What Do Names Reveal?
Impacts of Blind Evaluations on Composition and Quality

Haruka Uchida*

March 2024

Abstract

Concealing candidate identities during evaluations, or "blinding", is often proposed as a tool for combatting discrimination. I study how blinding impacts candidate selection and quality, and the forms of discrimination driving these effects. I conduct a natural field experiment at an academic conference, running each submitted paper through both blind and non-blind review. Four years after the experiment, I collect proxy measures of paper quality—citations and publication statuses—for each paper and link it to the experimental data. Blinding significantly reduces scores for traditionally high-scoring groups, and consequently alters the composition of applicants who are accepted to the conference. Despite these compositional changes, blinding does not worsen the conference's ability to select high-quality papers. I develop a model of evaluator discrimination that allows me to rationalize these effects and decompose non-blind disparities into two distinct forms of discrimination: accurate statistical discrimination and bias.

*University of Chicago. uchida@uchiago.edu. Previous versions of this paper circulated under the title, "Reducing and Widening Disparities with Blind Evaluations: Evidence from a Field Experiment." I am grateful to Tomas Hromadka, Srdjan Ostojic, Anne-Marie Oswald, and the committee members of the Cosyne conference. These parties are in no way responsible for the analyses and interpretations presented in this paper. I thank Jesse Backstrom, Alec Brandon, Christina Brown, Leonardo Bursztyn, Katie Coffman, Michael Guy Cuna, Michael Dinerstein, Michael Greenstone, Justin Holz, Kilian Huber, Alex Imas, Juanna Joensen, Xianglong Kong, Clara Kyung, John List, Ariel Listo, Claire Mackevicius, Magne Mogstad, Jack Mountjoy, Matthew Notowidigdo, David Novgorodsky, Devin Pope, Evan Rose, Sally Sadoff, Heather Sarsons, Andrew Simon, Dan Svirsky, Min Sok Lee, Germán Villegas Bauer, Rebecca Wu, Hellary Zhang, and participants at the University of Chicago Experimental Workshop, University of Chicago Labor Seminar, University of Chicago Microeconomics Third Year Seminar, Economic Science Association Meeting, Southern Economic Association Annual onference, and Advances with Field Experiments at the University of Chicago for helpful comments and discussion. Izzy Allum, Dushan Arsov, Zack Fattore, Ashley Folts, Sonoe Fitzsimonds, Owen Humphries, John Kim, Yujin Lee, Zack Olson, Riley Osborn, Victor Sengpiel, Steven Shi, and Kaixin Wang provided outstanding research assistance. This research was conducted with approval from the University of Chicago Institutional Review Board, IRB19-1675. This RCT was registered as AEARCTR-0005139.

1 Introduction

Women and racial minorities are persistently underrepresented at top ranks of corporate, academic, and political career ladders (Bertrand and Duflo, 2017). Disparities in academia have been documented across fields and at every stage from undergraduate studies to professorship (Buckles, 2019). Given that one possible driver is discrimination, a common policy proposed to address such disparities is to conceal candidate identities during evaluations, "blinding" (e.g. Goldin and Rouse, 2000). If gaps arise because evaluators discriminate against particular demographics, then blinding should improve outcomes for traditionally disadvantaged groups.

However, it remains unclear whether blinding successfully combats evaluator discrimination, and which forms of discrimination it alleviates. Answering these questions is essential for understanding not just why blinding has effects, but also the incentives to adopt it. Canonical models of animus (Becker, 1957) predict that if evaluators use candidate identities to cater to biased preferences, blinding can cause evaluations to select better-qualified candidates. On the other hand, if evaluators rationally use candidate identities as informative signals of underlying candidate quality (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977), then blinding can interfere with this prediction process and generate the reverse effect. In this sense, the nature and extent of discrimination can jointly determine the direction and magnitude of how blinding affects the demographic composition and quality of selected candidates. These questions are empirically challenging to test because they require blind status to be distributed as-good-as-randomly across both candidates and evaluators (e.g. Blank, 1991; Behaghel et al., 2015), and data on underlying ex-post candidate quality is typically unobserved.

This paper studies the impacts of blinding on candidate selection and quality, and the forms of discrimination that explain these effects. I tackle the challenges mentioned above by conducting a natural field experiment at Computational and Systems Neuroscience ("Cosyne"), an academic conference that is one of the landmark conferences of its field. In past years of the conference, reviewers were randomly assigned (conditional on subfield) to score papers using noisy signals of underlying paper quality: submission content (a two-page summary of the paper) and author names.[1] This setting offers me two key advantages. First, I randomly assign reviewers into either blind or non-blind review, and run every submitted paper through both blind and non-blind review. This cleanly iso-

[1] Reviewers were explicitly told to score papers based on submission content. This is in contrast to other review systems that ask reviewers to incorporate aspects of author identity into evaluation scores; for instance, those that place importance on admitting young researchers.

lates the impact of blinding on evaluator decisions. Second, I collect data on proxies of paper quality for each submitted paper four years after the experiment.[2] This novel data linkage is necessary for understanding not just whether, but *how*, reviewers use author identities. Although citations are not the only dimension of paper quality, this provides a useful benchmark for accurate statistical discrimination. I additionally present robustness to other functional forms and measures of paper quality, and measures collected at other points in time. In this way, combining these three pieces of information—non-blind scores, blind scores, and ex-post quality—for each submitted paper allows me to answer several new questions on how blinding affects evaluator discrimination, and the forms of discrimination it alleviates.

First, I test how blinding changes the allocation of scores across demographic groups. Blinding substantially reduces scores for traditionally high-scoring groups: applicants who are more senior (non-students), from better ranked institutions, and male.[3] Most stark is the reduction in the score gap between students and non-students. While non-blind reviewers score student papers nearly 0.25 standard deviations lower than non-student papers, this gap shrinks by 75 percent among blind reviewers. This implies that the non-blind score disparity between students and non-students was not fully driven by differences in submission content.

My study design ensures that these patterns are not driven by applicant selection into blind evaluations, or behavioral changes by applicants. In contrast to settings that allow for differential selection into blinding,[4] both blind and non-blind reviewers received exactly the same set of papers. Because submissions were online documents, all reviewers for a given submitted paper received the same exact information, aside from author names. This ensures that any effects I observe are driven by reviewer behaviors. To address concerns that blind review is not truly blind because reviewers could search for papers online (Charness et al., 2022), I repeat the main analysis using the 83% of my sample of submissions that could not be found through an online search at the time of review (Section 3.2).

Next, I examine whether and how impacts on score disparities translate to changes in who is accepted to the conference. The relationship is not mechanical, since changes in scores may have been driven by infra-marginal candidates who were not on the margin of acceptance. The

[2] Previous versions of this paper showed results for citations two years after the experiment, which produced results that were generally qualitatively similar to the ones presented in the main text, which use citations four years after.

[3] These traits were pre-registered.

[4] For example, studies that compare disparities across journals that choose to be blind with those in journals that choose to be non-blind (e.g. Crane, 1967; Ferber and Teiman, 1980; Budden et al., 2008).

conference, both in the past and in the experimental year, determined acceptances accordingly: choose an overall acceptance rate given venue capacity constraints, and select the papers with the highest scores.[5] In the experimental year, the conference accepted roughly 60% of papers, and used all reviewer scores—both blind and non-blind—to determine final acceptances (this also ensured that every reviewer was incentivized to score truthfully). To test how blinding affected acceptances, I compare the papers that scored in the top 60% of blind scores, with those that scored in the top 60% of non-blind scores. I find that blinding meaningfully altered the composition of accepted papers. Using only blind review would have eliminated the student acceptance gap. Simulating acceptance outcomes for other acceptance thresholds (beyond the 60% that was used in reality) shows that this effect persisted across overall acceptance rates, implying that blinding benefitted students at essentially every margin. Blinding also reduced the applicant institution rank and gender acceptance gaps, though these changes were not statistically significant.

Third, I test whether blinding changes the relationship between reviewer scores, conference acceptances, and underlying paper quality. To do so, I collect each paper's number of citations and publication status, including the journal of publication, four years after the experiment. Reviewer scores are generally informative of paper quality: both blind and non-blind score rankings positively correlate with paper quality rankings, though the correlations are far from one. Blind scores are not worse predictors of paper quality than non-blind scores. Consequently, I do not find any significant quality differences between papers that would be admitted to the conference under blind review, and those that would be admitted under non-blind review. This suggests that even entities that prioritize selecting the most qualified candidates over equalizing representation may consider adopting blinding. These results are likely not driven by the causal effects of the conference on my measures of paper quality: the patterns persist even when subsetting to papers whose conference acceptance statuses would have remained the same regardless of blind regime.[6]

I follow past work in using a paper's citations (four years later) as a proxy for its underlying quality (e.g. Smart and Waldfogel, 1996; Card et al., 2020). It is still possible that other dimensions of paper quality—such as popularity with conference attendees—are not fully captured by my quality measures. It is not obvious whether omitting these dimensions causes me to over- or underestimate the impacts of blinding on quality. Nonetheless, I proceed using citations, given that they provide a useful benchmark for statistical discrimination, and additionally show results using other observable

[5] This decision rule indeed accurately predicts 95% of the conference acceptances from the prior two years.

[6] I discuss this further in Section 4.

quality proxies instead. I discuss these implications further in Section 5.5.

Insignificant impacts of blinding on quality, despite significant changes in composition, suggest that accurate statistical discrimination alone cannot explain non-blind disparities. I formalize this using a simple model of reviewer decision-making, and use the additional model structure to decompose disparities under non-blind review into distinct forms of reviewer discrimination. Consider a reviewer who assigns scores based on predictions of underlying paper quality, given the information at the time of review: author identities, and submission content (e.g. the research idea explained in the submission). Non-blind reviewers—unlike blind ones—may use author identities to (1) accurately statistically discriminate based off of underlying subgroup differences in paper quality (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977) and to (2) engage in bias, defined as any deviation from accurate statistical discrimination, and includes animus (Becker, 1957) and belief-based discrimination (Bordalo et al., 2019; Bohren et al., 2019). In contrast, blind reviewers cannot use author identities, so blind scores depend only on submission content. Blind scores therefore proxy reviewers' perceptions of submission content aside from author identities.

I take this model to the data, which is possible because I observe blind scores, non-blind scores, and proxies of paper quality for each submitted paper. My approach allows for my observed measures of quality to be measured with error. Model estimates imply that disparities in paper quality between students and non-students, conditional on submission content (blind scores), cannot fully explain their differences in non-blind scores. I reject that accurate statistical discrimination can explain more than 25% of the non-blind score gap between students and non-students, although this magnitude depends on how reviewers weight paper quality. If reviewers place greater importance on identifying the highest-performing papers, a greater portion of the gap is attributable to accurate statistical discrimination. This is because non-students are slightly more likely than students to reach the highest levels of paper quality. In contrast, if reviewers place greater importance on rejecting the lowest-performing papers, than a smaller portion of the gap is attributable to accurate statistical discrimination. Point estimates suggest that over 44% of the non-blind institution rank score gap can be explained by accurate statistical discrimination, though the confidence interval for this estimate is quite large. This suggests that the mechanism driving blinding effects for student status differs from the one driving effects for institution rank. These results help rationalize why blinding differentially affects demographic groups without altering admit quality: removing both reviewers' abilities to accurately statistically discriminate and engage in bias can generate competing forces on conference quality. In my context, they offset.

Most similar to this paper is Pleskac et al. (2024), who also investigate the effects of blind review at an academic conference using a within-paper design, and link reviewer scores with papers' publication outcomes two years later. This current paper differs by formulating and estimating a model of reviewer decision-making to understand mechanisms. This allows for decomposing baseline (non-blind) disparities into two distinct mechanisms, statistical discrimination and bias, which hold differing policy implications.

In sum, the contributions of this paper are threefold. First, my within-paper design cleanly identifies the impacts of blinding on reviewer behavior, while holding all else, including submission content, fixed. I show, in a natural field setting with real stakes, that blinding affects reviewer discrimination during evaluations. I illustrate how the group that benefits the most from blinding (students) does so across the full distribution of potential acceptance thresholds. This builds on the blinding literature which has largely focused on across-candidate comparisons (e.g. Blank, 1991; Krause et al., 2012) by collecting within-candidate data, and is similar to Tomkins et al. (2017) and Huber et al. (2022) who also collect blind and non-blind evaluation outcomes for a given paper, though the latter collects much of its data from non-incentivized surveys, and neither examine distributional impacts. More generally, these results speak to the large body of work testing how changes in available candidate information impacts evaluation outcomes (e.g. Bertrand and Mullainathan, 2004; Sarsons, 2017; Agan and Starr, 2018). While my experiment takes place in a specific context, my approach provides a general framework to answer similar questions in other contexts.

