American Economic Review 2019, 109(10): 3395–3436
https://doi.org/10.1257/aer.20171829

The Dynamics of Discrimination: Theory and Evidence[†]

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

We model the dynamics of discrimination and show how its evolution can identify the underlying source. We test these theoretical predictions in a field experiment on a large online platform where users post content that is evaluated by other users on the platform. We assign posts to accounts that exogenously vary by gender and evaluation histories. With no prior evaluations, women face significant discrimination. However, following a sequence of positive evaluations, the direction of discrimination reverses: women's posts are favored over men's. Interpreting these results through the lens of our model, this dynamic reversal implies discrimination driven by biased beliefs. (*JEL* C93, D83, J16, J71)

A rich literature has documented discrimination in a wide range of contexts (Bertrand and Duflo 2017). These empirical studies have mostly focused on static settings: individuals are evaluated based on the quality of a single piece of output or a single interaction, with no information on prior evaluations in similar contexts. As prior work has noted, it is difficult to identify the underlying source of discrimination from such static settings, as different sources generate the same patterns of observable behavior (Fang and Moro 2011). In this paper, we develop a theoretical framework to show how the dynamics of discrimination can be used to identify its underlying source, and test these predictions in a field experiment on a large online platform.

Consider a setting where individuals repeatedly perform tasks that generate output, and in the process, produce an observable history of evaluations on these tasks. For example, a man and a woman are employed at a firm and are promoted based on

* Bohren: University of Pennsylvania, 133 South 36th Street, Philadelphia, PA 19104 (email: abohren@sas.upenn.edu); Imas: Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 (email: aimas@andrew.cmu.edu); Rosenberg: CarGurus, 2 Canal Park, Cambridge, MA 02141 (email: rosenberg.michael.m@gmail.com). Stefano DellaVigna was the coeditor for this article. We thank Nageeb Ali, Linda Babcock, Michael Callen, Hanming Fang, Uri Gneezy, Polina Imas, Nicola Gennaioli, Gregor Jarosch, Emir Kamenica, Gene Kucher, George Loewenstein, Kristof Madarasz, Craig McIntosh, Margaret Meyer, Siqi Pan, Sally Sadoff, Lise Vesterlund, Leeat Yariv, and seminar and conference participants for helpful comments. We thank Mustafa Dogan, Daniel Hauser, Conrad Kosowski, Catherine Lillis, Suneil Parimoo, Jaesung Son, and Lucia Zhang for excellent research assistance. Finally, we would like to thank the team at the online forum for helpful and illuminating discussions, and for providing us with the data for our analyses. The project received IRB approval at Carnegie Mellon and University of Pennsylvania. The experiment was pre-registered in the AEA RCT Registry (AEARCTR-0000950). The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[†] Go to https://doi.org/10.1257/aer.20171829 to visit the article page for additional materials and author disclosure statements.

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

The answer to this question depends critically on the underlying source of discrimination. If the source is belief-based—for example, the quality of output is imperfectly observed and evaluators believe that on average, men have higher abilities than women—then observing prior evaluations will reduce discrimination against women, relative to men with similar evaluations. This dynamic effect operates through two channels. First, prior evaluations provide signals of a worker's ability, which reduces the impact of perceived group statistics (e.g., beliefs about average ability) on how the worker's subsequent output is evaluated.[1] This mitigates discrimination between males and females with similar evaluation histories. Second, and particular to a social learning setting, the informational content of these signals is *endogenously* determined by the behavior of prior evaluators. When initial beliefs favor men, to overcome this disparity, a woman needs to produce *higher* quality output than a man to receive a similar evaluation: for example, to be promoted or have her output accepted. This speeds up the mitigation of discrimination for evaluators who are aware of the higher standard for women. These evaluators may even come to believe that the woman is of higher ability than a man with a similar evaluation history, thereby favoring her output over the man's and *reversing* the direction of discrimination in subsequent periods. In fact, observing a reversal can help disentangle whether evaluators' models are correct or misspecified; we show theoretically that a reversal provides evidence for bias. In contrast to belief-based sources, if discrimination is caused by a taste or preference against rewarding or interacting with women (Becker 1957), then a woman who receives a similar sequence of evaluations to a man will continue to face discrimination in future periods.

Our theoretical framework formalizes the relationship between the dynamic pattern of discrimination, which is based on observable evaluations, and the sources of discrimination, which are unobservable and depend on underlying preferences and beliefs. The literature on belief-based sources has generally focused on correct beliefs (e.g., rational expectations), where evaluators are partial toward a group based on true differences in the underlying distributions of the relevant attribute (Fang and Moro 2011; Phelps 1972; Altonji and Pierret 2001; Knowles, Persico, and Todd 2001). However, recent research has demonstrated that systematic biases in judgment can lead to incorrect stereotypes against a particular group (Schwartzstein 2014;

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

Fryer and Jackson 2008; Bordalo et al. 2016). Therefore, we allow for three potential sources: (i) belief-based with correct beliefs, (ii) belief-based with incorrect, biased beliefs, and (iii) preference-based. We show that these sources make contrasting dynamic predictions. When discrimination is based on common knowledge of correct beliefs, then observing similar sequences of evaluations for a man and a woman will mitigate discrimination, but will never lead to a reversal. Therefore, observing a dynamic reversal provides evidence for a belief-based source with *bias*, since it is also inconsistent with standard preference-based sources. We also illustrate how one form of bias, where some evaluators hold incorrect stereotypes against a group and other evaluators are aware of these stereotypes, can lead to a dynamic reversal.

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

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

In our experiment, we posted original mathematics questions on created accounts that exogenously vary in the gender of the username and the reputation of the user. Our setting is well suited for exploring the dynamics of discrimination because we are able to exogenously vary the evaluation histories of users, as summarized by their publicly observable reputations. We posted one-half of the questions to novice accounts that did not have prior evaluations. We manually built the reputations of the remaining accounts by posting content until the reputations reached the top

twenty-fifth percentile on the forum. We then randomly reassigned the gender of the username to avoid endogeneity issues and ensure that the underlying informational content of reputation is the same for both genders. Finally, we posted the remaining questions to these advanced accounts. We compare the pattern of discrimination between novice and advanced question posts to test the dynamic predictions of the different sources of discrimination.

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

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

In addition to our experimental results, we exploit two additional data sources: a proprietary dataset that contains additional information about the users who evaluated the content from our experiment, and a large observational dataset of all posts on the forum. We used these datasets to run additional robustness tests and to rule

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

out other potential explanations for the observed reversal, such as gender differences in attrition or the variance of ability. We also compare discrimination by type of post and reputation in the observational data. We find analogous patterns to the experiment, including both the dynamic reversal between questions posted by novice and advanced users and the lack of discrimination for answers.

The findings presented here highlight the importance of studying discrimination in dynamic settings, as discrimination in favor of a certain group, or a lack thereof, at any given stage can either be a function of or precursor to discrimination against that same group at a different stage. Both in academic and popular discourse, a common argument used to illustrate the *lack* of discrimination against a group is to point to individuals from that group who have made it to positions of prominence. Our theoretical framework and empirical evidence highlight the flaw of this argument: if individuals are aware that members of a group face discrimination at an earlier stage, there may be Bayesian foundations for favoring members of that group at later stages. For example, in a much-discussed paper, Williams and Ceci (2015, p. 5361) find that accomplished female academics in STEM fields are favored over male academics. The authors state that "these results suggest it is a propitious time for women launching careers in academic science." In contrast, other work has found significant discrimination against female students in STEM (Reuben, Sapienza, and Zingales 2014; Moss-Racusin et al. 2012). While these sets of findings appear contradictory, our results suggest that discrimination in favor of accomplished female professors may actually be a function of discrimination *against* women earlier in the pipeline.

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

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

these evaluations will be suboptimal, particularly when future evaluators are also biased or are not aware of the bias of prior evaluators.[3]

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

Related Literature.—Discrimination has been documented in a wide range of settings, including hiring (Riach and Rich 2006; Bartoš et al. 2016), housing (Ewens, Tomlin, and Wang 2014), and service markets (Gneezy, List, and Price 2012). It has also been documented against group identities based on race (Bertrand and Mullainathan 2004; Parsons et al. 2011), ethnicity (Fershtman and Gneezy 2001; Milkman, Akinola, and Chugh 2012) and gender (Moss-Racusin et al. 2012; Goldin and Rouse 2000). The few studies that use observational data to attempt to identify the source of discrimination typically compare the evaluations of a state (for example, whether output is accepted or rejected) to the true underlying value of that state (i.e., whether the output was actually high or low quality). For example, Knowles, Persico, and Todd (2001) compare decisions of law enforcement to search a motor vehicle to the success rate of the search; similarities in success rates across races led the authors to conclude that higher search rates for African American drivers are due to statistical rather than preference-based discrimination (see, for similar tests, Anwar and Fang 2006 and Arnold, Dobbie, and Yang 2018). In recent work, Sarsons (2017) uses an event study approach for matched samples of surgeons to explore belief-based gender discrimination in physician referrals. She concludes that the observed pattern of gender discrimination is not consistent with Bayesian learning with respect to accurate beliefs about the distribution of surgeon ability. However, in many observational settings, it is difficult or impossible to construct matched samples or to observe the true value of the underlying state at an individual level. Further, observational data often face endogeneity issues that preclude the causal identification of discrimination.

Due to endogeneity issues, many researchers have employed field experiments to study discrimination. Field experiments have been successful in causally identifying the incidence of discrimination, but most cannot identify the source of this discrimination (Bertrand and Duflo 2017). One notable exception is List (2004), who documents that minorities receive inferior initial and final offers when bargaining in a market for sports cards. He supplements these data with a series of artefactual and framed field experiments to identify the source of this discrimination. Data from

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

dictator games, market, and auction experiments provide support for belief-based discrimination, and rule out preference-based sources. This method demonstrates how eliciting the true value of the underlying state for different groups (e.g., the distribution of reservation prices across races) can identify the source of discrimination in a static setting. We provide a complementary approach that illustrates how dynamic data and variation in the subjectivity of judgment can be used to achieve the same goal.

