewarded if asked by a high reputation user due to shifting standards. In our experiment, randomization ensures that the average quality of questions posted to low reputation accounts is approximately the same as that of questions posted to high reputation 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: if the answer is correct or not. 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 of 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? The post has dozens of responses, including: is the question well-written, is it non-trivial, is it insightful, am I curious about the same question, has the poster made me curious about what they are asking, do I think it's 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. It should be noted that this response had one of the highest number of upvotes on the forum.

We created a second set of 140 Low Reputation accounts (i.e. no prior posts), with 70 male usernames and 70 female usernames. Questions on the forum are answered fairly quickly, and late answers often receive little attention, so posting answers to other questions required swift timing. 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 work on an answer to it. The 'answerer' would then send the answer and a link to the question to their partner, 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 person 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 7 days after posting for each answer, both in numerical form and as screenshots. A total of 5 of the 140 answers were dropped due to errors, e.g. question was closed before the 7 day window concluded.

The theory in Section 2 predicts that subjectivity in judgment, modeled as the level of 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, capturing relevant metrics for the experiment and to 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

Questions. Table 3 presents our results on the effect of reputation on discrimination. We first examine the changes in reputation and upvotes received for questions posted to Low Reputation accounts.[18] We find significant initial discrimination against females. Regressing the number of upvotes or the change in reputation on the gender of the poster reveals that questions posted to accounts with female usernames received significantly *fewer* upvotes (Column (1)) and accumulated significantly *fewer* additional reputation points (Column (2)) than questions posted to accounts with male usernames. These differences correspond to roughly 40% of a standard deviation for average number of upvotes and average change in reputation.

In contrast to Low Reputation accounts, questions posted to High Reputation accounts with female usernames received significantly *more* upvotes (Column (3)) and accumulated significantly *more* reputation points (Column (4)) than questions posted to High Reputation accounts with male usernames. These differences correspond to roughly 60% of a standard deviation for average number of upvotes and average change in reputation. We test the difference in the estimated coefficients of the male gender dummy between the Low Reputation and High Reputation regressions and find that this difference is significant for both upvotes ($\chi^2(1) = 10.03$; $p < .01$) and change in reputation ($\chi^2(1) = 9.67$; $p < .01$). Therefore, the male advantage is significantly larger at Low Reputations, compared to High Reputations.

Columns (5) and (6) present regression results for Low and High Reputation accounts within the same model. In Column (5), we regress the number of upvotes per question on dummies corresponding to the gender of the poster, the reputation level

[18] The change in reputation corresponds to the number of upvotes earned multiplied by 5 minus the number of downvotes earned multiplied by 2. Downvotes were very rare in our sample, and all results hold when using upvotes net of downvotes as the dependent variable.

Table 1. The Effect of Prior Evaluations on Discrimination

	Low Rep		High Rep		Low & High Rep		
	Upvotes	ΔRep	Upvotes	ΔRep	Upvotes	ΔRep	Binary
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Male	0.57**	2.86**	−0.64**	−3.16**	0.57**	2.86**	0.17**
	(.27)	(1.32)	(.27)	(1.37)	(.27)	(1.36)	(.08)
Rep					0.45*	2.33*	0.09
					(.27)	(1.35)	(0.08)
Male*Rep					−1.20***	−6.02***	−0.40***
					(.38)	(1.91)	(.11)
Constant	0.97***	4.68***	1.42***	7.01***	0.97***	4.68***	0.56***
	(.19)	(.93)	(.19)	(0.97)	(0.19)	(.96)	(.56)
# Obs	135	135	138	138	273	273	273

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$; Standard errors from OLS regressions reported in parentheses below each estimate; Male=1 if male username, 0 otherwise; Rep=1 if High Reputation account, 0 otherwise.

of the poster, and their interaction. This analysis confirms that the reversal of discrimination between the Low Reputation and High Reputation accounts: there is a negative and significant interaction between gender and reputation level. The same pattern of results hold for the change in reputation per question (Column (6)). To ensure that these results are not driven by outliers or subsequent voters herding on the first upvote, we perform the analysis using a binary variable that is equal to 1 if the question receives at least one upvote, and 0 otherwise. As shown in Column (7), the results are robust to this binary specification. Consistent with shifting standards, the average number of upvotes and change in reputation, pooled across both genders, does not significantly differ between Low and High reputation accounts.