Second, I collect proxy data on each paper's ex-post quality and combine this with data on blind and non-blind evaluation outcomes. This novel data linkage allows me to answer several new questions in the literature. It first enables me to directly test how blinding affects the ability to select qualified candidates, which speaks to debates on whether blinding policies equalize representation at the expense of quality. This builds on past work that relies on naturally occurring variation in blind statuses, for instance Laband and Piette (1994) who compare the quality (citations) of papers accepted by journals that use blind review, with the quality of papers accepted by journals that use non-blind review. These methods are vulnerable to confounds, given that blinding is likely not distributed as-good-as-randomly in the real world, both across applicants and across evaluators. My approach in contrast identifies the effects of blinding through two stages of random assignment. This analysis on quality also contributes to work that examines the consequences of policies and movements that change the permissible extent of bias, such as Huber et al. (2021) who find that biased expulsions of Jewish managers during the rise of Nazi Germany worsened firm outcomes.

Finally, I connect my results to a model of reviewer decision-making to understand the mechanisms driving blinding effects. This approach contributes to work beyond the literature of blinding. Although understanding the impacts of changes on disparities alone are important, uncovering why these effects arise is crucial for understanding optimal policies. My unique data linkage makes it possible to estimate the model and decompose disparities under non-blind review into two distinct forms of discrimination. My model borrows insights from the large literature on identifying and disentangling discrimination (List, 2004), particularly the "outcomes test" literature (Becker, 1957) that tests for biased decision-making by comparing subgroup differences in post-evaluation outcomes. I contribute by showing how incorporating the notion of blind review can help sidestep typical assumptions on the (lack of) subgroup differences in the distributions of submission content. Conditioning directly on blind scores ensures that I test for differential outcomes across comparable submissions: otherwise, disparities in non-blind review may arise solely due to subgroup differences in submission content. My approach uses revealed preference—consequential decisions by reviewers and subsequent quality proxy measures—to disentangle mechanisms, in contrast to past work that relies on stated preferences (e.g. Behaghel et al., 2015).

2 Experimental Design

The experiment was conducted in the 2020 Computational and Systems Neuroscience ("Cosyne"), an annual international academic computational neuroscience conference.[7] It is one of the landmark conferences of its field, and draws around 1,000 attendees each year. Conference submissions are generally early works. Before the submissions were due in the experimental year, the conference committee (truthfully) told the public that the conference would try out blind review but that some reviewers would not be blinded, in order to ease the logistical transition to blind review. Neither applicants nor reviewers were told about the existence of an experiment (but knew about the existence of both review processes), which minimizes behavioral biases that can arise when participants are aware of being in an experiment (Levitt and List, 2007).

I do not expect the Covid pandemic to affect reviewer behaviors during the experiment, since the review process occurred during October through early December of 2019.

[7] The conference itself took place in the end of February 2020.

As in prior years, applicants submitted information on the paper that they would present if accepted. This comprised of the title and abstract, a 300-word description, two pages of more detailed explanation with potentially figures and tables, a list of the relevant subfields, and a list of all authors. The change in the experimental year was that applicants were instructed to anonymize their submission documents such that aside from the author list, their documents did not include any identifying information, including author names, author affiliations, or acknowledgements. Identifying information was entered separately in the application form, in the same format as previous years. Applicants were never informed of the identities of their assigned reviewers.

Essentially every part of the review process during the experiment was unchanged from previous years, except for two new levels of randomization. First, all reviewers were randomly assigned treatment status by a coin flip. A reviewer was either "Non-Blind" and received the author lists associated with assigned papers, or was "Blind" and did not receive the author lists. A reviewer was either always Blind or always Non-Blind to minimize salience of treatment status and preserve naturalness of the task at hand. To incentivize all reviewers to report truthfully, all reviewer scores—both Blind and Non-Blind—were used to determine acceptance outcomes, and reviewers were told this upfront. Second, every application was randomly assigned to four reviewers from its relevant subfield: two "Non-Blind" reviewers and two "Blind" ones.[8] All reviewers received the submission documents (the title, 300-word description, and 2 page explanation), in the same format as previous years. In both the experimental year and previous, reviewers did not receive any additional information about authors besides names—other traits such as affiliated institution were left to be inferred from names.

As in previous years, reviewers were assigned on average 10 papers, and scored their assigned papers on an integer scale of 1 through 10, inclusive. Importantly, reviewers were asked to grade based on paper quality. Specifically, reviewers were instructed: "*Please read each 2-page pdf in your list of assigned abstracts, and evaluate it with respect to the criteria of: (1) Significance: how much does the study advance the state of the field? (2) Originality: how novel are the concepts, approach and/or techniques? (3) Clarity: are the addressed questions and the obtained results clearly presented? (4) Relevance to the Cosyne audience.*" This accentuates the subjective nature of the review process, so that reviewers may plausibly use author identities to inform their own predictions about underlying paper quality, but also use the information to show impartiality towards certain

[8] In prior years, each paper was assigned to three Non-Blind reviewers of its relevant subfield.

types of authors.[9] This makes the setting an ideal one to study how changes in identity information shape reviewer behaviors.

Reviewers were given all of their assigned papers at once, and could review papers at any order. Reviewers were recruited in the same way as prior years, which was a mix between committee recruitment and volunteers. Reviewers are typically academics from the field. Reviewer identities were never revealed to authors.

As stated in my pre-analysis plan (AEARCTR-0005139), I collect information on each paper's applicant (the individual who submits the application and will present at the conference if accepted) and the principal investigator (referred to as the PI here forward).[10] This information was collected through self-reported forms and later verified.[11] Specifically, I collect each applicant and PI's gender, rank of affiliated institution, and the applicant's student status (non-students are of higher status, such as post-docs or assistant professors). Institution rankings were taken from the the 2020 US News Best Global University Rankings. Because majority of applicants and PIs were from the same institution, I focus only on the applicant's institution rank going forward. As I also state in my pre-analysis plan, during the time of review, I collect information on whether or not each paper can be found through an online search. This data is used in a robustness check to address concerns that blinding is not truly blind.

While not mentioned in my pre-analysis plan, I also collect the number of historical cumulative citations that each applicant and PI is associated with at the time of submission, as exploratory work. I generally find that the main results presented do not change whether I control for these citation counts or not (Appendix B.5). Two and four years after the experiment, I collect, for each paper, information regarding whether it is available online, is published, journal of publication, and the number of citations it is associated with (Section 4).[12] Appendix A elaborates on the entire data collection process.

[9] Past work often finds that agents behave more biased in situations with greater ambiguity (e.g. Bowles et al., 2005).

[10] Non-Blind reviewers were informed of the applicant's identity, so that reviewers knew which author would present at the conference if accepted.

[11] Reviewers were never explicitly told about traits corresponding to authors. I inspect disparities along these dimensions, however, since Non-Blind reviewers can perceive them upon receiving author names.

[12] I show results using measures from four years after the experiment. Results are qualitatively similar when using measures from two years after.

3 Effects of Blinding on Scores and Acceptances

The experimental sample consisted of 657 paper submissions and 245 reviewers, translating to 2591 unique paper-reviewer observations.[13] Table 1 summarizes author traits (at the paper level), which by construction represents the traits in both treatment groups, since each paper was scored both Blind and Non-Blind. Around half of applicants were students. Among papers submitted by applicants who were affiliated with an institution rank (553 out of 657 papers), the median rank was 21 (see Appendix B.1 for more). Going forward, I bin institution rank by whether it is top 20 or not ("lower rank").

Applicants were majority male (76%) and PIs even more (82%). Applicant gender was not significantly correlated with student status nor institution rank (Table A1). Applicant and PI genders were not significantly correlated. Papers on average had four coauthors, and over 97% of papers had more than one author. Student applicants were significantly less likely to submit a solo-authored paper than non-students. There is no significant gender difference—for both applicants and PIs—in coauthorship behavior.

As expected given random assignment, Blind reviewers did not systematically differ in observed traits from Non-Blind reviewers (Table 2).

3.1 On Scores

What disparities would arise if the conference was run "business as usual", without blinding? Figure 1 reports scores by each trait. On average, when reviewers received author identities, student applicants scored 0.49 points (≈ 0.25 SD) worse than their senior counterparts. Applicants from top 20 ranked institutions (better than the median ranking) scored 0.77 points (≈ 0.40 SD) higher than those from lower ranked institutions. Female applicants and PIs received lower scores than their male peers, though the difference for applicants is not statistically significant. The conditional score gaps—when I consider each trait while controlling for the rest—exhibit similar patterns (Table A3).

The notion that some subgroups score worse than others under Non-Blind review alone does not reveal differential treatment, since it could be driven by differences in submission content. For this, I compare disparities between Blind and Non-Blind scores for the same set of papers. Recall that the conference explicitly instructed reviewers to evaluate submissions based on perceived paper

[13] Not all reviewers submitted scores for all of their assigned papers, so that 6% of papers ended up with three scores instead of four. I address potential reviewer endogeneity in Section 3.2.

quality. Since both Blind and Non-Blind reviewers receive the same exact submission aside from author names, differences across Blind and Non-Blind scores reflect changes that reviewers make based on author identity.

Blinding directionally reduced the correlation between the scores given by a paper's two same-treatment reviewers (Non-Blind correlation of 0.26, Blind 0.23).[14] Blinding did not significantly change the variation in reviewer scores ($p = 0.652$ for a variance ratio test).

I find that blinding significantly reduced scores for traditionally better-scoring applicants: applicants who were senior (non-students), from top 20 institutions, and male (Figure 1).[15] The estimates for PI gender are noisy, and in terms of magnitude, female PIs scores decreased by a greater magnitude than male PIs.

To consider unconditional disparities, I estimate, for paper p assigned to reviewer r:

$$Y_{p,r} = \beta_0 + \beta X_p * \mathbb{1}\{Blind\}_r + \lambda_r + \gamma_p + \varepsilon_{p,r} \tag{1}$$

where $Y_{p,r}$ is the score that paper p received from reviewer r, X_p is a vector of paper-level applicant and PI traits, $\mathbb{1}\{Blind\}_r$ is an indicator for whether reviewer r is Blind, λ_r are reviewer fixed effects, γ_p are paper fixed effects, and $\varepsilon_{p,r}$ is an error term. β captures the average change in a score gap due to blinding. While the paper and reviewer fixed effects are not necessarily to identify the impacts of blinding due to my randomization, I include them since they can improve precision if reviewers or papers differ in average scores they give or receive, respectively.[16]

Table A4 summarizes, and the results are consistent with the unconditional means in Figure 1. Blinding significantly reduced the student score gap by 75% ($p = 0.05$ after multiple hypothesis testing adjustment). Under blind review, student applicants were statistically indistinguishable from non-student applicants. This is not driven by a small number of applicants: the distribution of treatment effects (difference between a paper's average Blind and Non-Blind score) for students stochastically dominated the distribution for non-students (Figure A2), revealing that students were generally greatly disadvantaged by reviewers knowing their identities. This could be driven

[14] Interestingly these within-paper score correlations are similar in magnitude to Blank (1991).

[15] Blind reviewers were on average significantly harsher than Non-Blind reviewers, which is consistent with previous work (e.g. Blank, 1991).

[16] The paper fixed effects are identified because each paper was scored by multiple reviewers, some Blind and some Non-Blind. The reviewer fixed effects are identified because a given reviewer scored multiple papers. I do not include X_p alone in the regression as it is collinear with the paper fixed effects. Similarly, $\mathbb{1}\{Blind\}_r$ alone is collinear with the reviewer fixed effects.

by various reasons. For instance, reviewers may use student status as information beyond the submission to update beliefs on underlying paper quality. Reviewers may also over-penalize students during this belief updating process, relative to the truth. Reviewers may also hold self-doubt towards criticizing a senior author (note that concerns of retaliation are minimized, since reviewer identities are never disclosed to authors). I explore these potential mechanisms in Section 5.

Blinding reduced the institution rank score gap, such that the difference between applicants affiliated with a top 20 ranked institution and those affiliated with a lower ranked institution decreased by around 25%. However, this effect is marginally significant and is not statistically significant after a multiple hypothesis testing adjustment ($p = 0.22$). Moreover, the institution rank score gap persisted under blind review, suggesting that at least some portion of the non-blind disparity can be explained by subgroup differences in submission content.