Our findings shed light on the mechanism behind previously documented discrimination reversals. In labor market settings, Groot and van den Brink (1996); Booth, Francesconi, and Frank (1999); Lewis (1986); and Petersen and Saporta (2004) find discrimination against women at the initial hiring stage for promotable jobs, but conditional on being hired, they find that women are more likely to be promoted. Rosette and Tost (2010) document a female leadership premium, showing that in contrast to women at lower levels within an organization, women in high positions are seen as more effective than men at similar positions.[4] In a field experiment, Ayalew, Manian, and Sheth (2018) show that workers are more likely to follow a man's advice than a woman's; however, this result reverses when they are informed that the woman or man has achieved a high level position in a job outside of the experiment. In the art market, Bocart, Gertsberg, and Pownall (2018) document that while female artists are less likely to transition from primary to secondary art markets, those who do command a 4.4 percent premium on artworks sold. In academia, Mengel, Sauermann, and Zölitz (2019) find that junior female instructors systematically receive lower teaching evaluations compared to male instructors for similar courses, but at the senior level, female instructors receive higher evaluations than male instructors. While these results could be driven by institutional factors, our theoretical and empirical findings suggest that the reversals may be driven by belief-based discrimination with bias (for example, biased priors or *stereotypes*). Consistent with this mechanism, Mengel, Sauermann, and Zölitz (2019) find that initial discrimination against females is higher in courses with math-related content, where distorted gender stereotypes are more likely to play a role (Coffman 2014).

The paper proceeds as follows. Section I presents the theoretical model, Section II presents the experiment and analysis of observational data, while Section III discusses the implications for policy and concludes. All proofs are in the Appendix.

I. A Dynamic Model of Discrimination

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

[4] Leslie, Manchester, and Dahm (2017) argue that this leadership premium extends to perceived potential as well. In their paper, women who are perceived to be able to rise through the ranks are judged to add more value to the company than men with similarly high potential; in contrast, low potential women are judged to add less value than low potential men. Importantly, substantially *fewer* women are judged as being able to rise through the ranks than men. Gornall and Strebulaev (2019) use a field experiment to show that promising female entrepreneurs receive significantly more interested replies from venture capitalists than male entrepreneurs pitching identical projects. See also Beaman et al. (2009) for the effect of exposure to female leaders on perceived effectiveness.

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

A. *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 $t = 1, 2, \ldots$. Each task has hidden quality $q_t = a + \epsilon_t$, where $\epsilon_t \sim N(0, 1/\tau_\epsilon)$ is an independent random shock with precision $\tau_\epsilon > 1$. Ability is fixed across time, and higher ability generates higher expected quality.

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

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

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

An evaluator is *partial* toward men if she favors male workers, either through her subjective belief about the distribution of ability by gender, which we refer

to as *belief-based partiality*, or through preferences, which we refer to as *preference-based partiality*. In the first case, an evaluator has a "taste" favoring male workers, meaning that she has a disamenity value associated with tasks produced by female workers.

DEFINITION 1 (Preference-Based Partiality): *An evaluator of type θ_i has preference-based partiality toward men if $c^i_F > 0$.*

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

DEFINITION 2 (Belief-Based Partiality): *An evaluator of type θ_i has belief-based partiality toward men if $\hat{\mu}^i_M > \hat{\mu}^i_F$. This partiality is unbiased if $\hat{\mu}^i_M = \mu_M$ and $\hat{\mu}^i_F = \mu_F$, and otherwise is biased.*

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

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

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

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

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

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

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

$$v_i(h,s,g) = \hat{E}_i[q \,|\, h,s,g] - c_g^i. \tag{2}$$

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

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

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

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

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

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

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

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

In the following sections, we explore how the different forms of partiality impact discrimination. We use these insights to illustrate how observable behavior (i.e.,

[6] Given histories $h' = (v_1,\ldots,v_m)$ and $h = (v_1',\ldots,v_n')$, we say $h' \subset h$ if $m < n$ and the histories have the same first m evaluations, i.e., $v_t = v_t'$ for all $t \leq m$.

evaluations) can be used to identify the *source* of discrimination (i.e., preferences, beliefs).

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

Misspecified Models of Inference: The setup for the evaluator's model of inference builds on the framework of social learning with model misspecification developed in Bohren and Hauser (2018). This framework can capture broad classes of model misspecification, including an incorrect model of the relationship between ability and gender and an incorrect model about other evaluators' preferences or beliefs. For example, the setting where all evaluators have common knowledge that they share the same preferences and beliefs is captured by a single type θ_1, $\pi(\theta_1) = 1$, who correctly believes that all other evaluators are this type, $\hat{\pi}_1(\theta_1) = 1$. This type's subjective belief about average ability by gender may or may not be correct. We analyze the dynamic behavior in this setting in Proposition 2.

Alternatively, there may be heterogeneity in evaluators. For example, some evaluators may use a heuristic to form beliefs about the relationship between ability and gender, while other evaluators have a correct belief about average ability by gender. Evaluators who use a heuristic are likely not aware of their bias—otherwise, they would correct for it—and believe that other evaluators form beliefs in a similar manner (as in the case of the bias blind spot (Pronin, Lin, and Ross 2002) or the false consensus effect (Ross, Greene, and House 1977)). Our framework can model this setting using a type θ_1 that has an incorrect subjective belief about average ability by gender and an incorrect subjective belief that other evaluators are the same type, $\hat{\pi}_1(\theta_1) = 1$. The other type θ_2 has a correct subjective belief about the ability distribution for males and females, and can either accurately anticipate the presence of the biased type, $\hat{\pi}_2(\theta_1) = \pi(\theta_1)$, be unaware of the biased type, $\hat{\pi}_2(\theta_1) = 0$, or under- or overestimate its frequency, $\hat{\pi}_2(\theta_1) < \pi(\theta_1)$ or $\hat{\pi}_2(\theta_1) > \pi(\theta_1)$. This type could also be aware that some evaluators are biased, but not understand the exact extent of the bias. Importantly, when there is heterogeneity in evaluators' subjective beliefs about the relationship between gender and ability, then at least one type has a misspecified model of inference. We explore the dynamic behavior in this setting in Proposition 3.

Belief-Based Discrimination: Theories of belief-based discrimination have typically focused on rational, or *statistical*, discrimination, where evaluators hold correct beliefs about aggregate group differences. These models fall into two broad categories that differ primarily in how group differences in beliefs arise, whether (i) group differences are exogenous and discrimination is due to imperfect information (Phelps 1972), or (ii) group differences are "self-fulfilling" and discrimination is an equilibrium effect (Arrow 1973).[7] In the first class of models, evaluators hold prior beliefs about workers' abilities that differ by group identity and use these group statistics to infer individual ability (Altonji and Pierret 2001, Aigner and Cain 1977,

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

Lundberg and Startz 1983). Our model with correctly specified evaluators falls into this class. In the second class of models, ex ante *identical* workers decide whether to engage in costly and unobservable skill acquisition. Discrimination arises when workers from different groups coordinate on equilibria with different levels of skill acquisition (Coate and Loury 1993, Fryer 2007). In contrast to the first class of models, there are also always equilibria in which both men and women acquire the same level of skill and evaluators treat them identically.

Belief-based discrimination can also arise from systematically incorrect, or *biased*, beliefs. Here, an evaluator engages in *inaccurate* statistical discrimination because her decision is based on a misspecified model of group differences in the distributions of ability (see Bohren et al. 2019 for evidence). Several models provide microfoundations for how such biased beliefs about group differences can arise and persist. Evaluators may form biased stereotypes of ability as a result of using the representative heuristic that exaggerates empirical reality (Bordalo et al. 2016), due to selective attention that discounts how, for example, context affects behavior (Schwartzstein 2014), or because of coarse categorization of experiences with a particular group (Fryer and Jackson 2008). In our setting, such stereotyping corresponds to distortions in the subjective belief about average ability, $\hat{\mu}_g$. As also noted in Schwartzstein (2014), the discrimination literature has tended to classify discrimination driven by distorted stereotypes as taste-based.[8] However, we demonstrate that biased beliefs lead to patterns of discrimination that substantially differ from those that arise in taste-based models in which evaluators have animus toward a particular group (i.e., preference-based partiality). This is one reason we clearly distinguish between discrimination due to incorrect beliefs and discrimination due to preferences.

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

We model the level of subjectivity in judgment as the precision of the signal of quality, τ_η. Factors that increase subjectivity, such as uncertainty over the evaluation criteria and noisier information sources, decrease the precision of the signal.

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

Our theoretical results match the empirical findings on subjective judgment: we will show that a decrease in signal precision leads to greater reliance on beliefs about group statistics to assess quality, and therefore, greater scope for belief-based discrimination.

B. *Initial Discrimination*

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

Consider the evaluation of the first task from a worker of gender g by an evaluator who has subjective prior beliefs $(\hat{\mu}_F, \hat{\mu}_M)$ about average ability, preference parameter c_F, and observes signal s_1. Given these prior beliefs about ability, the evaluator's prior belief about quality is normally distributed with mean $\hat{\mu}_g$ and precision $\tau_q \equiv \tau_a \tau_\epsilon / (\tau_a + \tau_\epsilon)$, i.e., $q_1 \sim N(\hat{\mu}_g, 1/\tau_q)$. The initial signal has conditional distribution $s_1 \,|\, q_1 \sim N(q_1, 1/\tau_\eta)$. Given the prior belief and signal distribution, the evaluator's posterior belief about quality conditional on observing s_1 is also normally distributed, $q_1 \,|\, s_1 \sim N\left(\frac{\tau_q \hat{\mu}_g + \tau_\eta s_1}{\tau_q + \tau_\eta}, \frac{1}{\tau_q + \tau_\eta}\right)$. From (2), the optimal evaluation is equal to

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

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

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

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

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

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

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

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

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

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

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

C. *Dynamics of Discrimination*

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

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

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

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

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

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

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

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

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

A key feature of social learning settings, such as our model, is the endogenous informational content of evaluations. In particular, the signal required to receive a given evaluation is decreasing in the prior belief about average ability, as shown in (6). Therefore, the prior belief about average ability impacts the posterior distribution of ability through two channels, which move in opposite directions: (i) the prior distribution of ability directly enters the Bayesian update, and this distribution has an increasing monotone likelihood ratio in its mean; and (ii) the distribution of the

signal that yields a given evaluation has a decreasing monotone likelihood ratio in the prior belief about average ability. The proof of Proposition 2 lies in establishing that the first effect dominates. Therefore, the posterior mean is increasing in the prior mean, so if evaluators believe that males have a higher prior mean than females, then following a given signal, they also believe that males have a higher posterior mean.[10]

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

Suppose there are two types of evaluators. The first type θ_1 uses a heuristic to form beliefs about the relationship between ability and gender, which leads to belief-based partiality that favors men, $\hat{\mu}_F^1 < \hat{\mu}_M^1$. This type is not aware of its bias, and believes that other evaluators have the same beliefs, $\hat{\pi}_1(\theta_1) = 1$, such as in the case of the bias blind spot (Pronin, Lin, and Ross 2002) or false consensus effect (Ross, Greene, and House 1977). With probability $p \in (0,1)$, an evaluator is type θ_1. We refer to this type as the *heuristic* type. The second type θ_2 has no belief-based partiality, $\hat{\mu}_F^2 = \hat{\mu}_M^2$, but is aware that some evaluators do; it has a correctly specified model of the type distribution, $\hat{\pi}_2(\theta_1) = p$. We refer to this type as the *impartial* type. To close the model, assume that both types have the same belief about the average ability of male workers, $\hat{\mu}_M^1 = \hat{\mu}_M^2$, and let $\hat{\mu}_M$ denote this belief. Importantly, this heterogeneity in the subjective belief about the average ability of female workers implies that at least one type has incorrect beliefs.[11]

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

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

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

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

Following evaluation v_1, let $\hat{\mu}_F^i(v_1)$ denote an evaluator of type θ_i's posterior belief about the average ability of a female worker and $\hat{\mu}_M(v_1)$ denote the posterior belief about the average ability of a male worker (which is the same for both types).