Taken together, these results provide evidence that initial discrimination is driven by belief-based partiality with bias: not only is discrimination mitigated by reputation, but the direction *reverses.*

Answers. Table 4 presents our results for the effect of subjectivity on discrimination. We find no evidence of gender discrimination on answers. Regressing the number of upvotes or the change in reputation per answer on gender reveal no significant gender

Table 2. The Effect of Subjectivity on Discrimination.

	Answers Only		Questions & Answers	
	Upvotes	Δ Rep	Upvotes	Δ Rep
	(1)	(2)	(3)	(4)
Male	−0.20	−1.38	−0.20	−1.38
	(.17)	(.97)	(.23)	(1.16)
Question			0.17	0.08
			(.23)	(1.16)
Male*Question			0.77**	4.24**
			(.32)	(1.64)
Constant	0.81***	4.60***	0.81***	4.60***
	(.12)	(.69)	(.16)	(.82)
# Obs	135	135	270	270

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$; Standard errors from OLS regressions reported in parentheses; Male=1 if male username, 0 otherwise; Question=1 if question post, 0 if answer; Low Reputation accounts only.

differences in the evaluation of answers, as presented in Columns (1) and (2).[19] We test the difference in the estimated coefficients of the male gender dummy between the Low Reputation question and answer regressions and find that this difference is significant for both upvotes ($\chi^2(1) = 5.83$; $p = .02$) and change in reputation ($\chi^2(1) = 6.34$; $p = .01$). The male advantage is significantly larger for questions, compared to answers.

Columns (3) and (4) present regression results for Low Reputation questions and answers within the same model. We regress the number of upvotes or change in reputation per post on dummies corresponding to gender, type of post (question or answer) and their interaction. 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.

These results are consistent with our theoretical prediction on how the level of subjectivity affects discrimination stemming from belief-based partiality. They are inconsistent with discrimination due to preference-based partiality.

Robustness Checks. To test the robustness of the results and provide further evidence for our assumptions, we obtained a private dataset from the forum that provides

[19] Recall that there are only Low Reputation answer accounts.

additional information about the evaluators in our experiment. Specifically, the dataset allows us to uniquely identify the users who evaluated our content by voting on the questions and answers in the experiment, as well as to track their historical activity on the forum.[20] The dataset also includes time stamps for all actions taken by the users.

We first used 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 on one of our posts from each evaluator, and re-ran the analyses from Tables 3 and 4. Our findings are robust to excluding repeat evaluators. The results are presented in Appendix C.

We also explored whether users systematically differed in the type of content they evaluated on the forum. Specifically, we sought to determine whether the users who evaluated our posts specialized in the type of content they usually evaluated by either evaluating mostly questions or mostly answers, or whether most users evaluated both. To examine this question, 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, depending on the type of posts (questions versus answers). Pairwise comparisons revealed no significant differences ($p > .4$ for all comparisons).

3.4 Analysis of Observational Data

Next, we analyzed an observational dataset of evaluation behavior on the forum to complement our experimental results. We use this data to estimate relevant statistics that are publicly available to users on the forum.

Description of Data. The observational dataset is compiled and made publicly available by the forum. 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 between July 2010 and March 2017. To code gender, we ran an algorithm to classify the gender of the usernames. We followed the gender resolution

[20] The usernames of the evaluators were unique but anonymized for privacy.

approach of Vasilescu, Capiluppi, and Serebrenik (2014), who both developed the algorithm and validated its accuracy through secondary data collection on online Q&A forums.[21] 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.'

Analysis. We examine how evaluations of posts in the observational data vary with reputation, inferred gender of the user and type of post. For each user, we calculate the change in reputation per post. This variable corresponds to the evaluation of quality of the user's posted content. As in our experiment, we look at the evaluation of questions posted to low and high reputations, and the evaluation of answers posted to low reputations. We define low and high reputations similar to the experiment: Low Reputation corresponds to accounts with no prior reputation or posts, and High Reputation corresponds to accounts that attained reputations of 100 to 240 points (the range in our High Reputation experimental condition).