Interestingly, the notion that those from worse ranked institutions benefitted from blinding is largely driven by the fact that applicants from near-top (rank 6 through 20) institutions were worse off (Figure A3), though these differences across finer categories of institution rank are not statistically significant. This finding qualitatively matches those from Blank (1991), who also finds that blinding most affects (negatively) authors from near-top institutions, rather than those at the highest or lowest ranks.

The impacts on gender score gaps are imprecisely estimated. Point estimates imply that blinding reduced the applicant gender score gap by around 50%, while increasing the PI gender score gap, but neither are statistically significant.[17]

The results presented here are likely not driven by subgroup differences in coauthorships. When comparing the role of the traits of the rest of the coauthors (authors who are neither the applicant nor PI), I find strong evidence suggesting that score disparities are largely driven by applicant and PI traits. The effects reported above persist even after controlling for coauthor traits (Appendix B.4) or the number of historical citations associated with each applicant and PI (Appendix B.5).

Despite possible reviewer heterogeneity (Welch, 2014), I do not find significant evidence of

[17] Taking these point estimates literally suggests that female applicants are better off by blinding, but female PIs are worse off, though neither effects are statistically significant. There are many possible reasons why the effects of blinding on the gender gap may go in opposite directions across applicants and PIs (though I do not find this pattern with precision). For instance, applicants are generally the ones that would present the paper if it were chosen, which means they hold a very visible role. PIs on the other hand may represent the available resources or networks available to the paper's authors. Another distinction is that PIs are generally more established than applicants. Recent work suggests that discrimination may change with reputation and seniority (e.g. Bohren et al., 2019; Petersen and Saporta, 2004).

heterogeneity by reviewer gender and institution rank (Appendix B.6). Point estimates suggest that female reviewers are more favorable towards female applicants and PIs than male reviewers, and blinding reverses this gap, but these patterns are not statistically significant.

3.2 Robustness Checks

In this section, I run robustness checks to address three potential concerns with the main analysis: (1) clustering at the reviewer-level as opposed to the paper-level, (2) that Blind reviewers were not truly blind since they could have learned author identities elsewhere, (3) not every reviewer submitted their assigned reviews. Table 3 summarizes the results from each (see Table A11 for selection into these subsamples).

First, my main results cluster at the reviewer-level since treatment (blind status) was assigned at the reviewer-level. Whether I cluster at the reviewer-level or paper-level does not change my results.

Second, a common concern with blind evaluations is that "Blind" may not truly be blind. This may be heightened due to the rise in online working papers, such that reviewers may easily find the papers and thus author identities, even if the conference norm is to present unpublished papers (Goldberg, 2012; Charness et al., 2022). It is unclear in which direction this would bias estimates. If Blind reviewers knew author identities and acted as they would have if they had received the author list (i.e., been "Non-Blind"), then this would underestimate my estimates of blinding reviewers to author identity. On the other hand, if Blind reviewers were aware of author identities and behaved more favorably towards groups that they otherwise would not have, my estimates of blinding effects would be over-inflated. With this in mind, I repeat the main analyses with the subsample of papers that were not available online during the time of review. When the experiment was occurring, at least two research assistants searched for each paper online, using the title, authors, and abstract together, which included examining author webpages when available. 83% of papers (551 out of 657) in my sample did not appear in an online search using these steps. In general, I find that conclusions from this subsample analysis are consistent with the main results. While I cannot ensure that the reviewers of this subsample were blind to author identity, it provides suggestive evidence by removing the papers most likely to be identifiable by reviewers.

Third, I run two additional robustness checks to address the fact that some papers did not receive all 4 reviews. 37 of the 657 papers (6%) ended up with three reviewer scores instead of four.

No paper was missing more than one of its four reviews. To address potential endogeneity concerns of missing reviews, I re-do the main analysis in two different ways. First, I re-run the analyses using the subsample of papers that received scores from all of its assigned reviewers (Column 4 of Table 3). Second, I implement inverse probability weighting, using a paper's applicant and PI traits to predict the likelihood that it is missing reviews as in Table A11 (Column 5 of Table 3). Both produce results consistent to the main results, which suggests that the results are likely not driven by some reviewers' choices to only score particular papers.

3.3 On Acceptances

3.3.1 Under the Realized Overall Acceptance Rate

Average changes in score gaps will influence disparities in acceptances, presumably the outcome of true interest, only if it causes the advantaged groups to be crowded out. The overall effect depends on who is at the margin, the distribution of scores (Figure A1), and the effects of blinding along the score distribution (Figure 4).

The conference each year determines acceptances directly from reviewer scores: the conference chooses an overall acceptance rate given that year's venue capacity constraints, and then selects the papers with the highest reviewer scores. Figure A4 corroborates this relationship between reviewer scores and acceptances.[18] In the experimental year, the conference accepted around 60% of submissions (402 out of 657 papers), using both a paper's Non-Blind and Blind scores to determine acceptances.[19] I first test for disparities in realized acceptances by estimating, for paper p:

$$Y_p = \alpha_0 + \alpha X_p + \epsilon_p \tag{2}$$

where Y_p is an indicator for whether the paper was accepted, X_p is a vector of paper-level applicant and PI traits, and ϵ_p is an error term. Column 1 of Table 4 summarizes. Students were 7 percentage points less likely to be accepted than non-students, and applicants from lower ranked institutions were 19 percentage points less likely to be accepted than those from top 20 ranked institutions.

Did the use of blind review impact acceptance outcomes? To answer this question, I investigate which papers would have been accepted if the conference had used only Non-Blind scores, and

[18] This decision rule accurately predict 95% of the conference acceptances from the prior two years.

[19] Historically, the conference had an overall acceptance rate of around 35% (1 year prior, i.e. Spring of 2019) and 55% (2 years prior, i.e. Spring of 2018). The covid pandemic did not affect the conference's acceptance decisions in the year of the experiment, since the decisions were finalized in the winter of 2019.

which would have been accepted if it used only Blind scores. Because I observe both each paper's ranking in Blind scores and its ranking in Non-Blind scores, I simulate which papers would have been accepted under one state of the world (Blind) but not the other (Non-Blind), and vice versa. I then estimate Equation 2, now using these simulated acceptance statuses as the outcome variable.

Table 4 summarizes. If only Non-Blind scores were used to determine acceptances, then under the realized acceptance rate, applicants who were students, from lower ranked institutions, and female, would have been 12, 18, and 9 percentage points, respectively, less likely to be admitted than their counterparts (column 2 of Table 4). Blinding meaningfully altered conference acceptances. If only Blind scores were used, 30% of papers that would have been accepted under Non-Blind review would have been crowded out (Table A10). These changes were not substituting across similar authors. In fact, using only Blind scores to determine acceptances would have eliminated the student acceptance rate gap (column 3 of Table 4). Blinding did not substantially change the applicant institution rank acceptance gaps; even if only blind scores were used to determine acceptances, applicants from lower ranked institutions would have been 16 percentage points less likely to be accepted to the conference than those from top 20 institutions. Blinding also reduced the gender acceptance rate gaps, and slightly worsened the PI gender acceptance rate gap, though these were not statistically significant. Ultimately, under the given overall acceptance rate, the incorporation of blind review in the conference selection process significantly crowded in students.

3.3.2 Under Simulated Overall Acceptance Rates

Next, I explore whether conclusions vary across acceptance margins, by simulating acceptance outcomes for various overall acceptance rates.[20] Figure 2 illustrates. Students uniformly benefit from blinding: regardless of conference selectivity, students are generally around 4 percentage points more likely to be accepted into the conference with blinding. Disparities by institution rank exhibit similar patterns. The impacts on gender gaps are more ambiguous. I find suggestive evidence of heterogeneity for the PI gender gap. Female PIs directionally benefit from blinding at less selective thresholds (low overall acceptance rates), but are crowded out by blinding at highly selective margins, although these are not statistically significant. This is likely because there are fewer women relative to men at the margin of high thresholds and more women relative to men at less selective thresholds (Figure A1h), rather than heterogeneous effects of blinding on scores along the submission

[20] This is similar to Kessler et al. (2019) who analyze distributional effects in callback rate gaps. They find that employers prefer applicants with prestigious internships at essentially every point on the callback threshold distribution.

content distribution (Figure 4).

Ultimately, these figures may not capture the full effects of varying the overall acceptance rate on disparities, as they hold the applicant pool fixed. It is possible that changing the overall acceptance rate affects acceptance rate gaps through many endogenous channels, such as through shifts in applicants' decisions to apply in the first place,[21] or reviewers' effort and attitudes towards papers they guess (accurately or inaccurately) to be infra-marginal. There is reason to believe that these channels, particularly the latter, would be relatively small in this context, though I cannot rule it out. Neither applicants nor reviewers in this experiment were aware of the overall acceptance rate (60%), particularly given that the conference annually changes location and therefore capacity constraints. Historically, the conference had an overall acceptance rate of around 35% (1 year prior) and 55% (2 years prior). Nonetheless, it is possible that changing the overall acceptance rate impacts disparities through alternative channels. These simulations are instead meant to illustrate that even when holding all else fixed, including the applicant pool, conclusions on the efficacy of blinding can still vary with the overall acceptance rate. This offers a potential reconciliation for why prior work on blinding has found mixed, or seemingly inconsistent, results.

4 Effects of Blinding on Quality

Given that blinding changes the allocation of scores to papers, are blind scores better or worse than non-blind scores in predicting underlying paper quality? One argument against blinding is that it can worsen evaluators' abilities to predict candidate quality. To assess this claim, I collect each paper's number of citations and publication status, including the journal of publication, four years after the experiment. I follow past work in interpreting these measures as proxies for underlying paper quality (e.g. Smart and Waldfogel, 1996; Card et al., 2020). While true paper quality is likely unobserved and multidimensional, these measures serve as proxies of quality that capture dimensions of a paper's influence and perceived quality, and I here forward refer to them as quality measures for expositional ease. I discuss further the implications of using these measures as paper quality in Section 5.5. Table 5 shows the raw means. Nearly 80% of the papers were available online. 68% of papers had at least one citation, 52% were published.

First, I benchmark the predictive power of reviewer scores against random paper acceptance.

[21] For instance, applicants' propensities to apply can be differentially impacted by systematic changes in diversity initiatives or statements (Niederle et al., 2013; Leibbrandt and List, 2018).

If papers were randomly selected, then accepting $X\%$ of papers corresponds to admitting around $X\%$ of the cumulative number of citations associated with all submitted papers. Figure 3 shows the fraction of citations that are associated with admitted papers, relative to the total number of citations associated with the full sample of papers, depending on whether admission is based off of Blind scores or Non-Blind scores. Both lines are more outwards relative to the random benchmark, which suggests that reviewer scores perform better than random acceptance at admitting highly-cited papers, and are informative of paper quality.

Regressing papers' quality percentile ranks on its reviewer score percentile ranks confirms that reviewer scores contain meaningful information on paper quality (Table A12). Predicting citations using author observables (applicant student status, institution rank, gender, PI gender), and comparing its predictive power with reviewer scores shows that reviewer scores provide significantly more information than these observable traits alone.

Moreover, I do not find any significant difference in the predictive power of Blind scores relative to Non-Blind scores. This is captured both by the overlapping lines in Figure 3 and the insignificant difference in Blind and Non-Blind score coefficients in Table A12. This suggests that Blind scores perform as well as Non-Blind scores in predicting paper quality. These relationships are not driven by differential selection into subfields, since they remain unchanged with the inclusion of subfield fixed effects. This pattern also persists when considering a paper's within-reviewer ranking instead of its within-sample one (Table A13).

Consequently, using only Blind review admits just as high-quality papers as using only Non-Blind review. Comparing the citations that would have been accepted under blind review and those that would have been under non-blind review, generally across overall acceptance rates, I reject that blinding caused the conference to select lower-quality candidates by more than 5 citations (≈ 0.12 standard deviations) (Figure A6). In other words, blind review does not pose a significant tradeoff between changing representation and reducing quality.

One possible concern is that the above patterns are driven by the fact that conference acceptance itself influenced citation and publication outcomes. This is likely not the case. To check this, I remove the set of papers that were marginal to the blinding policy, by subsetting to the papers that scored either both in the top 60% under Blind and Non-Blind review, or both in the bottom 40% under Blind and Non-Blind. I re-construct percentile rankings among these 417 papers, and the same patterns emerge (Table A14). Additionally, using the fact that the conference's acceptance rule creates a discontinuity in reviewer scores that determines acceptance, I conduct a regression discontinuity and find no evidence that acceptance affects citations 4 years later (Figure A5).