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

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

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

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

Increasing the prevalence of heuristic evaluators impacts second period discrimination through two channels. First, it increases the difference in the impartial type's second period beliefs about the average ability of a male and a female. This is because a larger share of heuristic evaluators means that it is more likely that the female faced initial discrimination and received the higher signal required to receive a given evaluation from the heuristic type, rather than the lower signal required to receive this evaluation from the impartial type. Second, it increases the probability that the second period evaluator is a heuristic type. Since the heuristic type still discriminates against females in the second period, it is more likely that a female will continue to face discrimination. The first effect dominates for low p, while the latter effect dominates for high p. This leads to a non-monotonicity in how second period discrimination changes with respect to p. Further, discrimination is always zero at $p = 0$, as all evaluators are impartial, and discrimination is always positive at $p = 1$, as this corresponds to a single type of evaluator with belief-based partiality. Figure 1 illustrates the possibility of a reversal.

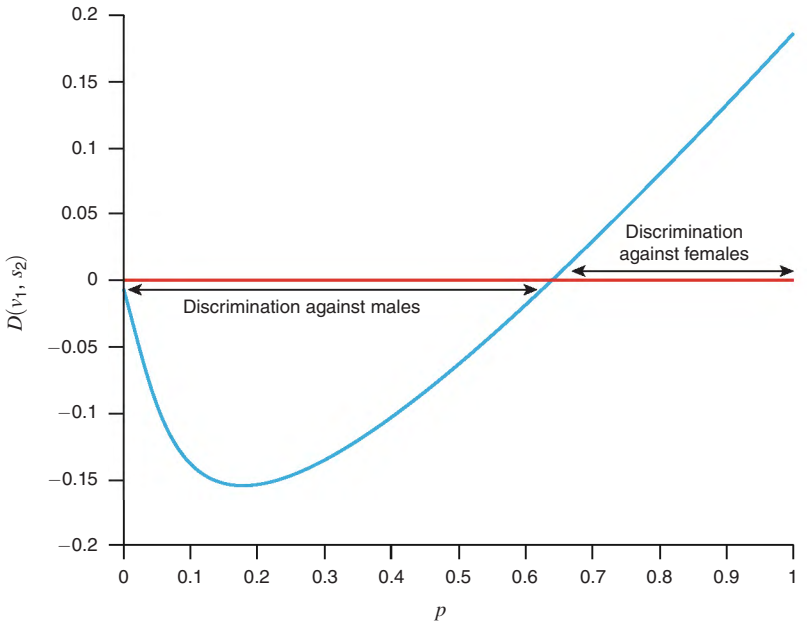

FIGURE 1. SECOND PERIOD DISCRIMINATION, AS A FUNCTION OF THE PROPORTION OF HEURISTIC EVALUATORS

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

D. *Discussion of Results*

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

Robustness

Alternative Distributions of Ability: We combine a partial analytical derivation with numerical analysis to illustrate that the impossibility of a reversal in the correctly specified model (Proposition 2) is robust to other distributions of ability, including the beta distribution, exponential distribution, gamma distribution and a setting with

binary ability and quality (see online Appendix A).[12] The key feature that drives the impossibility of a reversal is that the ability distribution satisfies the monotone likelihood ratio property (MLRP) with respect to the parameter that varies by gender. The MLRP is commonly assumed in information economics (Milgrom 1981) and holds for many other families of distributions. Propositions 1 and 3 are also robust to alternative distributional assumptions. It is straightforward to extend Proposition 1 analytically. By similar intuition to Proposition 3, it is possible to generate reversals in misspecified models with other ability distributions.

Coarse Evaluations: We assume that the space of possible evaluations is isomorphic to the space of beliefs about expected quality. In reality, the space of possible evaluations may be coarser than the evaluator's belief about expected quality. For example, the evaluator may only be able to accept or reject a task, or rate it on a scale of 1 to 5. When this is the case, it will not be possible to perfectly infer the signal an evaluator observed from the reported evaluation and information will be lost. In online Appendix B.1, we show that an analogue of Proposition 2 holds for coarse evaluations. In particular, a discrimination reversal does not occur between the first and second period when evaluators have common knowledge of the same beliefs about the distribution of ability for men and women.

Shifting Standards: Another relevant feature for our setting is how the standard of evaluation may change with respect to reputation. Higher reputation often leads to increased responsibilities and privileges, which require greater ability to manage effectively. As such, individuals may be subject to increasingly higher benchmarks as their level of seniority increases to avoid erroneously granting responsibility to someone who is unprepared. Our framework can easily be adapted to capture shifting standards (Biernat, Vescio, and Manis 1998) with respect to reputation. We say a worker faces *shifting standards* if, conditional on receiving a positive initial evaluation, the worker faces a stricter standard in the second period: a higher signal is required to receive a given evaluation, relative to the signal required for the same evaluation in the first period. We explore this extension in online Appendix B.2.

Alternative Models

Attrition: Suppose that workers exit the worker pool with positive probability after completing each task, and lower ability workers exit at a higher rate than higher ability workers. In this case, the content and the *length* of the worker's evaluation history provide information about ability. If male workers exit at a lower rate than female workers, conditional on sharing similar evaluation histories, then the length of the evaluation history has different informational content for male and female workers. If evaluators' subjective prior beliefs about ability favor males, then this differential attrition will shrink these initial differences. It can even lead to a reversal

[12] Analytical results are possible for the normal distribution, as a normal ability distribution is the conjugate prior for a normal signal distribution. This means that the posterior distribution of ability is also normally distributed, which allows for a recursive representation of the belief-updating process and a closed-form characterization of the evolution of beliefs. When the conjugate prior property does not hold, as is the case for the other distributions we consider, a combination of analytical and numerical analysis is necessary.

when low-ability women exit at a fast enough rate that distributions following longer histories favor women. In contrast to Proposition 3, such a reversal can occur even when all evaluators have correctly specified models.

In Section IID we empirically test for differential attrition by gender. We use observational data from the forum to show that males and females with similar evaluation histories leave the market at similar rates. Therefore, there is no evidence for differential attrition in our experimental setting. Evaluators may incorrectly believe that attrition differs by gender, but in this case, they have a misspecified model.

Differential attrition may drive discrimination reversals in other labor market settings. It is therefore important to empirically measure attrition in order to rule out differential attrition as a potential driver of the reversal. Our empirical analysis demonstrates how one could measure attrition based on observable data from the setting of interest and test whether it differs by gender.

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

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

Self-Fulfilling Beliefs: As discussed in Section IA, self-fulfilling beliefs are another form of belief-based discrimination. Fryer (2007) explores how

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

II. A Field Experiment

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

A. *Description of Forum*

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

The forum offers written guidelines for evaluating posts, and these guidelines are actively discussed on the forum's message boards. Voting is meant to serve a dual purpose: (i) upvoting is meant to highlight a quality post while downvoting is meant to discourage low quality posts, and (ii) upvoting rewards the *user* for a high quality

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

[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.

post while downvoting punishes him or her for a low-quality post.[16] The second point stems from the fact that users earn publicly observable reputation points from the votes they receive for their posts. An upvote earns 5 reputation points on questions and 10 reputation points on answers, while a downvote deducts 2 reputation points for both questions and answers.[17] Reputation unlocks privileges, such as the ability to edit and comment on others' posts or tag questions as duplicates. It can also be used as a currency through the assignment of "bounties": users can *spend* their reputation points to post a question with a bounty that will be awarded to the highest quality answer, as determined by the question poster, to increase the quality of answers.

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

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

B. *Experimental Design*

The ability to exogenously vary the gender and reputation associated with a user makes this an ideal setting for testing the dynamic predictions of different sources of discrimination. Comparing evaluations of question and answer posts allow us to test the predictions of how discrimination varies with the level of subjectivity in judgment.

Posting Questions.—We generated a series of original mathematics questions and posted them under male and female usernames on accounts with low and high reputations. We opened 280 new accounts, with 140 male usernames and 140 female

[16] While we use votes as our primary criterion of differential evaluation, discrimination can also manifest in less objectively measurable ways, such as the language used in response to a post. In a companion paper (Bohren, Imas, and Rosenberg 2018), we use natural language processing to document systematic differences in the ways in which users respond to male versus female posters.

[17] It is not possible for a user's reputation to fall below 1.

usernames.[18] Each account was associated with its own email address, username, and password. Of these accounts, 140 (70 with female usernames and 70 with male usernames) were left as new accounts; these comprised the Novice accounts. For the other 140 accounts (70 male and 70 female), we manually built up the reputation to the top twenty-fifth percentile of reputation on the forum; at the time of the experiment, this corresponded to a reputation of at least 100. Research assistants earned reputation on each account by posting content until the accumulated reputation reached 100. Once an account reached at least 100, the research assistant stopped posting content. Because reputation was accumulated through the actions (votes) of other users on the forum, we could not control the exact number of reputation points associated with each account: the mean reputation on these accounts was $M = 155.23$. These accounts comprised the Advanced accounts.