This analysis comes with several important caveats. First, there is the obvious endogeneity problem in not being able to control for quality of question posts at either reputation level. Second, there may be substantial gender-based selection between the low and high reputation levels. Finally, the number of posts that generated a user's reputation is relevant for inferring ability. The issue of different numbers of posts resulting in similar reputations is controlled for in our experiment through randomization; it is less straightforward to control for this issue in the observational data.[22]

[21] To increase accuracy, 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. The first step is preprocessing the data in order to obtain (*name, country*) tuples for each user when such information is available. This involves eliminating special characters and converting Leet to Latin (e.g. w3513y to Wesley). The preprocessed data is then fed into a Python tool that classifies the tuple as 'male,' 'female,' or 'x' (when gender cannot be inferred). When inferring gender, the tool goes through 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.

[22] In the analysis of observational data, we attempt to address the issue by looking at High Reputation accounts which 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. 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 reach the reputation threshold. Increasing

Keeping these caveats in mind, the results from the observational data are similar to the effects documented in the experiment. First looking at Low Reputation accounts, questions posted by accounts with male usernames receive significantly more reputation points than those posted by accounts with female usernames ($\beta = 1.33$, $p < .01$). This differential evaluation reverses direction for High Reputation accounts: questions posted by accounts with male usernames receive significantly fewer reputation points than those posted by accounts with female usernames ($\beta = -1.57$, $p = .02$). Comparing the coefficients reveals a significant difference: the estimated coefficient on the male gender dummy in the regression on questions posted to Low Reputation accounts is significantly larger than the estimated coefficient on the male gender dummy in the regression on questions posted to High Reputation accounts ($\chi^2(1) = 12.28$; $p < .01$).

Next, we compare the evaluation of questions and answers. As in the experiment, we find no significant differences by gender for Low Reputation accounts; a similar number of reputation points are earned for answers posted to accounts with male and female usernames ($\beta = -0.31$, $p = .35$). The estimated effect of gender on the change in reputation is significantly different for answers posted to Low Reputation accounts than for questions posted to Low Reputation accounts ($\chi^2(1) = 7.23$; $p < .01$). The analysis of observational data corroborates our experimental results in suggesting that systematically biased beliefs are a significant driver of the documented discrimination.

Stereotyping. As shown in Section 2, 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 more exacting standards. In this subsection, we use publicly available statistics from the observational dataset to explore one potential mechanism that could lead to these 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

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.

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.[23]

In the Appendix D, we explore how 'representativeness' can lead to biased beliefs in our setting. We examine the distribution of users' reputations per answer post over the *entire* range of reputations. 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. Comparing the reputation earned per post of those with male and female usernames, we find that the difference in means is fairly small and only marginally significant. However, we show that even mild belief distortions of these true means due to 'representativeness' can quickly exacerbate the small underlying difference, and lead to large differences in the perceived means of ability ($\hat{\mu}_F$ and $\hat{\mu}_M$ from the theory model). While the perceived means are fairly similar when the distortion is minimal, under the moderate levels of distortion that are consistent with other studies (Arnold, Dobbie, and Yang 2017), the difference in perceived means triples. 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.

4 Discussion and Conclusion

In this paper, we develop a model of discrimination, and explore both how it evolves dynamically and how it responds to the degree of subjectivity in judgment. We show that the observable patterns of discrimination along these two dimensions depend critically on the underlying source – which we term *partiality*. The analysis yields an impossibility result: discrimination does not dynamically reverse if it is driven by partiality with correctly specified beliefs. In contrast, a reversal can occur if some evaluators hold biased stereotypes, while others are aware of the bias and account for it at later stages. Finally, we 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

[23] 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.

more objective. Therefore, manipulating the level of subjectivity can be used to further identify the underlying source of discrimination.