5 Disentangling Sources of Discrimination

My results above show that blinding changes the demographic composition of those accepted to the conference. This implies that reviewers use author names during evaluations when the information is given to them. Simultaneously, I find that blind scores are no worse at predicting paper quality than non-blind scores, which suggests that accurate statistical discrimination alone cannot explain the blinding effects. To conclude, I interpret these results using a simple model of reviewer decision-making. The additional structure of the model allows me to decompose baseline (Non-Blind) disparities into two mechanisms through which reviewers may use information on author identities: (1) accurate statistical discrimination (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977), wherein reviewers "efficiently" use author group memberships to update beliefs on latent paper quality, and (2) other determinants of disparities, including animus (Becker, 1957) and belief-based discrimination (Bordalo et al., 2019; Bohren et al., 2019).

I build on the framework of Canay et al. (2020), who present a cost-benefit based decision model to interpret the literature on "outcome tests", which compare subgroup differences in post-evaluation outcomes to identify evaluator bias (Becker, 1957). While these works focus on gathering insights from non-blind evaluations and subsequent outcomes, I add in the notion of blinding, which helps identify model components that were previously unobserved.

5.1 A Model of Reviewer Scores

Consider a setting where reviewers are asked to score papers based on rational predictions of underlying (unobserved) paper quality, Q_p. First, consider Non-Blind reviewers. These reviewers receive author lists and submission content, and can use this information to inform predictions of paper quality. Non-Blind reviewers may also use this information to engage in bias. Following previous work, define bias as deviations from accurate statistical discrimination, so that it includes both animus (Becker, 1957) and belief-based discrimination (Bordalo et al., 2019; Bohren et al., 2019).

Let x_p be a vector of author traits that are observable to both the reviewer and researcher through the author list (e.g. applicant's student status, gender, subfields). Let v_p capture a paper's submission content that is observed by the reviewer at the time of review (e.g. how compelling the idea described in the submission is), but not directly observed by the researcher. For ease of discourse, assume that higher values of v_p represent "better" submissions (e.g. more novel idea). Define $\mathcal{X}$ as the set of all possible x_p and $\mathcal{V}$ the set of all possible v_p. Following this notation, paper p receives, in expectation across reviewers, a Non-Blind score that is given by:

$$S_p^{NB} = \mathbb{E}[Q_p|x_p, v_p] - \tau(x_p, v_p) \tag{3}$$

where $\tau(x_p, v_p)$ captures variation in reviewer scores that cannot be explained by variation in expected paper quality. Note that the score equation captures behavior of the average reviewer. Appendix D.1 explains the link between the average reviewer and a single reviewer's decision problem.

In contrast to Non-Blind reviewers, Blind reviewers cannot observe author information, x_p. Blind reviewers must therefore predict underlying paper quality using only submission content, v_p. Assume that for a given paper, Blind reviewers observe the same submission content as Non-Blind reviewers.[22] As a result, paper p's Blind score is given by:

$$S_p^{B} = \mathbb{E}[Q_p|v_p] - \tau(v_p) \tag{4}$$

where again $\tau(v_p)$ captures the residual variation in Blind scores that is not attributed to expected paper quality. Note that Equation 4 can allow for the possibility that reviewers form predictions about author characteristics using submission content, and imbed these predictions into scores: the key difference from Non-Blind review is that all papers with the same submission content receive the same score, regardless of actual author traits x_p. This immediately highlights the potential value of data on Blind scores, as it can give insight on submission content, v_p, an attribute that is typically taken to be unobserved.

Define $S^{NB}(x, v)$ as the mapping from a paper's author traits and submission content to its expected Non-Blind score: $S^{NB}(x, v) = \mathbb{E}[Q|x, v] - \tau(x, v)$. Consider two distinct vectors of author traits, $x, x' \in \mathcal{X}$. The average difference in Non-Blind scores between papers by authors of trait x

[22] These assumptions are violated if, for instance, blinding not only removes author identities from reviewers' information sets, but also causes reviewers to disengage with the review process and pay less attention to submission content.

and papers by authors of trait x', after accounting for differences in submission content, is then:

$$\underbrace{\mathbb{E}_V\Big[S^{NB}(x,V) - S^{NB}(x',V)\Big]}_{\text{Average difference in Non-Blind score}}$$

$$= \underbrace{\mathbb{E}_V\Big[\mathbb{E}[Q|x,V] - \mathbb{E}[Q|x',V]\Big]}_{\text{Average difference in expected paper quality}} - \underbrace{\mathbb{E}_V\Big[\tau(x,V) - \tau(x',V)\Big]}_{\equiv\Delta_{x,x'}=\text{Average bias}} \qquad (5)$$

where the expectation is taken over the marginal distribution of submission content, V. In other words, the average difference in scores can be decomposed into two components: average subgroup differences in expected paper quality, and a residual gap. This latter component captures subgroup differences in the "penalty" to scores that is not explained by accurate statistical discrimination:

Definition 5.1. Subgroup $x \in \mathcal{X}$ faces *bias* relative to subgroup $x' \in \mathcal{X}$ at submission content $v \in \mathcal{V}$ if $\tau(x,v) > \tau(x',v)$.

When reviewers are unbiased, subgroup differences in Non-Blind scores correspond directly to subgroup differences in expected paper quality.

It is worth making a few notes regarding what "bias" here represents. First, it can capture forms of discrimination that are distinct from accurate statistical discrimination: this includes animus and belief-based discrimination. Potential mechanisms can also include attention discrimination (Bartoš et al., 2016), if reviewers pay less attention to papers by certain authors under non-blind review. Second, while the term "bias" may intuitively place blame on reviewer for intending to behave a particular way, I will not identify reviewer motives, which can differ from what reviewer behavior is consistent with. Third, while the model assumes that reviewers base scores off of paper quality, if reviewers' true objective functions differ, then subgroup differences in these "omitted payoffs" (Kleinberg et al., 2018) can affect the magnitude of estimated bias. In my context, although reviewers in the experiment were instructed to evaluate papers solely based on paper content, reviewers in reality may incorporate other aspects of author identities, such as how good of a speaker an applicant may be. These concerns persist when testing for discrimination in contexts beyond this paper and outside of academic review more generally. For example, an employer may base hiring decisions off of expected worker productivity, but may also consider whether the applicant would get along with co-workers, or provide good mentorship. Finally, specification errors and systematic measurement error that varies across the space of (x,v) can affect the estimates of bias (Canay et al., 2020). I discuss each of these notions further when interpreting my results in Section 5.4.

Define the latter component of Equation 5, the average bias that authors of trait x face relative to authors of x', as $\Delta_{x,x'}$. As Equation 5 highlights, comparing subgroup differences in Non-Blind scores alone does not identify $\Delta_{x,x'}$, since it can be driven by both subgroup differences in expected paper quality and bias. The goal going forward will be to disentangle the two. The usual problem in doing so is that even when quality measures are observed, submission content remains unobserved to the researcher.

Assume from here forward that a paper's expected Non-Blind and Blind scores are both strictly increasing in its ranking in submission content:

Assumption 1. *Define the functions $S^{NB}(x,v) \equiv E[Q|x,v] - \tau(x,v)$ and $S^{B}(v) \equiv E[Q|v] - \tau(v)$. Assume that the functions $S^{NB}(x,v)$ and $S^{B}(v)$ are both strictly increasing and continuous in $v \in \mathcal{V}$, for all $x \in \mathcal{X}$.*

An example of when this assumption is violated is: if reviewers' biases against students is increasing in submission content so much so that the score that students with the best submissions receive is lower than the score that students with worse submissions receive. Note that the above assumption does not necessarily assume away heterogeneity in bias against a subgroup, since it does not restrict $\tau(x,v) - \tau(x',v)$ to be monotone in v. Reviewers may, for instance, display greater bias against students at higher levels of submission content than those at lower levels, just not so much that students at these levels are scored worse than students at lower levels of submission content. Additionally, this assumption still allows for some forms of attention discrimination (Bartoš et al., 2016), wherein non-blind reviewers pay little attention to certain papers by a subgroup.[23]

5.2 Identifying Mechanisms

Data on Non-Blind scores, Blind scores, and quality, for each paper, identify the components of the model. My approach differs from past work in three main ways. First, I can identify marginal candidates directly from the data, because I observe reviewer scores for each paper and the conference's decision rule. This avoids "infra-marginality bias", which is that simply comparing average subgroup differences need not identify discrimination because average differences are generally uninformative

[23] For example, non-blind reviewers may, upon seeing author names, pay little attention to papers submitted by students and therefore fail to reward high-quality student papers with higher scores, but do give higher scores to high-quality papers submitted by non-students.

of marginal ones. Past work on outcomes tests generally either only observe a binary outcome[24] or does not know evaluation decision rule, so that the marginal candidate cannot be identified without imposing additional assumptions.[25]

Second, I use Blind scores to learn about submission content, v_p, a component that the outcomes test literature generally takes as unobserved to the researcher.[26] This ensures that I am testing for discrimination using otherwise comparable papers, without imposing additional assumptions on the distribution of observed author characteristics and unobserved submission content.

Third, I am able to observe subsequent proxies of paper quality for every submission in my sample. Unlike work on outcomes tests, the literature on blinding typically does not include data on quality. Even within past work on outcomes tests, contexts generally observe outcomes for an endogenously selected group—for instance, misconduct rates among defendants who judges decided to release—but this outcome is by construction unobserved for defendants judges chose not to release.[27] In my setting, directly observing paper outcomes–citations and publication status– years after the experiment helps recover true expectations of underlying paper quality, without relying on matching or extrapolation models. Combining the model and my experimental data thus highlights the value of my study design, as it overcomes various challenges that past work has faced in identifying the determinants of disparities.

Turning to identification: first, subgroup differences in expected quality (conditional on submission content) are identified from data on Blind scores and quality. By Assumption 1, a paper's Blind score is strictly increasing in its submission content. Taken together with Equation 4, a paper's ranking in Blind score is equal to its ranking in submission content: $R(S_p^B) = R(v_p)$, where R is the rank function. Consequently, $\mathbb{E}[Q_p|R(S_p^B)] = \mathbb{E}[Q_p|R(v_p)]$. Subgroup conditional expectations

[24] For instance, whether a bail judge releases a defendant prior to trial when testing for racial bias in judge decisions.

[25] Most relevant to this context is Smart and Waldfogel (1996), who test for bias by journal editors by comparing the number of citations associated with published papers, conditional on the order that the editor places the paper in the journal, which authors interpret as capturing the editor's quality assessment and therefore revealing marginal papers. Knowles et al. (2001) present an equilibrium model that implies that the average is informative of the marginal. Others have utilized exogenous assignment of evaluators as an instrumental variable for evaluation leniency to identify the marginal (Arnold et al., 2018; Dobbie et al., 2021). Arnold et al. (2022) have taken a similar approach with instrumental variables, but also distinguish between accurate statistical discrimination and bias by estimating misconduct risk (analogous to paper quality in the setting of the current paper) by extrapolating across released defendants of judges who vary in overall release rates. This relies on accuracy of the extrapolation model.

[26] In the bail judge example, v_p represents relevant non-race characteristics that a judge observes but not the researcher, which may be attributes such as past criminal history or family structure.

[27] Arnold et al. (2022) estimate misconduct risk (analogous to paper quality in the setting of the current paper) by extrapolating across released defendants of judges who vary in overall release rates. This approach relies on accuracy of the extrapolation model. Other work has generated models to predict application quality, for instance by matching applications using the text analysis to past applications and its subsequent outcomes (e.g. Li, 2017). This again hinges on the matching process to accurately pair applications, otherwise application quality is mis-measured.

in quality, $\mathbb{E}[Q_p|x_p, R(v_p)]$, are therefore identified since mean quality among papers of characteristics $(x_p, R(S_p^B))$ is directly observed. Then, subgroup differences in these conditional expectations identify the extent to which there is scope for accurate statistical discrimination.

Next, adding in data on Non-Blind scores identifies bias. Given that the rank transformation is a strictly monotonic one, let $t(x, R(v))$ be the mapping from $(x, R(v))$ to $\tau(x, v)$ for all $x \in \mathcal{X}$ and $v \in \mathcal{V}$. Since conditional expectations in quality can be identified from the data, $t(x_p, R(v_p))$ can be identified by taking the difference between paper's conditional expectation in paper quality and its Non-Blind score (eq. 3). Subgroup differences in $t(\cdot)$, conditional on Blind score, implies bias.