Critically, upon achieving a high reputation, we re-randomized the gender of the username on the Advanced accounts: 35 accounts that were built up under male usernames were switched to female, and 35 female accounts were switched to male; the remaining 70 accounts received a new username of the same gender. Importantly, when a username is switched, all past and future activity on the account became associated with the new username. That is, all *previous* posts now reflect the new username, and no public record of the name change is available. Re-randomizing the gender of the usernames avoids issues of endogeneity associated with, for example, female accounts requiring different quality posts to achieve the same level of reputation as male accounts. After reassigning usernames, the new female and male accounts had similar reputation levels ($M = 155.89$ versus $M = 154.57$, respectively, $p = 0.82$).

Our goal was to write high-quality questions that would be well received on the forum. Content on the forum ranges from high school arithmetic to upper-level graduate mathematics. Questions are tagged by topic, e.g., real analysis, combinatorics. Users are discouraged from posting questions directly from textbooks or duplicating content that is already posted; such posts are flagged and routinely closed by moderators. In order to minimize the chance that our content was flagged, we wrote 280 novel mathematics questions ranging in level of difficulty from upper-level undergraduate to early graduate. These questions were randomly assigned to one of the four conditions: male novice, female novice, male advanced, or female advanced.

We posted questions on a predetermined schedule to avoid altering the usual activity on the forum, i.e., flooding the forum with content. Research assistants posted one question at least 20 minutes apart between 5 PM and 10 PM, Monday through Thursday. Data on the community response to the questions, e.g., upvotes, downvotes, number of answers, were collected seven days after posting for each question, both in numerical form and as screenshots. A total of 7 of the 280 questions were dropped from our analysis due to forum moderators prematurely closing the questions before the end of the seven-day window or due to errors in the posting of the questions (i.e., two questions posted to the same account).

We measure discrimination as either the average change in reputation points per post (ΔRep) or the average number of upvotes net of downvotes per post (Net Votes).

[18] 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.

The dynamic pattern of discrimination provides a test of the theoretical predictions outlined in Section I. Conditional on observing discrimination between male and female Novice accounts, a mitigation in its intensity for Advanced accounts is consistent with belief-based partiality, including the case of statistical discrimination where beliefs are correct, while a *reversal* of discrimination for Advanced accounts is evidence for biased belief-based partiality.

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

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

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

To post answers, we created a second set of 140 Novice accounts with no prior posts, split between 70 male usernames and 70 female usernames. We needed to post answers in real time, as questions on the forum are answered fairly quickly and late answers generally receive little attention. To do so, research assistants worked in pairs. One member of the pair, the "answerer," would find a newly posted question

that had not been answered yet and write an answer for it. The "answerer" would then send the answer and a link to the question to the other research assistant, the "poster," who would assign the answer to one of our accounts and post it. The order of accounts that the answer would be posted to was predetermined: known to the "poster" but not the "answerer." As such, the research assistant writing the answer did not know the gender of the account that the answer would be posted to, and therefore, could not be subconsciously influenced by whether the answer would be posted to a male or female account. As with the questions, answers were posted between 5 PM and 10 PM, Monday through Thursday. Data were collected seven days after posting the answer, both in numerical form and as a screenshot. A total of 5 of the 140 answers were dropped due to errors, e.g., the question was closed before the seven-day window concluded.

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

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

C. *Experimental Results*

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

Subjectivity of Judgment.—We first examine the change in reputation (ΔRep) for answers posted to male versus female accounts (i.e., the reputation points earned on the post). Column 1 of Table 1 shows that regressing ΔRep per answer on gender reveals no significant difference in the evaluation of answers at conventional levels. This result is illustrated in panel A of Figure 2, which shows the average ΔRep by gender, and panel B, which plots the distributions of ΔRep by gender. Column 2 of Table 1 repeats the analysis using net votes per post as the dependent variable.[19] Together, these results suggest that there is little evidence for gender discrimination on answers.

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

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

	Answers only		Questions only		Answers and questions	
	ΔRep (1)	Net votes (2)	ΔRep (3)	Net votes (4)	ΔRep (5)	Net votes (6)
Male	−1.38	−0.31	2.86	0.58	−1.38	−0.31
	(0.97)	(0.17)	(1.32)	(0.27)	(1.16)	(0.22)
Question					0.08	0.09
					(1.16)	(0.22)
Male × question					4.24	0.89
					(1.64)	(0.32)
Constant	4.60	0.79	4.68	0.88	4.60	0.79
	(0.69)	(0.12)	(0.93)	(0.19)	(0.82)	(0.16)
Observations	135	135	135	135	270	270

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

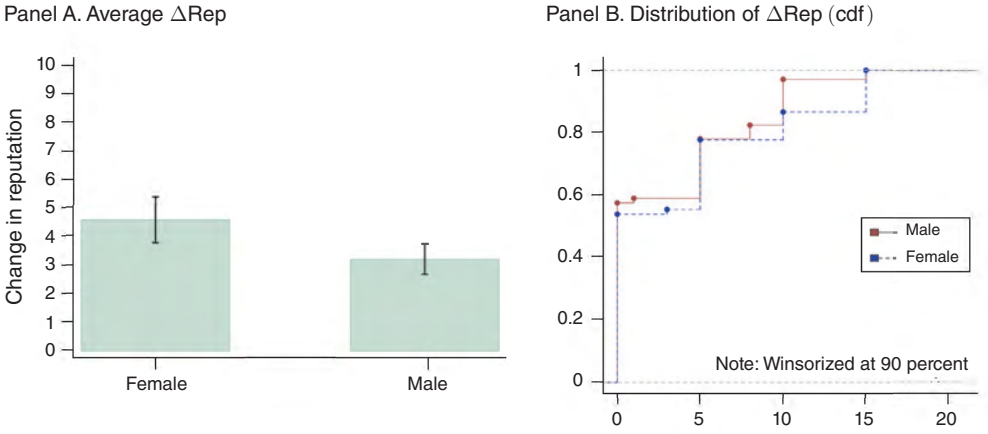

Figure 2. Change in Reputation for Answers

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

Next, we directly compare responses to answer versus question posts by gender. We first test the difference in the estimated coefficients of the male gender dummy between the question and answer regressions and find that this

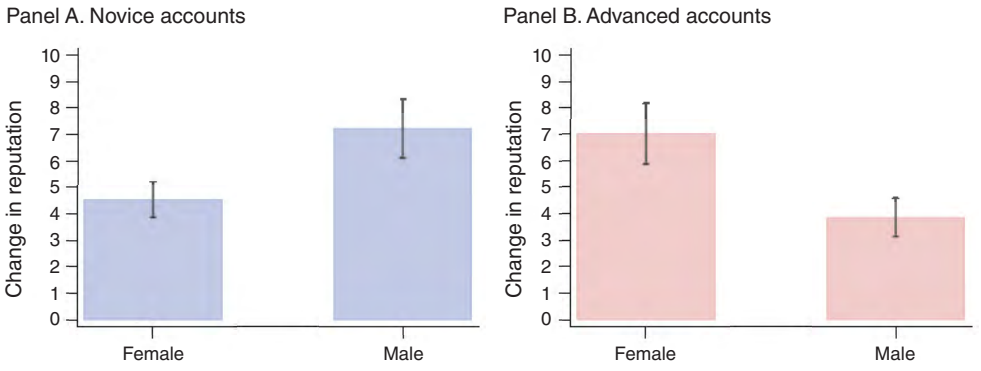

FIGURE 3. AVERAGE CHANGE IN REPUTATION FOR QUESTIONS

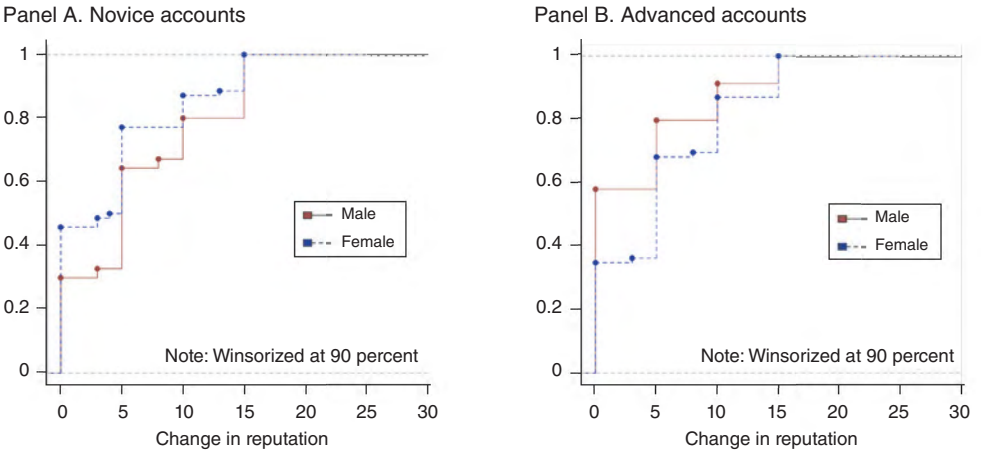

FIGURE 4. DISTRIBUTION OF CHANGE IN REPUTATION FOR QUESTIONS (CDF)

difference is significant for both ΔRep ($\chi^2(1) = 6.70$; $p = 0.01$) and net votes ($\chi^2(1) = 7.87$; $p = 0.005$). We then present regression results for question and answer posts within the same model. We regress ΔRep and net votes on dummies corresponding to gender, type of post (question or answer) and the interaction of gender and type of post (Table 1, columns 5 and 6). There is a significant mitigation of discrimination against female accounts for answers, relative to questions: the interaction effect between gender and type of post is positive and significant in both specifications. This implies that the male advantage is significantly larger for questions, compared to answers.

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

Dynamics of Discrimination.—Next, we examine the dynamics of discrimination by comparing discrimination toward novice and advanced users. As shown in panel B of Figure 3, questions posted to advanced female accounts accumulated more

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

	Advanced		Novice and advanced		
	ΔRep	Net votes	ΔRep	Net votes	Binary
	(1)	(2)	(3)	(4)	(5)
Male	−3.16	−0.62	2.86	0.58	0.17
	(1.37)	(0.28)	(1.36)	(0.27)	(0.08)
Advanced			2.33	0.49	0.09
			(1.35)	(0.27)	(0.08)
Male × advanced			−6.02	−1.20	−0.40
			(1.91)	(0.38)	(0.11)
Constant	7.01	1.38	4.68	0.88	0.56
	(0.97)	(0.20)	(0.96)	(0.19)	(0.06)
Observations	138	138	273	273	273

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

reputation points, ΔRep, than those posted to advanced male accounts. This contrasts with questions posted on novice female accounts, which accumulated fewer reputation points than those posted to novice male accounts (panel A of Figure 3). In other words, we observe a dynamic *reversal* of discrimination between novice and advanced accounts: questions from male users are favored at low reputations, while questions from female users are favored at high reputations. Figure 4 illustrates this reversal in the distributions of ΔRep: panel A shows the distribution of ΔRep on questions posted to novice accounts, while panel B shows the distribution of ΔRep for advanced accounts.