We present results from a field experiment exploring discrimination along the two dimensions outlined in the theory. We post questions and answers on an online forum to accounts that vary in the gender of the usernames and their reputation on the forum. An account's reputation is generated endogenously through upvotes on previously posted content. This allows us to examine discrimination at different stages – the initial stage, when the reputations of users is low, and more advanced stages, after users have accumulated higher reputations. We document three main results: (i) significant gender discrimination *exists* at the initial stages, in the form of fewer upvotes and less reputation gained on questions posted to low reputation female accounts than to male accounts; (ii) discrimination *reverses* at the more advanced stage, in that more upvotes and more reputation are gained on questions posted to high reputation female accounts than to male accounts; and (iii) discrimination *mitigates* at the initial stage for answers, where judgment of quality is less subjective relative to questions. We also analyze observational data from the forum. Using an algorithm to infer gender from usernames, we provide additional evidence for the main findings from the experiment.

Taken together, these results are consistent with discrimination driven by belief-based partiality with some form of misspecification. Using publicly available group statistics from the forum, we show that even a small degree of distortion that stems from using a 'representativeness' heuristic (Bordalo et al. 2016b) to form beliefs leads to significantly biased stereotypes, where male users are perceived to be of higher average ability than the underlying distributions suggest. As demonstrated in our theoretical framework, if some evaluators hold such distorted beliefs, this could lead to the observed reversal in discrimination.

Our results help reconcile seemingly contradictory findings in the literature on gender discrimination. Reuben, Sapienza, and Zingales (2014) find that students performing arithmetic problems were less likely to be hired in an experimental market if they were female than male. In contrast, Heikensten and Isaksson (2016) show that female students selected to serve as 'knowledge lifelines' in fields such as arithmetic and biology were more likely to be chosen by game show contestants than male students. While Milkman et al. (2012) finds that female graduate students are less likely to receive a response from a faculty member in many academic settings, Williams and Ceci (2015) documents discrimination in the opposite direction for tenure-track job applicants in STEM fields: female applicants are ranked higher than their male counterparts 2:1. While these conclusions seem contradictory if each result is treated in isolation, our findings suggest that these studies may have captured discrimination at different stages of a dynamic process. While Reuben et al. (2014) drew participants from the general student population, the students in Heikensten and Isaksson (2016) had to pass stringent thresholds before qualifying to participate. Milkman et al. (2012) document discrimination against women at the initial stages of graduate study, while Williams and Ceci (2015) demonstrate reverse discrimination for highly accomplished candidates with numerous publications.

In Williams and Ceci (2015), the authors conclude that "it is a propitious time for women launching careers in academic science." Our findings suggest that this conclusion may be premature: if the dynamic nature of discrimination is taken into account, a preference for women at the later stages could be a function of discrimination *against* them at the initial ones. Moreover, if biased beliefs are playing a role in driving discrimination, as our analysis suggests, then the reversal at later stages does *not* make up for the initial discrimination against women. Women may be inefficiently selected out of the pipeline in earlier stages. Additionally, conditional on making it to the later stages, these women should be receiving higher evaluations than they actually are.

Our findings shed light on the mechanism behind previously documented discrimination reversals. Booth et al. (1999); Groot and van den Brink (1996) show a reversal in discrimination against women at different stages of the hiring and promotion process. In recent work, Mengel, Sauermann, and Zolitz (2017) find that at the junior level, female instructors systematically receive lower teaching evaluations for similar courses, compared to male instructors, 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 may be indicative of belief-based discrimination with biased priors, e.g. 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 (Bordalo, Coffman, Gennaioli, and Shleifer 2016a; Coffman 2014). Additionally, consistent with our theoretical result on the possibility of a reversal, female students – who should be less likely to hold stereotypes against female teachers – drive the reversal in evaluations: male students discriminate against junior female faculty but not senior female faculty, while female students do not discriminate significantly against junior female faculty and *favor* senior female faculty.