5.3 Estimation

Estimating the model does not require me to directly observe underlying paper quality, Q_p, but does require observing an unbiased (in a statistical sense) measure of it. Suppose the data contains a proxy measure of quality, $\tilde{Q}_p$, for each paper, where $\tilde{Q}_p = Q_p + u_p$ and u_p is an independently distributed paper-specific error term. If $\tilde{Q}_p$ is unbiased, then it follows $\mathbb{E}[\tilde{Q}_p|x_p] = \mathbb{E}[Q_p|x_p]$. I rely on this equality in the following steps described below.

In the main specification, I use a paper's rank in citations as $\tilde{Q}_p$. In Section 5.5, I discuss the implications of using citations as my measure of paper quality, and explore other possible notions of paper quality that reviewers may consider, including when I use a paper's rank in journal-weighted publication status as $\tilde{Q}_p$, or other transformations of citations instead.

I take author traits, x_p, as the vector of applicant student status, institution rank, gender, PI gender, and the paper's relevant subfields reported by the applicant. Considering x_p as a vector allows me to estimate how a paper's expected quality and bias changes when changing a single trait (e.g. student status) while holding the rest fixed. The assumption going forward is that these dimensions of author traits capture the dimensions of what author names convey to reviewers.[28]

To conduct the decomposition in Equation 5, while accounting for conditional relationships (e.g. controlling for subfield), I impose functional forms on both the conditional paper quality expectation function, $\mathbb{E}[Q|x, R(v)]$, and bias function, $t(x, R(v))$. This allows me to integrate score and quality disparities over the marginal distribution of submission content.

[28] This is an assumption, since the experimental variation in this study is whether a reviewer received names or not. One may imagine other traits that are signaled through name that reviewers use during evaluations, such as the fame associated with an author. To address some of this concern, I find that the results presented are robust to controlling for the number of historical citations associated with each applicant and PI, by including it in x_p.

Specifically, I first estimate the average extent of accurate statistical discrimination by imposing that the conditional expectation function for paper quality is linear in submission content rankings:

$$\mathbb{E}[Q_p|x_p, R(v_p)] = \delta_0^Q + \delta_1^Q R(v_p) + x_p'\delta_2^Q + x_p'R(v_p)\delta_3^Q \tag{6}$$

The coefficients δ_2^Q and δ_3^Q capture subgroup differences. I estimate Equation 6 by regressing a paper's ranking in citations on the vector of author traits, Blind score percentile, and their interactions.[29] The predicted values, denote as $\widehat{EQ}_p$, from the regression estimate the conditional expectations of paper quality. Note that I do not rely on the assumption that directly observe a paper's latent quality, Q_p: I rely on the assumption that I observe an unbiased measure of it ($\tilde{Q}_p$), so that the predicted values from regressing $\tilde{Q}_p$ on $(x_p, R(v_p))$ recovers an unbiased estimate for $\mathbb{E}[Q_p|x_p, R(v_p)]$. Figure A7 shows the distribution of these predicted values.

To estimate the extent of bias, I first subtract a paper's Non-Blind score from its estimated conditional expectation in paper quality: $\hat{t}(x_p, R(v_p)) \equiv \widehat{EQ}_p - S_p^{NB}$. Note that I use a paper's expected quality rather than its realized quality, because following the model, the paper-specific outcome is not in the reviewer's information set at the time of review. This estimate assumes a mapping between the scale of reviewer scores and scale in citation percentile ranks: a ten percentile point increase in the expected percentile rank of paper quality translates to a one point increase in reviewer score.[30] I then impose linearity on the bias function:

$$t(x_p, R(v_p)) = \delta_0^\tau + \delta_1^\tau R(v_p) + x_p'\delta_2^\tau + x_p'R(v_p)\delta_3^\tau \tag{7}$$

The δ^τ coefficients are estimated by regressing $\hat{t}(x_p, R(v_p))$ on the above covariates. The coefficients δ_2^τ and δ_3^τ capture bias against subgroups.

Imposing functional forms on the quality expectation and bias functions allows me to integrate subgroup differences in expected paper quality and bias over the marginal distribution of submission content rankings. Let $\hat{F}(\cdot)$ denote the empirical distribution of submission con-

[29] Since scores are bounded between 1 and 10, I transform quality percentile rankings to be bounded between 1 and 10. Let Q be a paper's quality percentile ranking. I then calculate $Q \times 9/10 + 1$. I alternatively could have transformed reviewer scores to be bounded between 0 and 10, but transformed the quality measure to decompose the Non-Blind score gap in score units. The results are qualitatively the same if I only divide percentile rankings by 10, keeping Q as bounded between 0 and 10. In this case, papers of median quality in expectation receive a score of 5 from reviewers that only engage in accurate statistical discrimination.

[30] For instance, a paper expected to be in the median percentile of citations receives a score that is five points below a paper expected to be in the top percentile of citations.

tent rank, $R(v)$, as measured by Blind score rankings. The average level of accurate statistical discrimination contributing to the Non-Blind score percentile gap (first component of eq. 5) is captured by $\delta_2^Q + \int_v \delta_3^Q R(v)\, d\hat{F}(R(v))$. Similarly, the average contribution of bias ($\Delta_{x,x'}$) is $\delta_2^\tau + \int_v \delta_3^\tau R(v)\, d\hat{F}(R(v))$. The difference between the two estimates the total Non-Blind score gap.

5.4 Results

Before showing the estimated results as described in the previous section, I show the main ingredients that are used to estimate the model: disparities in Non-Blind scores and paper quality along the distribution of submission content, as measured by Blind scores. Figure 4 illustrates, using a kernel-weighted local polynomial regression. The figures in the left column show disparities in Non-Blind scores. Subgroup differences in the levels show that applicants who are students, or from lower ranked institutions, are scored significantly lower by reviewers than their counterparts, even after controlling for submission content. On average, Non-Blind reviewers scored students around 0.35 points ($\approx$ 0.18 SD) lower than non-students, after controlling for subgroup differences in submission content.

Can these gaps be rationalized by accurate statistical discrimination? The figures in the right column of Figure 4 show differences in average paper quality, and Table 6 shows the regression results from jointly considering each trait. First, as in the previous section, I find that a paper's submission content (percentile rank in blind scores) is informative of its paper quality (rank in citations). This can be gleaned from the upwards-sloped lines in the right panel of Figure 4. Additionally, when regressing paper quality on submission quality and its interactions, the p-value of a joint F-test on submission content and its interactions with author characteristics is less than 0.001 (column 8 of Table 6).

There is not a substantial difference in paper quality across student status, conditional on submission content. Non-Blind score disparities by student status cannot be fully explained by accurate statistical discrimination, because paper quality conditional on submission content does not significantly differ across student status (column 7 of Table 6). In fact, among papers of the best submission content, students are associated with slightly higher quality papers than non-students, though the difference is not statistically significant. Because these students are still scored significantly worse than non-students under non-blind review, I find that bias against students is highest among those with the best submission content (column 10 of Table 6). Under the model, if reviewers scored papers only based on paper quality, the gap in scores would be 0.13 points in favor of students, in contrast to the actual disparity of 0.35 points against students (Figure 5). This suggests that accurate statistical discrimination fails to explains the Non-Blind score gap. While my estimates are noisy, using 95% confidence intervals from clustered bootstraps drawn at the reviewer level, I reject that accurate statistical discrimination can explain more than 25% of the gap (Table 7), so that at least 75% of the gap is not explained by accurate statistical discrimination.

What does this unexplained gap in scores, "bias" against students, represent? One possibility is inaccurate beliefs, where reviewers over-inflate the conditional quality difference between student papers and non-student ones. Another is that reviewers are more likely to know non-students, and engage in favoritism towards papers submitted by these authors, whether consciously or not. Thirdly, as noted earlier, mis-specifications of the reviewer's objective function can also affect estimates of bias. Although the result that accurate statistical discrimination cannot fully explain the student score gap holds whether I use a paper's rank in citations or journal-weighted publication status, it is possible that reviewers incorporate other notions of paper quality. For instance, reviewers may act upon the belief (whether true or not) that non-student speakers draw more attendees to the conference relative to student ones. I am not able to test whether sessions with non-student speakers attract more participants than those with student speakers given data limitations, but I explore how using alternative transformations of citations—such as whether a paper reaches the top decile in citations—may affect interpretations in Section 5.5.

One additional concern may be omitted variables in the vector of author traits, x_p. I investigate whether these patterns by student status are driven by differences in applicants' past publications, a trait that is omitted from the baseline specification. I find that the pattern persists even after including the cumulative number of historical citations that a paper's applicant and PI are associated with at the time of review, suggesting that the result is not coming from reviewers preferring authors with a more prolific academic history.

In contrast to disparities on the basis of student status, accurate statistical discrimination can explain a sizable portion of the Non-Blind score gap between applicants from top ranked and lower ranked institutions. On average, the score gap between applicants from lower ranked institutions and those from top 20 ranked institutions was 0.50 points, after controlling for differences in submission content. Point estimates suggest that if reviewers were scoring solely based on expected paper quality, then the difference in scores would be 0.41 points, implying that around 18% of the gap remains unexplained by accurate statistical discrimination. This is consistent with institutional bias, but the confidence interval for this estimate is wide and I fail to reject that 100% of the gap can be explained by expected paper quality differences.

Turning to gender, I find that accurate statistical discrimination explains a sizable portion of both the applicant and PI gender disparities that arise from non-blind review. Taking the estimates literally implies that on average, female applicants and PIs receive favoritism, since average male-female differences in expected paper quality is of greater magnitude than the gap in Non-Blind scores. However, the confidence intervals on these estimates are very wide, which also reflects the fact that females are a minority in the sample.

Ultimately, the effects of blinding on composition and quality depend on how reviewers use author identities during evaluations. The above exercise illustrates, using a paper's rank in citations as a benchmark, that the distribution of Non-Blind reviewer scores is not always consistent with accurate statistical discrimination. While my model estimates are noisy, I illustrate how combining the framework with data on a paper's quality, and Blind and Non-Blind evaluation outcomes, allows for disentangling and testing for potential mechanisms. This is an important contribution, given that past work of blinding generally has not incorporated data on subsequent candidate quality, nor has used it to quantify underlying forms of discrimination.

5.5 Measuring Paper Quality

The analyses above use rankings in citations as an (statistically) unbiased measure of paper quality. Citations are often used as proxies for paper quality, both in past research (e.g. Smart and Waldfogel, 1996) and in consequential, real-world hiring and promotion decisions. In reality, true paper quality is likely multidimensional and unobserved. However, understanding the relationships between evaluation outcomes and such measures is still important, given that they do capture dimensions of a paper's perceived quality and prominence. Below, I discuss a few potential concerns one may have with measuring paper quality.

First, it is possible that citations and publication processes themselves contain evaluator biased (e.g. Jin et al., 2019; Card et al., 2020; Koffi, 2021), though empirically there is mixed evidence on this (e.g. Smart and Waldfogel, 1996). I find that the subgroups that perform worse under Non-Blind review (those from lower ranked institutions, female applicants and PIs) generally also have fewer citations than the subgroups that perform better under Non-Blind review. This could be at least partly because the citations process is also biased against authors of these subgroups. This would cause me to under-estimate the extent of bias relative to "true" paper quality. To examine the implications of this, I explore how results change if I inflate the citations of traditionally lower-scoring subgroups by 10 percent (Table A15). I find that if citations for student authors were deflated relative to true latent quality by 10 percent, at most 7 percent of the student score gap is attributable to accurate statistical discrimination.

Second, even if the measure of quality in the data is in line with the notion of quality that evaluators base scores on (e.g. citations), it is important to consider how these measures may be used by reviewers. For instance, while the section above proxied the reviewer's quality expectation using an un-weighted average over paper's rankings in citations, reviewers may actually assign scores in line with the intent of identifying the best papers, placing a greater weight on submissions with an exceptionally high expectation in future paper quality. This also reflects the possibility of reviewers' risk aversion. I conclude by exploring whether and how the student score gap decomposition results change with the aspects of a paper's citations that reviewers may optimize scores over. Table 7 summarizes.

To test whether and how the decomposition changes if reviewers were to place greater weight on rewarding the highest-citation papers, I use an indicator of whether a paper is in the top decile of citations as the proxy measure of paper quality, $\hat{Q}_p$. Returning to the model, this implies that reviewers assign scores based on the probability that a paper has many (in top decile) citations. I find that this change increases the magnitude of the student score gap that can be attributed to accurate statistical discrimination, which reflects how students are slightly less likely than non-students to reach the top decile in citations, though this difference is not precisely estimated.