For advanced accounts, regressing ΔRep or net votes per question on the gender of the poster reveals that questions posted to female accounts accumulated significantly more reputation points and net votes than questions posted to male accounts (Table 2, columns 1 and 2, respectively). These differences in evaluation correspond to roughly 0.6 standard deviations for both ΔRep and net votes. This contrasts with the significantly lower evaluation of questions posted to novice female accounts relative to novice male accounts, as reported in Table 1. Testing the difference in the estimated coefficients of the male gender dummy between the Novice and Advanced regressions reveals a significant difference for both ΔRep ($\chi^2(1) = 10.05; p = 0.002$) and net votes ($\chi^2(1) = 9.88; p = 0.002$).

Columns 3 and 4 of Table 2 present regression results for Novice and Advanced accounts within the same model. In column 3, we regress ΔRep on dummies corresponding to the gender of the poster, the reputation level of the poster (novice or advanced), and their interaction. The interaction between gender and reputation level is negative and significant, confirming the reversal of discrimination between the Novice and Advanced accounts. The same pattern of results holds for the net votes earned per question (column 4). To ensure that these results are not driven by outliers or subsequent voters herding on the first upvote, we replicate the analysis using a binary variable that is equal to 1 if the question receives at least one upvote, and 0 otherwise. As shown in column 5, the results are robust to this binary

specification.[20] Consistent with shifting standards, the average change in reputation and average number of net votes, pooled across both genders, does not significantly differ between Novice and Advanced accounts.

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

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

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

We also explored whether the users who evaluated questions in our experiment are similar to the users who evaluated answers. To determine whether users specialize in the type of content they evaluate by either evaluating mostly questions or mostly answers, or whether most users evaluate both, we tabulated each user's total number of votes by content type, and calculated the proportion of a given user's votes that were cast on questions versus answers. The proportions are very similar: on average, 48 percent of a user's votes were cast on questions and 52 percent were cast on answers, with a standard deviation of 0.21. This suggests that most users evaluated questions and answers in fairly equal proportions. We also examined whether the users who evaluated our content differed in their reputation levels and inferred genders, depending on the type of post.[21] Summary statistics are presented in online Appendix C.3; we found no significant differences in the characteristics of voters evaluating different types of posts.

D. *Observational Data*

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

[20] Results are also robust to winsorizing the dependent variable at 5 percent or 10 percent.
[21] Section IID outlines the process of inferring gender from a username.

Description of Data.—The observational dataset is compiled and made 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 from July 2010 to March 2017. We excluded all content posted as part of our experiment. To code gender, we ran an algorithm developed by Vasilescu, Capiluppi, and Serebrenik (2014) to classify the gender of the usernames (see online Appendix C.4.1 for a description of this algorithm). Each username is classified as "male," "female," or "x" (when gender cannot be inferred). In our sample, the gender was resolved for 55 percent of accounts, which we used in the analyses. Of these accounts, 19 percent were classified as "female" and the remaining 81 percent were classified as "male." Of accounts that had less than 100 reputation points, 21 percent were classified as "female"; of accounts that had between 100 and 240 reputation points—the Advanced range used in our experiment —13 percent were classified as female.

Attrition.—We studied posting behavior of users to determine whether there is differential attrition based on gender. To do so, we created a panel dataset of up to the first ten posts for each user, including whether each post was a question or an answer and the amount of reputation earned on the post.[22] For each user, we observed whether their first post was followed by a second post, whether their second post was followed by a third post, and so on. Differential attrition will lead to gender differences in the likelihood of observing a subsequent post conditional on receiving similar evaluations on the prior post.

We first examined whether there was differential attrition by gender following the first post. We ran a probit regression, regressing a dummy for whether a user generated a second post on the inferred gender of the username, the log of the reputation earned on the first post and their interaction, both pooling question and answer posts and analyzing each separately. We also split the reputation earned on the first post into quartiles and ran a similar regression. Neither the gender variable nor the interaction with reputation or reputation quartile is significant in any of these specifications. This suggests that female users were no less likely than male users to generate a second post conditional on a similar first post (see online Appendix Tables 6 and 7). We repeated a similar analysis to study differential attrition by gender following the second through ninth posts (see online Appendix Table 8). These results mirror the results following the first post: neither the gender variable nor the interaction with reputation is significant in subsequent periods.

Next, we ran a pooled analysis on all posts in our panel. In order to compare attrition rates for males and females with similar evaluation histories, we created a variable corresponding to the total reputation earned on all previous posts. For example, when looking at the likelihood of a fourth post, total reputation earned is the sum of the reputation earned on the third, second, and first posts. We ran a probit regression, regressing a dummy for whether a user generated a post t on the gender dummy, the log of the total reputation earned on posts 1 through $t - 1$, their

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

interaction, and a dummy of whether the previous post was a question. In two of the three specifications, we also controlled for how many posts it took to generate this total reputation. Neither the gender variable nor the interaction with total reputation is significant in any of these specifications (see online Appendix Table 9).

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

Variance.—The observational data also allow us to examine whether there are differences in variances of the ability distributions by gender. Since we did not find evidence for discrimination on answers, we use the evaluations of answers posted to new accounts to proxy for underlying ability. We then examine whether there are differences in the variance of these evaluations by gender. Running Levene's test of equal variances on the distributions of reputation points per first answer post (ΔRep) reveals no significant differences by gender ($p = 0.41$ using the mean, $p = 0.48$ using the median, $p = 0.46$ using the 10 percent trimmed mean).

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

Gender Differences in Evaluations.—We also use the observational data to examine how evaluations of posts vary with reputation, inferred gender of the user and type of post. As in our experiment, we focus on the evaluation of questions posted to novice and advanced accounts, and the evaluation of answers posted to novice accounts. We define posting to novice and advanced accounts similar to the experiment (see online Appendix C.4.4 for details).

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

Keeping these caveats in mind, we run regression analogous to Tables 1 and 2 using the reputation points earned per post (ΔRep) as the dependent variable. These results are presented in online Appendix Tables 10–12. The evaluation patterns by gender across the different types of posts are similar to those documented in the experiment, although the effect sizes vary depending on the specification. We document three main findings: (i) no significant evidence of gender discrimination on answers, (ii) questions posted by novice accounts with female usernames tend to earn fewer reputation points than those posted by novice accounts with male usernames, and (iii) questions posted to advanced accounts with female usernames tend to earn *more* reputation points than those posted to advanced accounts with male usernames.

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

III. Discussion and Conclusion

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

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

Taken together, our empirical results are consistent with discrimination driven by belief-based partiality with some form of misspecification.

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

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

Appendix

Throughout this section, let $\tau_{\epsilon\eta} \equiv \tau_\eta \tau_\epsilon / (\tau_\eta + \tau_\epsilon)$ denote the precision of the signal conditional on ability. We will also use the notation $\tau_a(t) \equiv \tau_a + (t-1)\tau_{\epsilon\eta}$ and $\tau_q(t) \equiv \tau_a(t)\tau_\epsilon / (\tau_a(t) + \tau_\epsilon)$, which we show in Lemma 1 denotes the precision of the belief about ability and quality, respectively, at time t.

PROOF OF PROPOSITION 1:

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

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

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

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

LEMMA 1: *Suppose an evaluator has a normally distributed subjective prior distribution of ability with mean $\hat{\mu}$ and precision τ_a for a worker of gender g, has no*

preference-based partiality, and believes that all other evaluators are also this type. Then following any history h_t for $t \geq 2$, the subjective posterior distribution of ability $f_a(a \mid h_t)$ is normally distributed with mean

$$\hat{\mu}(h_t) \equiv \frac{\tau_a \hat{\mu} + \tau_{\epsilon\eta} \sum_{n=1}^{t-1} s\big(v_n, \hat{\mu}(h_n), n\big)}{\tau_a + (t-1)\tau_{\epsilon\eta}} \tag{A1}$$

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

$$s\big(v, \hat{\mu}(h_t), t\big) \equiv \left(\frac{\tau_q(t) + \tau_\eta}{\tau_\eta}\right) v - \left(\frac{\tau_q(t)}{\tau_\eta}\right) \hat{\mu}(h_t) \tag{A2}$$

is the signal required to receive evaluation v at time t when the evaluator has belief $\hat{\mu}(h_t)$ and $\tau_q(t) \equiv \tau_a(t)\tau_\epsilon / \big(\tau_a(t) + \tau_\epsilon\big)$ is the precision of quality at time t.

PROOF:

Let $f_a(a \mid h)$ denote the subjective posterior distribution of ability following history h and $f_s(s \mid a)$ denote the signal distribution conditional on ability.

Suppose an evaluator has a normally distributed prior distribution of ability $f_a(a \mid h_1)$, with mean $\hat{\mu}$ and precision τ_a, no preference-based partiality, and believes that all other evaluators are also this type. Recall $s_t = a + \epsilon_t + \eta_t$, so $f_s(s \mid a)$ is normally distributed with mean a and precision $\tau_{\epsilon\eta}$. From (4), conditional on observing signal s_1, the first evaluation is

$$v_1 = \frac{\tau_q \hat{\mu} + \tau_\eta s_1}{\tau_q + \tau_\eta}. \tag{A3}$$

Therefore, the distribution of first period evaluations conditional on ability, denoted $f_v(v \mid a, h_1)$, is normally distributed with mean $\frac{\tau_q \hat{\mu} + \tau_\eta a}{\tau_q + \tau_\eta}$ and precision $\big(\tau_q + \tau_\eta\big)^2 \tau_{\epsilon\eta} / \tau_\eta^2$. It is possible to back out s_1 from observing v_1, $s_1 = s(v_1, \hat{\mu}, 1)$, where from (A2),

$$s(v_1, \hat{\mu}, 1) = \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) v_1 - \left(\frac{\tau_q}{\tau_\eta}\right) \hat{\mu}. \tag{A4}$$

Consider the posterior distribution of ability following history $h_2 = (v_1)$. From Bayes' rule,

$$f_a(a \mid h_2) = \frac{f_v(v_1 \mid a, h_1) f_a(a \mid h_1)}{\int_{-\infty}^{\infty} f_v(v_1 \mid a', h_1) f_a(a' \mid h_1)\, da'} = \frac{f_s\big(s(v_1, \hat{\mu}, 1) \mid a\big) f_a(a \mid h_1)}{\int_{-\infty}^{\infty} f_s\big(s(v_1, \hat{\mu}, 1) \mid a'\big) f_a(a' \mid h_1)\, da'},$$

where the second equality follows from $f_v(v \mid a, h_1) = \left(\frac{\tau_q + \tau_\eta}{\tau_\eta}\right) f_s\big(s(v, \hat{\mu}, 1) \mid a\big)$. The normal distribution is conjugate to itself for a normal likelihood function. Since $f_a(a \mid h_1)$ and $f_s(s \mid a)$ are normal, $f_a(a \mid h_2)$ is also normal with mean $\hat{\mu}(h_2) = \frac{\tau_a \hat{\mu} + \tau_{\epsilon\eta} s(v_1, \hat{\mu}, 1)}{\tau_a + \tau_{\epsilon\eta}}$ and precision $\tau_a(2) = \tau_a + \tau_{\epsilon\eta}$.