How discrimination dynamically evolves and varies with subjectivity of judgment has significant implications for policy. 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 misspecification, the welfare criterion is clear: incorrect beliefs are inefficient, so campaigns targeting beliefs would improve outcomes on both the efficiency and fairness dimensions. If discrimination is statistical, driven by belief-based partiality with correctly specified beliefs, then affirmative action policies have a trade-off: they increase fairness, while reducing efficiency. But making evaluations more objective will improve fairness while maintaining efficiency. If the policymaker also has a preference for equal outcomes, i.e. output from a woman should be as likely to be rewarded as output from a man, then the target should be investment in malleable dimensions of ability to equalize the distributions. If discrimination is driven by preference-based partiality, then the policymaker should target norms that may affect such preferences; manipulating objectivity of evaluations will not improve fairness.

The findings on dynamics also have implications for the timing of policy interventions. If the source of discrimination is belief-based, then interventions that change initial evaluations standards, such as affirmative action, may have the unintended consequences of pushing discrimination to later stages. Specifically, 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. Suggestive evidence of this phenomenon is reported in Bertrand and Mullainathan (2004), who find lower returns to education for resumes with African-American sounding names, compared to Caucasian names. In turn, policies that do not shift beliefs about initial thresholds may be more effective at mitigating discrimination both at the initial stages and down the road. For example, oversampling from discriminated groups at the initial stages would lead to more equal representation without shifting beliefs about the standards. As a result, evaluators at later stages may be less likely to perpetuate discriminatory practices.

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.

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.

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.

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.

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. 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.

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.

Heikensten, E. and S. Isaksson (2016): "In Favor of Girls: A Field Study of Adults' Beliefs in Children's Ability," *mimeo.*

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.

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.*

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

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.

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.

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.

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.

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.

A Appendix: Proofs from Section 2

The following lemma is used in the proofs of Propositions 3 and 4.

Lemma 1. *Suppose that a type of evaluator believes that a new worker's ability is normally distributed with mean $\hat{\mu}$ and precision τ_a, has taste parameter c_g, and believes that all other evaluators are also this type. Then following any history h_t, the perceived 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 $\tau_a(t) = \tau_a + (t-1)\tau_{\varepsilon\eta}$, where

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

for all $n < t$.

Proof. Suppose $f_{\hat{\mu}}(a|h_1) \sim N(\hat{\mu}, 1/\tau_a)$. From (2), conditional on observing signal s_1, the first evaluation is

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

It is possible to back out s_1 from observing v_1,

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

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 $\tau_q(t) \equiv \tau_a(t)\tau_\varepsilon/(\tau_a(t) + \tau_\varepsilon)$. The evaluator maximizes her expected payoff by choosing

$$v_t = \frac{\tau_q(t)\hat{\mu}(h_t) + \tau_\eta s_t}{\tau_q(t) + \tau_\eta} - c(r(h_t)) - c_g \tag{7}$$

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

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

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

$$\hat{\mu}(h_t) = \frac{\tau_a\hat{\mu} + \tau_{\varepsilon\eta}\sum_{n=1}^{t-1} s(v_n, \hat{\mu}(h_n), n)}{\tau_a + (t-1)\tau_{\varepsilon\eta}} \tag{8}$$

$$\tau_a(t) = \tau_a + (t-1)\tau_{\varepsilon\eta}. \tag{9}$$

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 (8) and (9). □

Proof of Proposition 3. 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

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

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,

$$\begin{aligned}\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) &= \left(\frac{\tau_a(t)}{\tau_a(t) + \tau_{\varepsilon\eta}}\right)(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)) \\ &\quad + \left(\frac{\tau_{\varepsilon\eta}}{\tau_a(t) + \tau_{\varepsilon\eta}}\right)(s_M(v_t, \hat{\mu}_M(h_t), t) - s_F(v_t, \hat{\mu}_F(h_t), t)) \\ &= \left(\frac{\tau_a(t)}{\tau_a(t) + \tau_{\varepsilon\eta}} - \frac{\tau_{\varepsilon\eta}\tau_q(t)}{(\tau_a(t) + \tau_{\varepsilon\eta})\tau_\eta}\right)(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)), \quad (10)\end{aligned}$$

which is positive if

$$\begin{aligned} & \frac{\tau_a(t)}{\tau_a(t) + \tau_{\varepsilon\eta}} - \frac{\tau_{\varepsilon\eta}\tau_q(t)}{(\tau_a(t) + \tau_{\varepsilon\eta})\tau_\eta} > 0 \\ \Leftrightarrow \quad & \frac{\tau_a(t)\tau_\eta - \frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon + \tau_\eta} \times \frac{\tau_a(t)\tau_\varepsilon}{\tau_a(t) + \tau_\varepsilon}}{(\tau_a(t) + \frac{\tau_\varepsilon\tau_\eta}{\tau_\varepsilon + \tau_\eta})\tau_\eta} > 0. \quad (11)\end{aligned}$$