Another possibility is that reviewers seek to prevent the lowest-quality submissions from being accepted. To test the implications of this, I define paper quality as whether it has any citations: reviewers score based on the probability that a paper has any citations at all. I find that this change reduces the share of the non-blind score gap that is attributable to accurate statistical discrimination. This is because papers submitted by students are not less likely to have citations than those by non-students (Table 5).

6 Generalizability

How do the results from this study translate to other contexts? This likely depends on numerous factors. For instance, in terms of gender composition, the experimental sample appears to be a more gender-imbalanced context relative to various other fields (Figure A8), and this may affect the underlying levels of subgroup quality differences (which determines the scope for accurate statistical discrimination), evaluator bias, and thus the effects of blinding. However, while the exact estimates from this paper may not replicate in another context, the importance of understanding how blinding impacts composition and quality persists in settings beyond the one studied here. My methods show the value of collecting Blind evaluation outcomes, Non-Blind evaluation outcomes, and quality measures, for each candidate for understanding how blinding impacts evaluator discrimination. The insights from this paper are also applicable to settings beyond blinding. I illustrate a way in which these ingredients can be used to estimate a model of discrimination, and quantify underlying forms of discrimination. I discuss generalizability in greater detail in Appendix E, following the SANS (Selection-Attrition-Naturalness-Scaling) conditions (List, 2020).

7 Conclusion

This paper studies the effects of blinding evaluations, using a natural field experiment that collects both Blind and Non-Blind scores for each paper submitted to an academic conference. Blinding alters the allocation of scores across demographic groups, such that previously under-performing groups, particularly applicants who are students and from lower-ranked institutions, benefit. These effects persist across the distribution of acceptance thresholds, and students are substantially more likely to be accepted to the conference under blind review. Despite these compositional changes, blind scores perform no worse in predicting subsequent paper quality than non-blind scores. Through the lens of a reviewer decision-making model, the results suggest that these changes in composition and lack of change in quality are because reviewers use author identities to both update beliefs about paper quality, and engage in bias. Blinding prevents accurate statistical discrimination, but also reduces scope for bias, and these two mechanisms pull the effects of blinding on quality in competing directions.

Tables and Figures

Table 1: Author Traits

	Mean	SD
Applicant Traits		
Student: Yes	0.51	0.50
Student: No	0.49	0.50
Student: Unknown	0.00	0.04
Has Institution Rank	0.84	0.37
Institution Rank | Have Rank	72.79	165.70
Institution Rank: Top 20	0.40	0.49
Institution Rank: 20+	0.44	0.50
Institution Rank: Not University	0.15	0.36
Institution Unknown	0.01	0.12
Gender: Female	0.23	0.42
Gender: Male	0.76	0.42
Gender: Unknown	0.00	0.04
PI Traits		
Has Institution Rank	0.84	0.37
Institution Rank | Have Rank	76.04	169.23
Institution Rank: Top 20	0.38	0.49
Institution Rank: 20+	0.46	0.50
Institution Rank: Not University	0.15	0.36
Institution Unknown	0.01	0.09
Gender: Female	0.17	0.37
Gender: Male	0.82	0.38
Gender: Unknown	0.01	0.09
Other		
Number of Authors	3.90	2.25
Solo Author	0.03	0.17
Share of Coauthors Female	0.08	0.14
Share of Coauthors Female | Has Coauthors	0.11	0.15
Observations (Papers)	657	

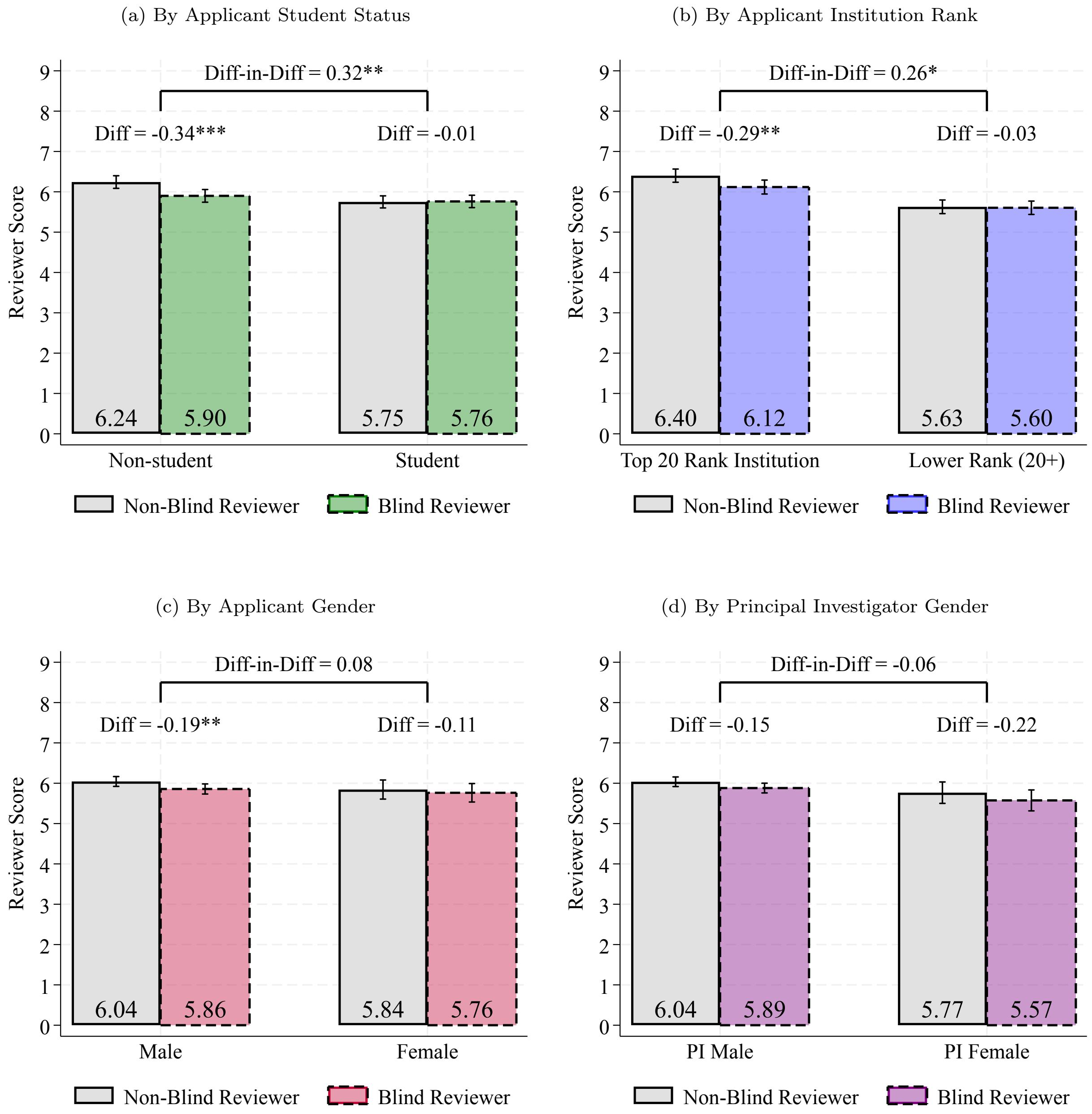
Notes. This table provides descriptive statistics for the full sample of papers submitted to the conference in 2020, which is the sample used in the experiment. Observations are at the paper level. Note that since each paper was assigned to Blind and Non-Blind reviewers, this represents the traits associated with both the Blind and Non-Blind papers.

Table 2: Reviewer Balance Table

	All		**Non-Blind**		**Blind**		**Difference**	
	Mean	SD	Mean	SD	Mean	SD	Diff	p-val
Gender: Female	0.35	0.48	0.36	0.48	0.33	0.47	-0.03	0.65
Gender: Male	0.65	0.48	0.63	0.48	0.67	0.47	0.04	0.55
Years Since PhD	6.56	4.49	6.51	4.79	6.61	4.19	0.10	0.88
Student	0.02	0.15	0.02	0.13	0.03	0.18	0.01	0.45
Inst: Top 20	0.37	0.48	0.39	0.49	0.35	0.48	-0.05	0.46
Inst: 20+	0.43	0.50	0.44	0.50	0.42	0.50	-0.02	0.80
Inst: Not Uni	0.18	0.38	0.14	0.35	0.21	0.41	0.06	0.19
Inst: Rank Unknown	0.02	0.15	0.03	0.16	0.02	0.15	0.00	0.94
Has Reviewed Pre-2020	0.92	0.27	0.92	0.28	0.92	0.27	0.00	0.89
N Reviewed Pre-2020	1.82	1.42	1.72	1.21	1.92	1.61	0.20	0.39
N	245		119		126			

Notes. This table shows summary statistics for Blind and Non-Blind reviewers. The final two columns show the mean difference and p-values from single hypothesis t-tests between the means of Non-Blind reviewers and means of Blind reviewers. “Has reviewed pre-2020” and “N reviewed pre-2020” correspond to whether the reviewer has reviewed in 2019 (one year before the experiment) or 2018 (two years before the experiment), and the number of times if so.

Figure 1: Reviewer Scores and Effects of Blinding


(a) By Applicant Student Status

(b) By Applicant Institution Rank

(c) By Applicant Gender

(d) By Principal Investigator Gender

Notes. Differences are calculated by a regression using paper fixed effects. * $p < .1$, ** $p < .05$, *** $p < .01$ P-values adjusted for multiple hypothesis testing, using Theorem 3.1 of List et al. (2019), for the difference-in-differences across each of the characteristics, estimated using equation 1, are 0.05, 0.22, 0.29, 0.79, respectively.

Table 3: Robustness for Effects of Blinding on Reviewer Scores

	Main	Clust.Paper	NotOnline	NotMiss	IPW
	(1)	(2)	(3)	(4)	(5)
Student × Blind	0.33**	0.33**	0.31**	0.33**	0.33**
	(0.13)	(0.14)	(0.14)	(0.14)	(0.13)
Lower Rank Inst. × Blind	0.28*	0.28*	0.20	0.27*	0.27*
	(0.15)	(0.15)	(0.16)	(0.16)	(0.16)
Female × Blind	0.23	0.23	0.32*	0.20	0.20
	(0.16)	(0.15)	(0.17)	(0.17)	(0.17)
Has Female PI × Blind	-0.06	-0.06	-0.04	-0.10	-0.11
	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)
Paper FE	×	×	×	×	×
Reviewer FE	×	×	×	×	×
N	2591	2591	2170	2480	2480
N Clusters	245	657	245	245	245
N Papers	657	657	551	620	620
R^2	0.57	0.57	0.59	0.57	0.57