Given the normality of the posterior belief about ability, we can define the evaluation and belief-updating processes recursively. Suppose that the distribution of ability following history h_t is normally distributed with mean $\hat{\mu}(h_t)$ and precision

$\tau_a(t)$. The evaluation process in period $t > 1$ is analogous to $t = 1$. The distribution of quality q_t conditional on observing signal s_t and history h_t is normal with mean $\hat{E}[q_t|h_t, s_t] = \frac{\tau_q(t)\hat{\mu}(h_t) + \tau_\eta s_t}{\tau_q(t) + \tau_\eta}$ and precision $\tau_q(t) + \tau_\eta$. Analogous to (A3), conditional on observing signal s_t, the evaluation in period t is equal to

$$v_t = \frac{\tau_q(t)\hat{\mu}(h_t) + \tau_\eta s_t}{\tau_q(t) + \tau_\eta}. \tag{A5}$$

Inverting this expression, the signal required to receive evaluation v_t is

$$s(v_t, \hat{\mu}(h_t), t) \equiv \left(\frac{\tau_q(t) + \tau_\eta}{\tau_\eta}\right) v_t - \left(\frac{\tau_q(t)}{\tau_\eta}\right)\hat{\mu}(h_t), \tag{A6}$$

which is equal to (A2). Belief-updating is also analogous to $t = 1$. For $t > 1$, the posterior distribution of ability following history $h_{t+1} = (h_t, v_t)$ is normally distributed with mean

$$\hat{\mu}(h_{t+1}) \equiv \frac{\tau_a(t)\hat{\mu}(h_t) + \tau_{\epsilon\eta} s(v_t, \hat{\mu}(h_t), t)}{\tau_a(t) + \tau_{\epsilon\eta}} \tag{A7}$$

and precision $\tau_a(t+1) = \tau_a(t) + \tau_{\epsilon\eta}$.

Initializing $\hat{\mu}(h_1) = \hat{\mu}$ and $\tau_a(1) = \tau_a$, and solving the recursive expressions for $\hat{\mu}(h_t)$ and $\tau_a(t)$ yields solution

$$\hat{\mu}(h_t) = \frac{\tau_a\hat{\mu} + \tau_{\epsilon\eta}\sum_{n=1}^{t-1} s(v_n, \hat{\mu}(h_n), n)}{\tau_a + (t-1)\tau_{\epsilon\eta}}, \tag{A8}$$

$$\tau_a(t) = \tau_a + (t-1)\tau_{\epsilon\eta}, \tag{A9}$$

for $t > 1$. Therefore, when the prior distribution of ability is normal, the distribution of ability following history h_t, i.e., $f_a(a \mid h_t)$, is also normal with mean $\hat{\mu}(h_t)$ and precision $\tau_a(t)$. ∎

PROOF OF PROPOSITION 2:

Suppose there is a single type of evaluator with belief-based partiality and no preference-based partiality. Given initial beliefs $\hat{\mu}_F < \hat{\mu}_M$, let $\hat{\mu}_F(h_t)$ and $\hat{\mu}_M(h_t)$ denote the subjective average ability of a female and male worker, respectively, following history h_t. We proceed by a series of lemmas.

LEMMA 2: *If $\hat{\mu}_F(h_t) < \hat{\mu}_M(h_t)$, then for all v_t, there is no belief reversal between periods t and $t+1$, $\hat{\mu}_F(h_{t+1}) < \hat{\mu}_M(h_{t+1})$, and the difference in beliefs about average ability decreases between periods t and $t+1$, $\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) < \hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)$.*

PROOF:

Suppose $\hat{\mu}_F(h_t) < \hat{\mu}_M(h_t)$. The difference in signals required for a male and a female worker to receive evaluation v_t is

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

where $s(v, \hat{\mu}, t)$ is defined in (A2). Therefore, given $\hat{\mu}_g(h_{t+1})$ defined in (A7), the difference in the belief about posterior average ability following evaluation v_t is

$$\begin{aligned}(\text{A10}) \quad \hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) &= \left(\frac{\tau_a(t)}{\tau_a(t) + \tau_{\epsilon\eta}}\right)\left(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)\right) \\ &\quad + \left(\frac{\tau_{\epsilon\eta}}{\tau_a(t) + \tau_{\epsilon\eta}}\right)\left(s(v_t, \hat{\mu}_M(h_t), t) - s(v_t, \hat{\mu}_F(h_t), t)\right) \\ &= \left(\frac{\tau_a(t) - \tau_{\epsilon\eta}\tau_q(t)/\tau_\eta}{\tau_a(t) + \tau_{\epsilon\eta}}\right)\left(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)\right).\end{aligned}$$

This is positive if $\tau_a(t) - \tau_{\epsilon\eta}\tau_q(t)/\tau_\eta > 0$, which is equivalent to $\tau_\epsilon^2 + \tau_\epsilon\tau_\eta + \tau_a(t)(\tau_\epsilon + \tau_\eta) > \tau_\epsilon^2$, which always holds since all precisions are positive. Therefore, $\hat{\mu}_M(h_{t+1}) > \hat{\mu}_F(h_{t+1})$, which establishes part (1). From (A10), $\hat{\mu}_M(h_{t+1}) - \hat{\mu}_F(h_{t+1}) < \hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)$ if and only if $\frac{\tau_a(t)}{\tau_a(t) + \tau_{\epsilon\eta}} - \frac{\tau_{\epsilon\eta}\tau_q(t)}{(\tau_a(t) + \tau_{\epsilon\eta})\tau_\eta} < 1$, which always holds since the first term is less than 1 and the second term is subtracting a positive number. This establishes part (2). ∎

Therefore, given prior beliefs $\hat{\mu}_F < \hat{\mu}_M$, by Lemma 2, the difference in beliefs about average ability for males and females with the same history decreases across periods but never reverses, i.e., $\hat{\mu}_F(h_t) < \hat{\mu}_M(h_t)$ for all h_t. The next lemma establishes that this rules out discrimination reversals.

LEMMA 3: *A discrimination reversal occurs between periods t and $t+1$ if and only if the beliefs about average ability of male and female workers reverse between periods t and $t+1$.*

PROOF:

Given beliefs $\hat{\mu}_M(h_t)$ and $\hat{\mu}_F(h_t)$, from (A5), discrimination in period t is equal to

$$(\text{A11}) \qquad D(h_t, s_t) = \left(\frac{\tau_q(t)}{\tau_q(t) + \tau_\eta}\right)\left(\hat{\mu}_M(h_t) - \hat{\mu}_F(h_t)\right).$$

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. ∎

Given $\hat{\mu}_F(h_t) < \hat{\mu}_M(h_t)$ for all h_t, by Lemma 3, there is no discrimination reversal between any periods t and $t+1$. This establishes Proposition 2. ∎

PROOF OF PROPOSITION 3:

Let $f_a(a \mid h)$ denote the subjective posterior distribution of ability following history h and $f_s(s \mid a)$ denote the signal distribution conditional on ability. Suppose evaluators are the heuristic type θ_1 with probability $p \in (0, 1)$ and the impartial type θ_2 with probability $1 - p$. Type θ_1's belief about male and female ability evolve as in Lemma 1, since this type believes that all other evaluators have the same beliefs as it. Type θ_2's

belief about male ability also evolves as in Lemma 1, since both types have the same prior belief about male ability. Therefore, type θ_1 evaluates all workers and type θ_2 evaluates male workers in the same way as was previously characterized. The novelty stems from characterizing how type θ_2 evaluates female workers in the second period.

We first characterize how type θ_2 updates its belief about a female worker's ability after the first evaluation. Suppose an evaluator of type θ_2 observes evaluation v_1 in period one for a female worker. The evaluator believes that with probability p, it is from a heuristic type θ_1 who observed signal $s\big(v_1, \hat{\mu}_F^1, 1\big) = \Big(\frac{\tau_q + \tau_\eta}{\tau_\eta}\Big)v_1 - \Big(\frac{\tau_q}{\tau_\eta}\Big)\hat{\mu}_F^1$, and with probability $1-p$, it is from an impartial type θ_2 who observed signal $s\big(v_1, \hat{\mu}_M, 1\big) = \Big(\frac{\tau_q + \tau_\eta}{\tau_\eta}\Big)v_1 - \Big(\frac{\tau_q}{\tau_\eta}\Big)\hat{\mu}_M$ (recall $\hat{\mu}_F^2 = \hat{\mu}_M$). Note $s\big(v_1, \hat{\mu}_F^1, 1\big) > s\big(v_1, \hat{\mu}_M, 1\big)$. Therefore, the distribution of first period evaluations conditional on ability a is a mixture of two normal distributions,

$$f_v(v \,|\, a, h_1) \;=\; \Big(\frac{\tau_q + \tau_\eta}{\tau_\eta}\Big)\Big(p f_s\big(s\big(v, \hat{\mu}_F^1, 1\big) \,|\, a\big) + (1-p) f_s\big(s\big(v, \hat{\mu}_M, 1\big) \,|\, a\big)\Big),$$

where $f_s\big(s \,|\, a\big)$ is normally distributed with mean a and precision $\tau_{\epsilon\eta}$. Consider the posterior distribution of ability following history $h_2 = \big(v_1\big)$. Since the prior belief $f_a\big(a \,|\, h_1\big)$ is normal with mean $\hat{\mu}_M$ and the likelihood function $f_v\big(v \,|\, a, h_1\big)$ is a mixture of two normal distributions, the posterior belief will be a mixture of two normal distributions,