This will be the case if the numerator of (11) 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 (10), $\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) < \hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)$ iff (11) 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 (7), discrimination in period t is equal to

$$D(h_t, s_t) = \frac{\tau_q(t)}{\tau_q(t)+\tau_\eta}(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)). \tag{12}$$

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 3.

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 4. Type θ^B'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^U'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^U's belief about female ability evolves.

When type θ^U observes evaluation v_1, she believes that with probability p, it is from a biased type who observed signal $s_1^B(v_1) = \left(\frac{\tau_q+\tau_\eta}{\tau_\eta}\right)v_1 - \left(\frac{\tau_q}{\tau_\eta}\right)\hat{\mu}_F^B$, and with probability $1-p$, it is from an unbiased type who observed signal $s_1^U(v_1) = \left(\frac{\tau_q+\tau_\eta}{\tau_\eta}\right)v_1 - \left(\frac{\tau_q}{\tau_\eta}\right)\hat{\mu}$. Note $s_1^B(v_1) > s_1^U(v_1)$. Therefore, the likelihood function for evaluation v_1 is a mixture of two normal distributions,

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

Since the prior belief $f_a(a) \sim N(\hat{\mu}, 1/\tau_a)$ 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

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

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

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

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. Therefore, in the second period, the unbiased type gives females evaluation

$$v_{2,F}^U(s_2, v_1) \;=\; \left(\frac{\tau_\eta}{\tau_{q,2}+\tau_\eta}\right)s_2 + \left(\frac{\tau_{q,2}}{\tau_{q,2}+\tau_\eta}\right)\left(\frac{pC_1D_1}{CD}\hat{\mu}_1(v_1) + \frac{(1-p)C_2D_2}{CD}\hat{\mu}_2(v_1)\right).$$

Define $\gamma(v_1) \equiv \frac{pC_1D_1}{CD}\hat{\mu}_1(v_1) + \frac{(1-p)C_2D_2}{CD}\hat{\mu}_2(v_1)$. In the second period, the biased type gives females evaluation

$$v_{2,F}^B(s_2, v_1) = \left(\frac{\tau_\eta}{\tau_{q,2}+\tau_\eta}\right)s_2 + \left(\frac{\tau_{q,2}}{\tau_{q,2}+\tau_\eta}\right)\hat{\mu}_{F,2}^B(v_1)$$

following initial evaluation v_1 and signal s_2, where from Lemma 1, $\hat{\mu}_{F,2}(v_1) = \frac{\tau_a\hat{\mu}_F^B + \tau_{\varepsilon\eta}s_1^B(v_1)}{\tau_a+\tau_{\varepsilon\eta}}$. Both types give males evaluation

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

following initial evaluation v_1 and signal s_2, where from Lemma 1, $\hat{\mu}_{M,2}(v_1) = \frac{\tau_a\hat{\mu} + \tau_{\varepsilon\eta}s_1^U(v_1)}{\tau_a+\tau_{\varepsilon\eta}}$.

Fixing v_1 and s_2, discrimination in period 2 is equal to

$$\begin{aligned} D(v_1, s_2) &= p(v_{M,2} - v_{F,2}^B) + (1-p)(v_{M,2} - v_{F,2}^U) \\ &= p\left(\frac{\tau_{q,2}}{\tau_{q,2}+\tau_\eta}\right)(\hat{\mu}_{M,2}(v_1) - \hat{\mu}_{F,2}^B(v_1)) + (1-p)\left(\frac{\tau_{q,2}}{\tau_{q,2}+\tau_\eta}\right)(\hat{\mu}_{M,2}(v_1) - \gamma(v_1)) \\ &= \left(\frac{\tau_{q,2}}{\tau_{q,2}+\tau_\eta}\right)(\hat{\mu}_{M,2} - p\hat{\mu}_{F,2}^B(v_1) - (1-p)\gamma(v_1)). \end{aligned}$$