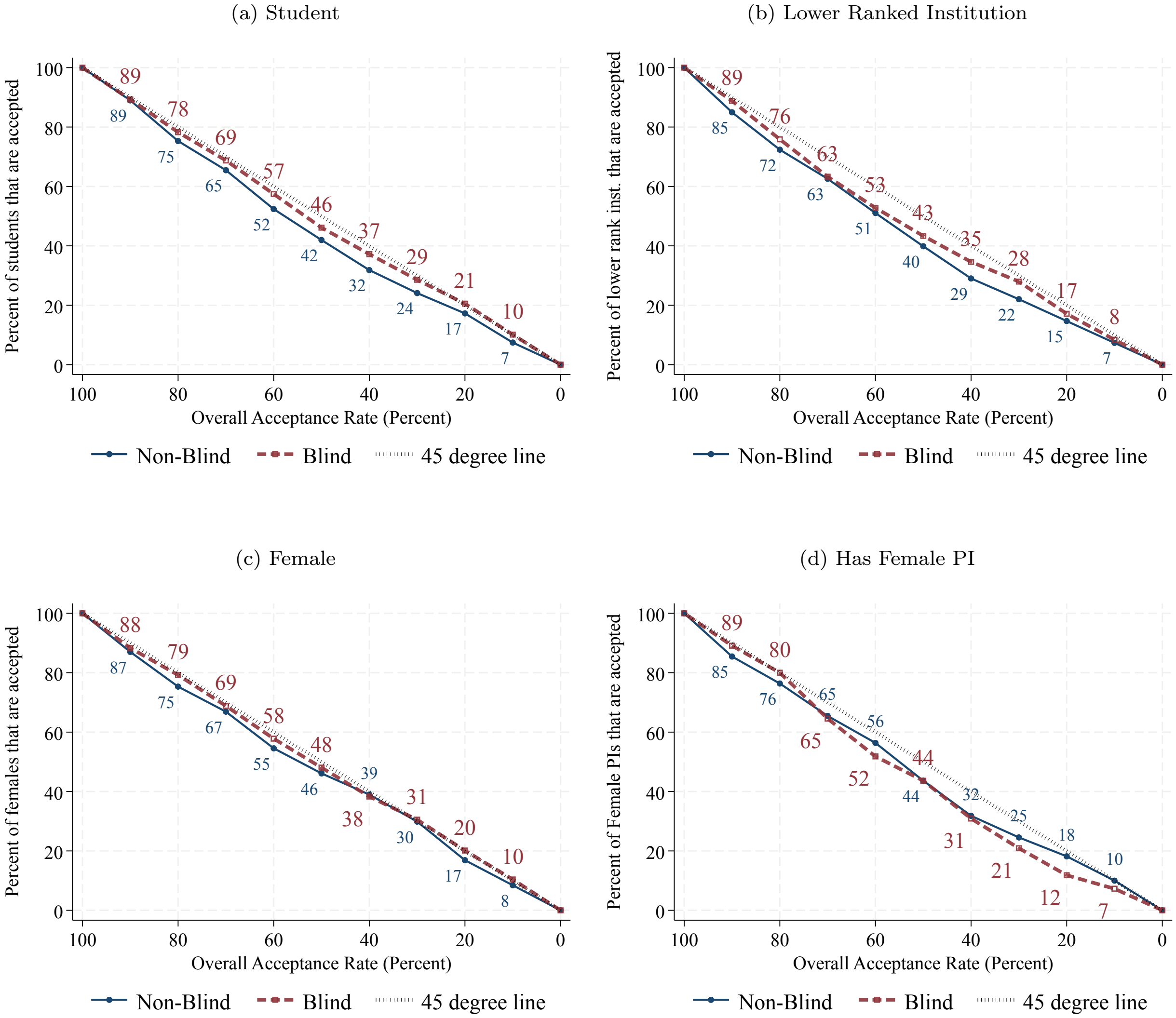
Notes. This table examines the robustness of blinding effects on reviewer scores. The first column presents the main effects from the main text. The subsequent columns are deviating from the main specification by: (2) clustering at the paper-level (3) sub-setting to papers that were not available online at the time of the review process (4) sub-setting to papers that are not missing a review (5) using inverse probability weighting to acknowledge the fact that some papers were more likely to be missing one of their four reviews. Observations are at the paper-reviewer level. Dependent variable is the score that a reviewer gave to a paper. Standard errors in parentheses, clustered at the reviewer level.

Table 4: Effects of Blinding on Acceptances

		Simulated		
	Actual	Non-Blind	Blind	Diff
	(1)	(2)	(3)	(4)
Student	-0.07*	-0.12***	0.00	0.12**
	(0.04)	(0.04)	(0.04)	(0.05)
Lower Rank Inst.	-0.19***	-0.18***	-0.16***	0.02
	(0.04)	(0.04)	(0.04)	(0.05)
Female	-0.01	-0.09*	-0.04	0.05
	(0.05)	(0.05)	(0.05)	(0.05)
Has Female PI	-0.03	-0.05	-0.07	-0.02
	(0.05)	(0.05)	(0.05)	(0.07)
N	657	657	657	657
R^2	0.06	0.08	0.05	0.04

Notes. This table tests the impacts of blinding on acceptance disparities. The first column shows disparities in realized acceptances, which was determined by both Non-Blind and Blind scores. Overall acceptance rate was 61% (402 out of 657 papers accepted). The dependent variable is whether the paper was admitted to the conference. The second column shows what disparities would have been if the conference only used Non-Blind scores to determine acceptances. Dependent variable for this column is whether a paper scored in the top 61% of Non-Blind scores. The third column shows what disparities would have been if the conference only used Blind scores to determine acceptances. Dependent variable for this column is whether a paper scored in the top 61% of Blind scores. The fourth column shows the effects of blinding on these acceptance gaps. The dependent variable for this column is the change in acceptance induced by blinding: the difference in the dependent variable from the second and third columns. Observations are at the paper-level. Heteroskedastic robust standard errors in parentheses.

Figure 2: Effects of Blinding on Composition, by Overall Acceptance Rate



(a) Student

(b) Lower Ranked Institution

(c) Female

(d) Has Female PI

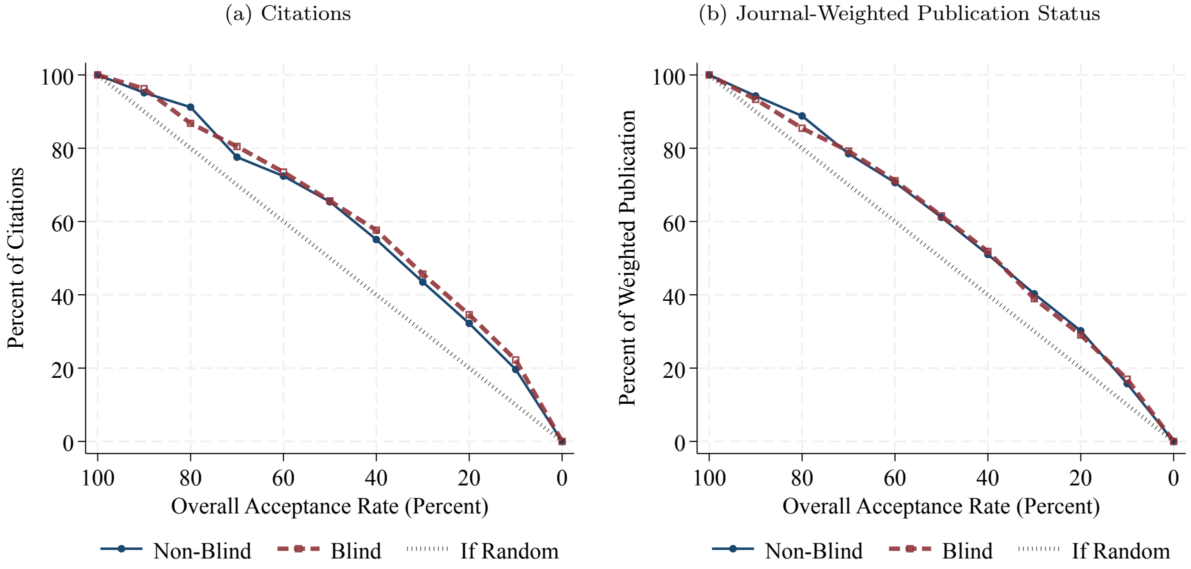
Notes. These figures present simulations of the impacts of blinding on acceptance rate disparities. The x-axis in each figure is the overall acceptance rate for which the acceptance outcome is simulated are for, so that the rightmost side corresponds to accepting fewer papers overall. The y-axis reflects the percentage of the relevant demographic that is accepted. "Lower ranked" institution corresponds to those below a rank of 20, which also corresponds to the median rank. The blue solid line corresponds to when acceptance outcomes are determined by Blind scores only. The gray dashed line corresponds to when acceptance outcomes are determined by Non-Blind scores only. All acceptance outcomes are simulated by assuming that papers that scored in the top $X\%$ of Blind scores are accepted under blinding for an $X\%$ overall acceptance rate, and similarly for Non-Blind. Acceptances are simulated using a paper's reviewer-residualized average Blind and average Non-Blind score.

Table 5: Paper Citations and Publication Statuses Four Years Later

	Citations			Publication Status		
	Has	N	N | Has	Online	Pub	Weighted
All	0.68	19.49	28.84	0.78	0.52	5.16
	(0.47)	(40.75)	(46.79)	(0.41)	(0.50)	(6.93)
			By Applicant Student Status			
Student	0.70	17.32	24.88	0.80	0.53	4.90
	(0.46)	(31.16)	(34.75)	(0.40)	(0.50)	(6.44)
Not Student	0.66	21.83	33.26	0.76	0.52	5.45
	(0.48)	(48.83)	(57.07)	(0.43)	(0.50)	(7.42)
Difference	0.04	-4.50	-8.39*	0.04	0.02	-0.55
	[0.04]	[3.18]	[4.43]	[0.03]	[0.04]	[0.54]
			By Applicant Institution Rank			
Lower Ranked	0.64	15.24	23.82	0.74	0.51	4.39
	(0.48)	(31.84)	(37.18)	(0.44)	(0.50)	(5.67)
Top 20	0.72	24.65	34.02	0.83	0.55	5.85
	(0.45)	(49.67)	(55.58)	(0.38)	(0.50)	(7.28)
Difference	-0.08**	-9.40***	-10.20**	-0.09**	-0.05	-1.46***
	[0.04]	[3.53]	[4.91]	[0.03]	[0.04]	[0.55]
			By Applicant Gender			
Female	0.65	13.03	20.06	0.75	0.50	4.27
	(0.48)	(21.59)	(24.04)	(0.44)	(0.50)	(5.68)
Male	0.69	21.51	31.40	0.79	0.53	5.44
	(0.46)	(44.88)	(51.30)	(0.41)	(0.50)	(7.26)
Difference	-0.04	-8.49**	-11.34**	-0.05	-0.03	-1.17*
	[0.04]	[3.74]	[5.29]	[0.04]	[0.05]	[0.64]
			By PI Gender			
Female	0.64	16.33	25.66	0.76	0.52	5.51
	(0.48)	(29.99)	(34.32)	(0.43)	(0.50)	(6.83)
Male	0.68	20.24	29.65	0.78	0.52	5.08
	(0.47)	(42.76)	(49.00)	(0.41)	(0.50)	(6.96)
Difference	-0.05	-3.91	-3.99	-0.02	-0.01	0.43
	[0.05]	[4.28]	[6.12]	[0.04]	[0.05]	[0.73]

Notes. This table shows summary statistics for the citation and publication statuses for each paper four years after the experiment. Observations are at the paper-level. Each column captures summary statistics for a unique outcome: an indicator for whether the paper has any citations, number of citations (including zeros), number of citations if has a strictly positive number of citations, whether the paper has citations in the top quartile of sample, whether the paper is available online, whether the paper is published, journal-weighted publication status. Each column uses the entire sample of papers besides for "N | Has" which subsets to the papers that have at least some citations. Journal-weighted publication status uses the impact factor associated with the journal that a paper was published in, and takes on value zero if the paper is unpublished. Standard deviations in parentheses and standard errors in brackets. The first row pools the entire sample of papers, and the following rows divide the sample by author traits: applicant student status, applicant institution rank, applicant gender, and principal investigator (PI) gender. "Lower ranked" institution corresponds to those below a rank of 20, which also corresponds to the median rank. "Difference" rows show the difference between the two preceding author traits (which are mutually exclusive), using a t-test comparison of means.