$$f_a\big(a \,|\, h_2\big) \;=\; p\left(\frac{C_{a,1}\big(h_2\big)}{C_a\big(h_2\big)}\right) f_{a,1}\big(a \,|\, h_2\big) + (1-p)\left(\frac{C_{a,2}\big(h_2\big)}{C_a\big(h_2\big)}\right) f_{a,2}\big(a \,|\, h_2\big),$$

where $f_{a,1}\big(a \,|\, h_2\big)$ and $f_{a,2}\big(a \,|\, h_2\big)$ denote the posterior distributions of ability conditional on observing signals $s\big(v_1, \hat{\mu}_F^1, 1\big)$ and $s\big(v_1, \hat{\mu}_M, 1\big)$, respectively, which are both normally distributed with corresponding means $\hat{\mu}_1\big(h_2\big) \equiv \frac{\tau_a \hat{\mu}_M + \tau_{\epsilon\eta} s\big(v_1, \hat{\mu}_F^1, 1\big)}{\tau_a + \tau_{\epsilon\eta}}$ and $\hat{\mu}_2\big(h_2\big) \equiv \frac{\tau_a \hat{\mu}_M + \tau_{\epsilon\eta} s\big(v_1, \hat{\mu}_M, 1\big)}{\tau_a + \tau_{\epsilon\eta}}$, respectively, and precision $1/\big(\tau_a + \tau_{\epsilon\eta}\big)$, and

$$\begin{aligned}
C_{a,1}\big(h_2\big) \;&\equiv\; \int_{-\infty}^{\infty} f_s\big(s\big(v_1, \hat{\mu}_F^1, 1\big) \,|\, a\big) f_a\big(a \,|\, h_1\big)\, da \\
&=\; \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\tau_a \tau_{\epsilon\eta}}{\tau_a + \tau_{\epsilon\eta}}} \exp\Big(-0.5\Big(\tau_a \big(\hat{\mu}_M\big)^2 + \tau_{\epsilon\eta} s\big(v_1, \hat{\mu}_F^1, 1\big)^2 \\
&\qquad\qquad - \big(\tau_a + \tau_{\epsilon\eta}\big)\hat{\mu}_1\big(h_2\big)^2\Big)\Big), \\
C_{a,2}\big(h_2\big) \;&\equiv\; \int_{-\infty}^{\infty} f_s\big(s\big(v_1, \hat{\mu}_M, 1\big) \,|\, a\big) f_a\big(a \,|\, h_1\big)\, da \\
&=\; \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\tau_a \tau_{\epsilon\eta}}{\tau_a + \tau_{\epsilon\eta}}} \exp\Big(-0.5\Big(\tau_a \big(\hat{\mu}_M\big)^2 + \tau_{\epsilon\eta} s\big(v_1, \hat{\mu}_M, 1\big)^2 \\
&\qquad\qquad - \big(\tau_a + \tau_{\epsilon\eta}\big)\hat{\mu}_2\big(h_2\big)^2\Big)\Big), \\
C_a\big(h_2\big) \;&\equiv\; p C_{a,1}\big(h_2\big) + (1-p)\, C_{a,2}\big(h_2\big)
\end{aligned}$$

are normalization coefficients.

We next characterize how type θ_2 updates its belief about a female worker's quality in period 2. Given history $h_2 = (v_1)$, let $f_q(q|h_2)$ denote the prior distribution of quality in the second period and let $f_q(q\,|\,h_2, s_2)$ denote the posterior belief about quality following h_2 and signal s_2. Recall $q_2 = a + \epsilon_2$, ϵ_2 is normally distributed and $f_a(a\,|\,h_2)$ is a mixture of two normal distributions. The convolution of a normal distribution with a mixture of two normal distributions is a mixture of two normal distributions. Therefore, $f_q(q\,|\,h_2)$ is a mixture of two normal distributions. Since the likelihood function f_s is normally distributed, posterior $f_q(q\,|\,h_2, s_2)$ is also a mixture of two normal distributions:

$$f_q(q\,|\,h_2,s_2) = p\left(\frac{C_{a,1}(h_2)\,C_{q,1}(h_2,s_2)}{C_a(h_2)\,C_q(h_2,s_2)}\right) f_{q,1}(q\,|\,h_2,s_2) + (1-p)\left(\frac{C_{a,2}(h_2)\,C_{q,2}(h_2,s_2)}{C_a(h_2)\,C_q(h_2,s_2)}\right) f_{q,2}(q\,|\,h_2,s_2),$$

where $f_{q,1}(q\,|\,h_2,s_2)$ and $f_{q,2}(q\,|\,h_2,s_2)$ are both normally distributed with corresponding means $\hat{\mu}_{q,1}(h_2,s_2) \equiv \frac{\tau_q(2)\,\hat{\mu}_1(h_2)+\tau_\eta s_2}{\tau_q(2)+\tau_\eta}$ and $\hat{\mu}_{q,2}(h_2,s_2) \equiv \frac{\tau_q(2)\,\hat{\mu}_2(h_2)+\tau_\eta s_2}{\tau_q(2)+\tau_\eta}$, respectively, and precision $1/(\tau_q(2)+\tau_\eta)$, and

$$C_{q,1}(h_2,s_2) \equiv \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\tau_q(2)\,\tau_\eta}{\tau_q(2)+\tau_\eta}}\exp\Big(-0.5\Big(\tau_q(2)\,\hat{\mu}_1(h_2)^2+\tau_\eta(s_2)^2 - \big(\tau_q(2)+\tau_\eta\big)\hat{\mu}_{q,1}(h_2,s_2)^2\Big)\Big),$$

$$C_{q,2}(h_2,s_2) \equiv \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\tau_q(2)\,\tau_\eta}{\tau_q(2)+\tau_\eta}}\exp\Big(-0.5\Big(\tau_q(2)\,\hat{\mu}_2(h_2)^2+\tau_\eta(s_2)^2 - \big(\tau_q(2)+\tau_\eta\big)\hat{\mu}_{q,2}(h_2,s_2)^2\Big)\Big),$$

$$C_q(h_2,s_2) \equiv p\left(\frac{C_{a,1}(h_2)\,C_{q,1}(h_2,s_2)}{C_a(h_2)}\right) + (1-p)\left(\frac{C_{a,2}(h_2)\,C_{q,2}(h_2,s_2)}{C_a(h_2)}\right)$$

are normalization coefficients.

Finally, we compute aggregate discrimination in period two. Consider a female worker with history $h_2 = (v_1)$ who generates signal s_2 in the second period. The impartial type θ_2 gives this worker evaluation

$$v_2(h_2,s_2,F) = \left(\frac{\tau_q(2)}{\tau_q(2)+\tau_\eta}\right)\gamma(h_2,s_2) + \left(\frac{\tau_\eta}{\tau_q(2)+\tau_\eta}\right)s_2,$$

where

$$\gamma(h_2,s_2) \equiv p\left(\frac{C_{a,1}(h_2)\,C_{q,1}(h_2,s_2)}{C_a(h_2)\,C_q(h_2,s_2)}\right)\hat{\mu}_1(h_2) + (1-p)\left(\frac{C_{a,2}(h_2)\,C_{q,2}(h_2,s_2)}{C_a(h_2)\,C_q(h_2,s_2)}\right)\hat{\mu}_2(h_2).$$

The heuristic type θ_1 gives this worker evaluation

$$v_1(h_2, s_2, F) = \left(\frac{\tau_q(2)}{\tau_q(2)+\tau_\eta}\right)\hat{\mu}_F^1(h_2) + \left(\frac{\tau_\eta}{\tau_q(2)+\tau_\eta}\right)s_2,$$

where from Lemma 1, $\hat{\mu}_F^1(h_2) = \frac{\tau_a \hat{\mu}_F^1 + \tau_{\epsilon\eta} s(v_1, \hat{\mu}_F^1, 1)}{\tau_a + \tau_{\epsilon\eta}}$. Both types give males evaluation

$$v(h_2, s_2, M) = \left(\frac{\tau_q(2)}{\tau_q(2)+\tau_\eta}\right)\hat{\mu}_M(h_2) + \left(\frac{\tau_\eta}{\tau_q(2)+\tau_\eta}\right)s_2,$$

where from Lemma 1, $\hat{\mu}_M(h_2) = \frac{\tau_a \hat{\mu}_M + \tau_{\epsilon\eta} s(v_1, \hat{\mu}_M, 1)}{\tau_a + \tau_{\epsilon\eta}}$. Therefore, aggregate discrimination in the second period is equal to

$$\begin{aligned} D(h_2, s_2) &= v(h_2, s_2, M) - p\, v_1(h_2, s_2, F) - (1-p)\, v_2(h_2, s_2, F) \\ &= \left(\frac{\tau_q(2)}{\tau_q(2)+\tau_\eta}\right)\left(\hat{\mu}_M(h_2) - p\hat{\mu}_F^1(h_2) - (1-p)\gamma(h_2, s_2)\right). \end{aligned}$$

Aggregate discrimination reverses at (h_2, s_2) if $D(h_2, s_2) < 0$. We know that at $p = 0$, $D(h_2, s_2) = 0$, as this is the case with no partiality, and at $p = 1$, $D(h_2, s_2) > 0$, as this is the case with a single type of evaluator with belief-based partiality from Proposition 2. Therefore, if the derivative of $D(h_2, s_2)$ with respect to p is negative at $p = 0$, discrimination will become negative for an interval $(0, \bar{p})$ before becoming positive. This derivative simplifies to showing that

$$1 < \left(\frac{\tau_\epsilon^2}{(\tau_\epsilon+\tau_\eta)(\tau_a+\tau_\epsilon)}\right)\left(1 + \frac{C_{a,1}(h_2)\, C_{q,1}(h_2, s_2)}{C_{a,2}(h_2)\, C_{q,2}(h_2, s_2)}\right). \tag{A12}$$

From the expressions above,

$$\begin{aligned} \frac{C_{a,1}(h_2)\, C_{q,1}(h_2, s_2)}{C_{a,2}(h_2)\, C_{q,2}(h_2, s_2)} = \exp\Big(&-0.5\tau_{\epsilon\eta}\big(s(v_1, \hat{\mu}_F^1, 1)^2 - s(v_1, \hat{\mu}_M, 1)^2\big) \\ &+ 0.5\big(\tau_a + \tau_{\epsilon\eta} - \tau_q(2)\big)\big(\hat{\mu}_1(h_2)^2 - \hat{\mu}_2(h_2)^2\big) \\ &+ 0.5\big(\tau_q(2) + \tau_\eta\big(\hat{\mu}_{q,1}(h_2, s_2)^2 - \hat{\mu}_{q,2}(h_2, s_2)^2\big)\big)\Big), \end{aligned}$$

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, fixing s_2, there exists a $\bar{v}$ such that for $v_1 > \bar{v}$, (A12) is satisfied. Similarly, fixing h_2, there exists an $\underline{s}$ such that for $s_2 < \underline{s}$, (A12) is satisfied. Therefore, for any prior beliefs about ability for each type, it is possible for discrimination to reverse in the second period. ∎

REFERENCES

Aigner, Dennis J., and Glen G. Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review* 30 (2): 175–87.