Discrimination reverses at (v_1, s_2) if $D(v_1, s_2) < 0$. We know that at $p = 0$, $D(v_1, s_2) = 0$ for all (v_1, s_2), as this is the case with no partiality, and at $p = 1$, $D(v_1, s_2) > 0$ for all (v_1, s_2), as this is the case with a single type of evaluator with belief-based partiality from Proposition 3. 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

$$1 < \left(\frac{\tau_\varepsilon^2}{(\tau_\varepsilon+\tau_\eta)(\tau_a+\tau_\varepsilon}\right)\left(1 + \frac{C_1D_1}{C_2D_2}\right).$$

From the expressions above,

$$\begin{aligned}\frac{C_1D_1}{C_2D_2} &= \exp(-0.5\tau_{\varepsilon\eta}(s_1^B(v_1)^2 - s_1^U(v_1)^2) + 0.5(\tau_a + \tau_{\varepsilon\eta} - \tau_{q,2})(\hat{\mu}_1(v_1)^2 - \hat{\mu}_2(v_1)^2) \\ &\quad +0.5(\tau_{q,2} + \tau_\eta(\hat{\mu}_1(v_1, s_2)^2 - \hat{\mu}_2(v_1, s_2)^2)),\end{aligned}$$

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 initial sets of beliefs for each type, it is possible for discrimination to reverse in the second period.

B Extensions

B.1 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

$$v(h_t, s_t, g) = \frac{\tau_{q,t}\hat{\mu}_g(h_t) + \tau_\eta s_t}{\tau_{q,t} + \tau_\eta} - c(r(h_t)) - c_g, \tag{13}$$

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

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

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

B.2 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

$$E[q_t | h_t, s_t, g] \geq c_g, \tag{14}$$

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 7. *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 7. 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

$$\begin{aligned}
\frac{f_\mu(a|v_1 = 1)}{f_\mu(a'|v_1 = 1)} &= \frac{P_\mu(v_1 = 1|a)}{P_\mu(v_1 = 1|a')} \cdot \frac{f_\mu(a)}{f_\mu(a')} \\
&= \frac{\Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a - \overline{s}(\mu, c))\right)}{\Phi\left(\sqrt{\tau_{\varepsilon\eta}}(a' - \overline{s}(\mu, c))\right)} \cdot \frac{\phi(\sqrt{\tau_a}(a - \mu))}{\phi(\sqrt{\tau_a}(a' - \mu))}. \qquad (15)
\end{aligned}$$

The goal is to show that (15) 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,

$$\frac{\partial^2}{\partial a \partial \mu} \log P_\mu(v_1 = 1|a) + \log f_\mu(a) \geq 0. \tag{16}$$

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 (16),

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

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 (16)

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

Therefore, need to show that for all x,

$$\begin{aligned}
\tau_a - \left(\frac{\phi(x)x}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2}\right)\left(\frac{\tau_q \tau_{\varepsilon\eta}}{\tau_\eta}\right) &\geq 0 \\
\Leftrightarrow \tau_x - \left(\frac{\phi(x)x}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2}\right) &\geq 0,
\end{aligned} \tag{17}$$

where $\tau_x \equiv \frac{\tau_a \tau_\eta}{\tau_q \tau_{\varepsilon\eta}}$. From Stack Exchange[24], we know that

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

From the definition of τ_x,

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

Therefore, (17) holds for all x. Therefore, for all $a > a'$, (15) 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,

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

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

which also means that

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

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

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

C Robustness Checks

Results after dropping votes from repeat evaluators.

Table 3. The Effect of Prior Evaluations on Discrimination

	Upvotes	*Δ* Rep	Upvotes	*Δ* Rep	Upvotes	*Δ* Rep	Binary
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Male	0.43* (.22)	2.17** (1.07)	-0.52** (.23)	-2.58** (1.14)	0.43* (.22)	2.17* (1.12)	0.10 (.08)
Reputation					0.31 (.22)	1.64 (1.11)	0.02 (0.08)
Male*Reputation					-0.95*** (.31)	-4.75*** (1.57)	-0.28** (.12)
Constant	0.66*** (.15)	3.13*** (.76)	.97*** (.16)	4.77*** (0.81)	0.66*** (0.16)	3.13*** (0.79)	0.44*** (0.06)
# Observations	135	135	138	138	273	273	273

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors from OLS regressions reported in parentheses below each estimate. Male=1 if male username, 0 otherwise; Reputation=1 if High Reputation account, 0 otherwise. Columns 1 and 2 report analyses for Low Reputation accounts. Columns 3 and 4 report analyses for High Reputation Accounts. Columns 5, 6, 7 report analyses for both Low and High Reputation Accounts.

Table 4. The Effect of Subjectivity on Discrimination.

	Upvotes	Δ Rep	Upvotes	Δ Rep
	(1)	(2)	(3)	(4)
Male	-0.20 (.16)	-1.15 (.82)	-0.20 (.19)	-1.15 (.96)
Question			-0.05 (.19)	-0.42 (.96)
Male*Question			0.63** (.27)	3.32** (1.35)
Constant	0.72*** (.12)	3.55*** (.58)	0.72*** (.14)	3.55*** (.68)
# Observations	135	135	270	270

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Standard errors from OLS regressions reported in parentheses below each estimate. Male=1 if male username, 0 otherwise; Question=1 if question post, 0 if answer. Columns 1 and 2 report analyses for answers posted to Low Reputation accounts. Columns 3 and 4 report analyses for both questions and answers posted to Low Reputation Accounts.

D Stereotyping

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

$$\pi_{t,g}^{st} \equiv \pi_{t,g} \frac{\left(\frac{\pi_{t,g}}{\pi_{t,-g}}\right)^{\theta}}{\sum_{s \in T} \pi_{s,g} \left(\frac{\pi_{s,g}}{\pi_{s,-g}}\right)^{\theta}}, \tag{18}$$

where $\theta \geq 0$ corresponds to the extent of the belief distortion. Incorrect stereotypes are most likely to form when there are group differences in the frequency of a particular type, but the overall type distributions are largely the same.This is consistent with recent empirical work that finds support for the model (Arnold et al. 2017; Bordalo et al. 2016a; Coffman 2014).

We explore how 'representativeness' can bias beliefs in our setting by examining the

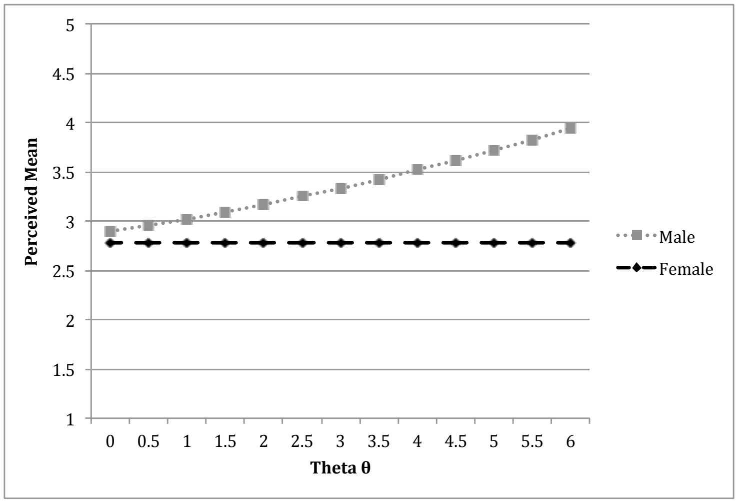

Figure 2. Perceived mean $\hat{\mu}_g$ as function of θ, by gender.

distribution of users' reputations per answer post. Since we do not find evidence for discrimination on answer posts in either the experiment or the analysis of observational data, we can view this as a proxy for ability. We divide the distribution of reputation 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 the small underlying difference. Figure 2 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 (θ=2.5) consistent with empirical estimates from other 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.