Altonji, Joseph G., and Charles R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *Quarterly Journal of Economics* 116 (1): 313–50.

Anwar, Shamena, and Hanming Fang. 2006. "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence." *American Economic Review* 96 (1): 127–51.

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics* 133 (4): 1885–1932.

Arrow, Kenneth. 1973. "The Theory of Discrimination." In *Discrimination in Labor Markets,* edited by Orley Ashenfelter and Albert Rees, 3–33. Princeton, NJ: Princeton University Press.

Ayalew, Shibiru, Shanthi Manian, and Ketki Sheth. 2018. "Discrimination from Below: Experimental Evidence on Female Leadership in Ethiopia." Unpublished.

Bartoš, Vojtech, Michal Bauer, Julie Chytilová, and Filip Matejka. 2016. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition." *American Economic Review* 106 (6): 1437–75.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *Quarterly Journal of Economics* 124 (4): 1497–1540.

Becker, Gary S. 1957. *The Economics of Discrimination.* Chicago: University of Chicago Press.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit Discrimination." *American Economic Review* 95 (2): 94–98.

Bertrand, Marianne, and Esther Duflo. 2017. "Field Experiments on Discrimination." In *Handbook of Economic Field Experiments,* Vol. 1, edited by Abhijit Vinayak Banerjee and Esther Duflo, 309–93. Amsterdam: North-Holland.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.

Biernat, Monica, Melvin Manis, and Thomas E. Nelson. 1991. "Stereotypes and Standards of Judgment." *Journal of Personality and Social Psychology* 60 (4): 485–99.

Biernat, Monica, Theresa K. Vescio, and Melvin Manis. 1998. "Judging and Behaving toward Members of Stereotyped Groups: A Shifting Standards Perspective." In *Intergroup Cognition and Intergroup Behavior*, edited by Constantine Sedikides, John Schopler, and Chester A. Insko, 151–76. Mahwah, NJ: Lawrence Erlbaum Associates.

Bocart, Fabian Y. R. P., Marina Gertsberg, and Rachel A. J. Pownall. 2018. "Glass Ceilings in the Art Market." Unpublished.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2019. "Inaccurate Statistical Discrimination." NBER Working Paper 25935.

Bohren, J. Aislinn, and Daniel N. Hauser. 2018. "Social Learning with Model Misspecification: A Framework and a Robustness Result." Unpublished.

Bohren, Aislinn, Alex Imas, and Michael Rosenberg. 2018. "The Language of Discrimination: Using Experimental versus Observational Data." *AEA Papers and Proceedings* 108: 169-74.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence: Dataset." *American Economic Review.* https://doi.org/10.1257/aer.20171829.

Booth, Alison L., Marco Francesconi, and Jeff Frank. 1999. "Glass Ceilings or Sticky Floors?" Unpublished.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics* 131 (4): 1753–94.

Coate, Stephen, and Glenn C. Loury. 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83 (5): 1220–40.

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *Quarterly Journal of Economics* 129 (4): 1625–60.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33 (1): 67–80.

Danilov, Anastasia, and Silvia Saccardo. 2017. "Discrimination in Disguise." Unpublished.

Ewens, Michael, Bryan Tomlin, and Liang Choon Wang. 2014. "Statistical Discrimination or Prejudice? A Large Sample Field Experiment." *Review of Economics and Statistics* 96 (1): 119–34.

Fang, Hanming, and Andrea Moro. 2011. "Theories of Statistical Discrimination and Affirmative Action: A Survey." In *Handbook of Social Economics,* Vol. 1, edited by Jess Benhabib, Matthew O. Jackson, and Alberto Bisin, 133–200. Amsterdam: North-Holland.

Fershtman, Chaim, and Uri Gneezy. 2001. "Discrimination in a Segmented Society: An Experimental Approach." *Quarterly Journal of Economics* 116 (1): 351–77.

Fiske, Susan T. 1998. "Stereotyping, Prejudice, and Discrimination." In *The Handbook of Social Psychology,* Vol. 1, edited by Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, 357–411. Boston: McGraw-Hill.

Fiske, Susan T., Donald N. Bersoff, Eugene Borgida, Kay Deaux, and Madeline E. Heilman. 1991. "Social Science Research on Trial: Use of Sex Stereotyping Research in Price Waterhouse v. Hopkins." *American Psychologist* 46 (10): 1049–60.

Fiske, Susan T., and Shelley E. Taylor. 1991. *Social Cognition.* New York: McGraw-Hill Education.

Fryer, Roland G. 2007. "Belief Flipping in a Dynamic Model of Statistical Discrimination." *Journal of Public Economics* 91 (5–6): 1151–66.

Fryer, Roland, and Matthew O. Jackson. 2008. "A Categorical Model of Cognition and Biased Decision-Making." *B.E. Journal of Theoretical Economics* 8 (1): Article 1935–1704.

Gneezy, Uri, John List, and Michael K. Price. 2012. "Toward an Understanding of Why People Discriminate: Evidence from a Series of Natural Field Experiments." NBER Working Paper 17855.

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90 (4): 715–41.

Gornall, Will, and Ilya A. Strebulaev. 2018. "Gender, Race, and Entrepreneurship: A Randomized Field Experiment on Venture Capitalists and Angels." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3301982 (accessed August 16, 2019).

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74 (6): 1464–80.

Groot, Wim, and Henriëtte Maassen van den Brink. 1996. "Glass Ceilings or Dead Ends: Job Promotion of Men and Women Compared." *Economics Letters* 53 (2): 221–26.

Kelley, Harold H. 1973. "The Process of Causal Attribution." *American Psychologist* 28 (2): 107–28.

Knowles, John, Nicola Persico, and Petra Todd. 2001. "Racial Bias in Motor Vehicle Searches: Theory and Evidence." *Journal of Political Economy* 109 (1): 203–29.

Kravitz, David A., and Judith Platania. 1993. "Attitudes and Beliefs about Affirmative Action: Effects of Target and of Respondent Sex and Ethnicity." *Journal of Applied Psychology* 78 (6): 928–38.

Leslie, Lisa M., Colleen Flaherty Manchester, and Patricia C. Dahm. 2017. "Why and When Does the Gender Gap Reverse? Diversity Goals and the Pay Premium for High Potential Women." *Academy of Management Journal* 60 (2): 402–32.

Lewis, Gregory B. 1986. "Gender and Promotions: Promotion Chances of White Men and Women in Federal White-Collar Employment." *Journal of Human Resources* 21 (3): 406–19.

List, John A. 2004. "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field." *Quarterly Journal of Economics* 119 (1): 49–89.

Lundberg, Shelly J., and Richard Startz. 1983. "Private Discrimination and Social Intervention in Competitive Labor Markets." *American Economic Review* 73 (3): 340–47.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. 2019. "Gender Bias in Teaching Evaluations." *Journal of the European Economic Association* 17 (2): 535–66.

Milgrom, Paul. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell Journal of Economics* 12 (2): 380–91.

Milkman, Katherine L., Modupe Akinola, and Dolly Chugh. 2012. "Temporal Distance and Discrimination: An Audit Study in Academia." *Psychological Science* 23 (7): 710–17.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. "Science Faculty's Subtle Gender Biases Favor Male Students." *Proceedings of the National Academy of Sciences* 109 (41): 16474–79.

Olson, James M., Robert J. Ellis, and Mark P. Zanna. 1983. "Validating Objective versus Subjective Judgment: Interest in Social Comparisons and Consistency Information." *Personality and Social Psychology Bulletin* 9 (3): 427–36.

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review* 101 (4): 1410–35.

Petersen, Trond, and Ishak Saporta. 2004. "The Opportunity Structure for Discrimination." *American Journal of Sociology* 109 (4): 852–901.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659–61.

Price, Joseph, and Justin Wolfers. 2010. "Racial Discrimination among NBA Referees." *Quarterly Journal of Economics* 125 (4): 1859–87.

Pronin, Emily, Daniel Y. Lin, and Lee Ross. 2002. "The Bias Blind Spot: Perceptions of Bias in Self versus Others." *Personality and Social Psychology Bulletin* 28 (3): 369–81.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How Stereotypes Impair Women's Careers in Science." *Proceedings of the National Academy of Sciences* 111 (12): 4403–08.

Riach, Peter A., and Judith Rich. 2006. "An Experimental Investigation of Sexual Discrimination in Hiring in the English Labor Market." *B.E. Journal of Economic Analysis and Policy: Advances in Economic Analysis & Policy* 6 (2): 1–20.

Rosette, Ashleigh Shelby, and Leigh Plunkett Tost. 2010. "Agentic Women and Communal Leadership: How Role Prescriptions Confer Advantage to Top Women Leaders." *Journal of Applied Psychology* 95 (2): 221–35.

Ross, Lee, David Greene, and Pamela House. 1977. "The 'False Consensus Effect': An Egocentric Bias in Social Perception and Attribution Processes." *Journal of Experimental Social Psychology* 13 (3): 279–301.

Sarsons, Heather. 2017. "Interpreting Signals in the Labor Market: Evidence from Medical Referrals." Unpublished.

Schwartzstein, Joshua. 2014. "Selective Attention and Learning." *Journal of the European Economic Association* 12 (6): 1423–52.

Snyder, Melvin L., Robert E. Kleck, Angelo Strenta, and Steven J. Mentzer. 1979. "Avoidance of the Handicapped: An Attributional Ambiguity Analysis." *Journal of Personality and Social Psychology* 37 (12): 2297–2306.

Vasilescu, Bogdan, Andrea Capiluppi, and Alexander Serebrenik. 2014. "Gender, Representation and Online Participation: A Quantitative Study." *Interacting with Computers* 26 (5): 488–511.

Williams, Wendy M., and Stephen 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 (17): 5360–65.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.