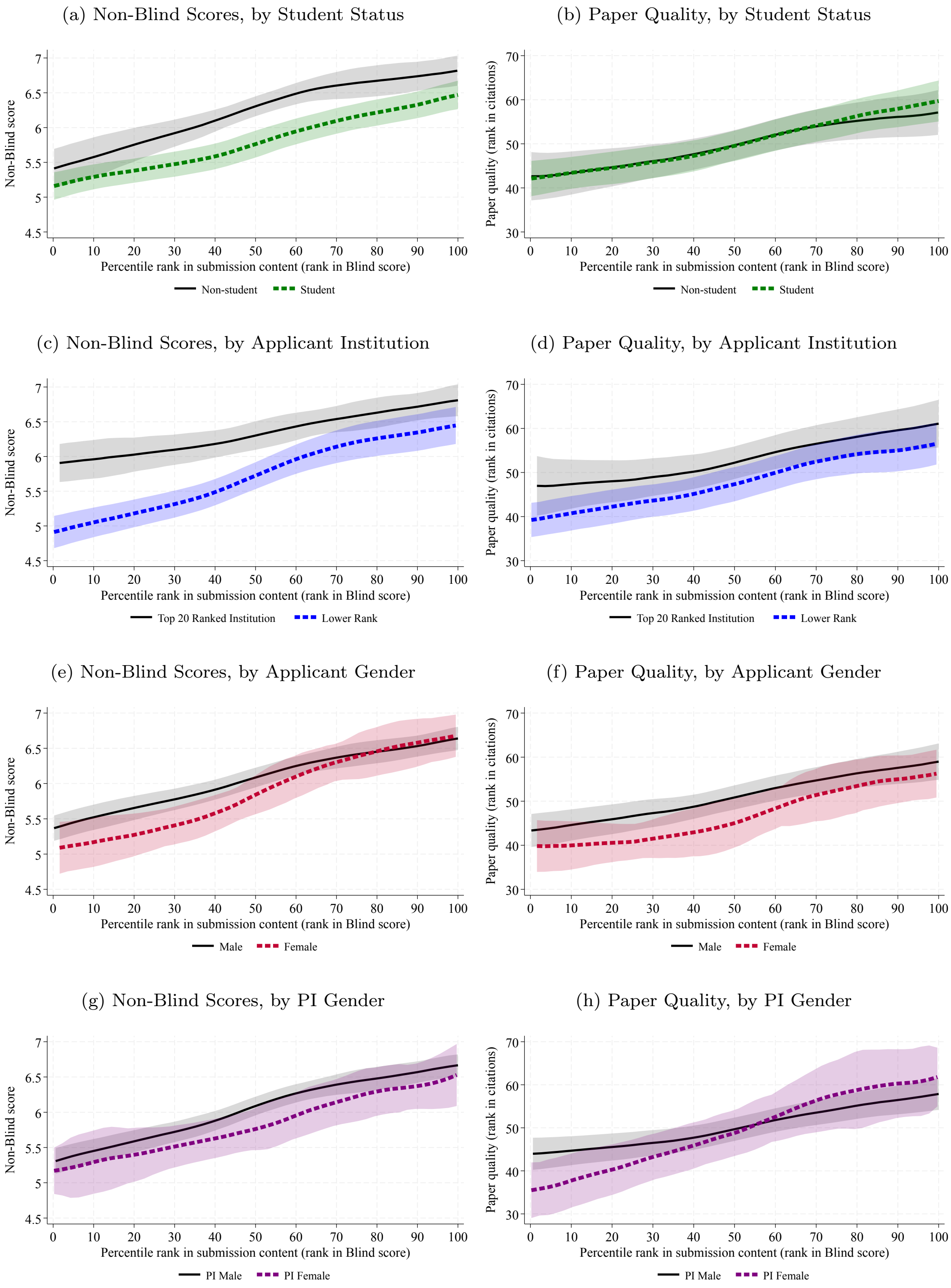
Figure 3: Effects of Blinding on Quality


(a) Citations

(b) Journal-Weighted Publication Status

Notes. These figures illustrate the share of (a) citations or (b) journal-weighted publication statuses that are attributable to accepted papers, for various overall acceptance rates. For instance, (a) shows the share of total citations associated with the papers that would be accepted under non-blind (blue solid line) or blind (red dash line) or random (gray dotted line). Journal-weighted publication status uses the impact factor associated with the journal that a paper was published in, and takes value zero if the paper is unpublished. The red solid line represents the shares when acceptance outcomes are determined by Blind scores only, and the blue solid line for when outcomes are determined by Non-Blind scores only. I assume that papers that scored in the top $X\%$ of Blind scores are accepted under blinding for an $X\%$ overall acceptance rate, and similarly for Non-Blind.

Figure 4: Disparities in Non-Blind Scores and Paper Quality



(a) Non-Blind Scores, by Student Status

(b) Paper Quality, by Student Status

(c) Non-Blind Scores, by Applicant Institution

(d) Paper Quality, by Applicant Institution

(e) Non-Blind Scores, by Applicant Gender

(f) Paper Quality, by Applicant Gender

(g) Non-Blind Scores, by PI Gender

(h) Paper Quality, by PI Gender

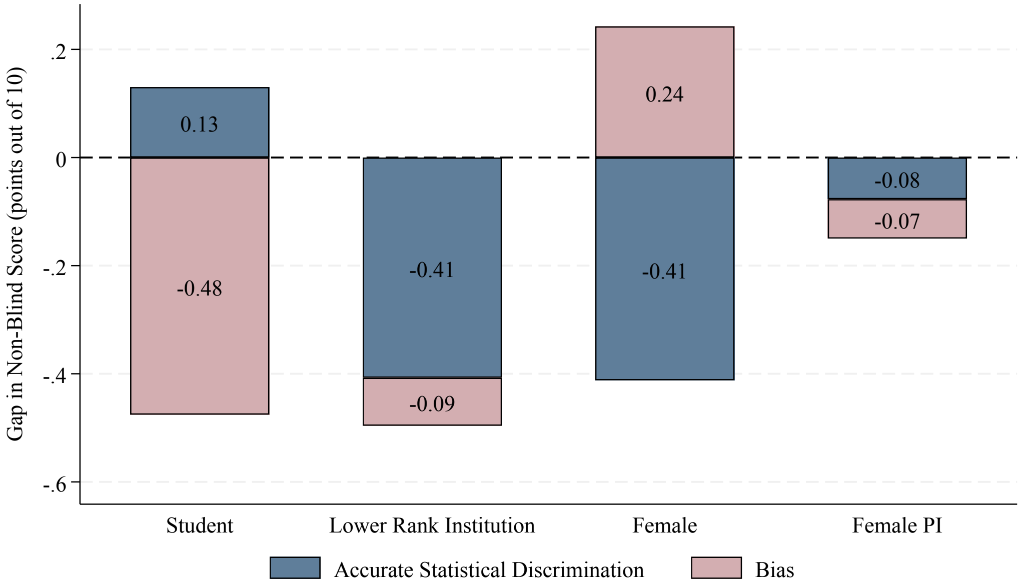
Notes. These figures show disparities in Non-Blind reviewer scores and paper quality along the distribution of submission content, using a kernel-weighted local polynomial regression. Figures on the left column plot a paper's Non-Blind score, after residualizing out reviewer fixed effects. Figures on the right column show paper quality, as measured by a paper's rank in the number of citations four years after the experiment. Submission content is measured by a paper's Blind score, after residualizing out reviewer fixed effects. Shaded areas correspond to 95% confidence intervals, from heteroskedastic robust standard errors. Observations are at the paper level.

Table 6: Estimating Accurate Statistical Discrimination and Bias

	Paper Quality								Bias	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student	-0.20				0.08	0.51	1.20	-0.70	4.64***	1.37
	(2.22)				(2.24)	(2.24)	(2.16)	(4.23)	(1.10)	(2.15)
Lower Rank Inst.		-6.61***			-6.71***	-6.20**	-4.11*	-6.83	0.84	1.75
		(2.42)			(2.44)	(2.44)	(2.39)	(4.36)	(1.24)	(2.20)
Female			-4.13		-4.05	-4.25*	-4.11*	-4.86	-2.38*	-0.15
			(2.52)		(2.53)	(2.55)	(2.45)	(4.65)	(1.33)	(2.53)
Has Female PI				-2.36	-2.45	-2.94	-1.65	-12.08**	-0.10	-10.01***
				(2.98)	(2.96)	(2.96)	(2.81)	(5.07)	(1.53)	(2.84)
Submission Content							0.23***	0.14**	0.04**	-0.01
							(0.04)	(0.07)	(0.02)	(0.03)
Student × Sub. Content								0.04		0.07*
								(0.07)		(0.04)
Lower Rank Inst. × Sub. Content								0.06		-0.02
								(0.07)		(0.04)
Female × Sub. Content								0.01		-0.05
								(0.08)		(0.04)
Has Female PI × Sub. Content								0.23**		0.21***
								(0.10)		(0.05)
Subfield FE						×	×	×	×	×
N	657	657	657	657	657	657	657	657	657	657
R^2	0.00	0.01	0.01	0.00	0.02	0.05	0.10	0.11	0.16	0.19

Notes. This table shows the model estimates for the expected quality and bias functions. The first 3 columns show how author traits and submission content correspond to underlying paper quality, which reflect the scope for accurate statistical discrimination. Dependent variable for columns 1 through 3 is paper quality, measured by a paper's percentile rank in the number of citations four years after the experiment. Submission content is measured by a paper's percentile rank in blind score. Paper quality is measured by a paper's percentile rank in citations. Precentile ranks take on a value between 0 and 100. Bias is estimated by the difference between a paper's expected paper quality (estimated using the predicted values of columns 1 through 3), and Non-Blind score (after residualizing out reviewer fixed effects) multiplied by 10 to match the upper bound of the paper quality measure. Columns 9 through 10 test for bias in Non-Blind scores, by regressing the estimate of bias on author traits and submission content. Predicted values from the specification in column 8 are used as the expected paper quality in columns 9 and 10. Negative valued coefficients on an author trait indicate bias against that subgroup. Columns 8 and 10 are used in the main model specification (Figure 5). Observations are at the paper level. Heteroskedastic robust standard errors in parentheses.

Figure 5: Disentangling Accurate Statistical Discrimination and Bias



Notes. This figure decomposes disparities in Non-Blind scores into the contributions of accurate statistical discrimination (in navy) and bias (in pink), after controlling for submission content (see Equation 5). Each bar uses the full sample of papers, and shows results from a unique decomposition that considers how score disparities are affected by changing the relevant demographic, while holding all other observable traits, including submission content and subfield, constant. Accurate statistical discrimination (navy) is estimated by the difference in mean paper quality, conditional on submission content and other observables (the rest of the observed author traits and subfield). To reach this estimate, I take the estimates from Table 6, integrate the coefficient on the relevant demographic (e.g. coefficient on Student) and the product of its interaction with submission content (e.g. multiply the coefficient on Student × Submission Content with each paper's submission content) over the full distribution of submission content. Bias (pink) is estimated by the gap in Non-Blind score differences that remains unexplained after subtracting out underlying paper quality differences. Negative values indicate bias against the relevant demographic. Submission content is measured by a paper's Blind score ranking, after residualizing out reviewer fixed effects. Paper quality is measured by a paper's ranking in its number of citations that it has four years after the experiment, divided by ten to to match the possible range of reviewer scores (as described in Section 5.3).

Table 7: Percentage of the Non-Blind Score Gap Attributable to Accurate Statistical Discrimination

	Student	Inst Rank	Gender	PI Gender
Citations: Rank	-38	82	243	51
	[-109,25]	[44,176]	[-1917,2861]	[-944,1003]
Citations : Top Decile	50	63	392	10
	[-3,131]	[29,139]	[-3438,4054]	[-799,827]
Citations : Top Quartile	-81	155	477	24
	[-202,8]	[92,317]	[-3999,5540]	[-1092,1095]
Citations: Has	-156	113	98	94
	[-359,-47]	[50,241]	[-949,1632]	[-1470,1964]
Weighted Publication: Rank	-19	47	275	-170
	[-81,38]	[18,110]	[-2246,2824]	[-1559,1291]
Is Published	-95	29	243	-82
	[-236,11]	[-24,123]	[-1843,2865]	[-1140,1540]
Is Online	-144	122	166	-39
	[-331,-49]	[65,256]	[-1608,2115]	[-970,1074]

Notes. This table decomposes disparities in Non-Blind scores into the contribution that is attributable to accurate statistical discrimination, after controlling for submission content (see Equation 5). Each pair of rows shows the decomposition using a different measure of paper quality. The first row shows the decomposition corresponding to the main specification, which uses a paper's ranking in its number of citations that it has four years after the experiment, re-scaled to match the range of potential reviewer scores (as described in Section 5.3). The subsequent rows show the results for when instead the measure of paper quality is: whether the paper has citations in the top decile of the sample, whether the paper has citations in the top quartile of the sample, whether the paper has any citations, the paper's ranking in its journal-weighted publication status, whether the paper is published, whether the paper is available online. Accurate statistical discrimination is estimated by the difference in mean paper quality, conditional on submission content and other observables (the rest of the observed author traits and subfield). To reach this estimate, I take the estimates from Table 6, integrate the coefficient on the relevant demographic (e.g. coefficient on Student) and the product of its interaction with submission content (e.g. multiply the coefficient on Student × Submission Content with each paper's submission content) over the full distribution of submission content. The displayed percentages divide this magnitude by the average gap in Non-Blind scores for the subgroup. 95% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the reviewer level.

References

Agan, A. and S. Starr (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics 133*(1), 191–235.

Aigner, D. J. and G. G. Cain (1977). Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review 30*(2), 175–187.

Arnold, D., W. Dobbie, and P. Hull (2022). Measuring racial discrimination in bail decisions. *American Economic Review*.

Arnold, D., W. Dobbie, and C. S. Yang (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics 133*(4), 1885–1932.

Arrow, K. J. (1973). *The theory of discrimination*, pp. 3–33. Princeton University Press.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017). Does the gender composition of scientific committees matter? *American Economic Review 107*(4), 1207–38.

Bartoš, V., M. Bauer, J. Chytilová, and F. Matějka (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review 106*(6), 1437–1475.

Becker, G. S. (1957). *The economics of discrimination.* University of Chicago press.

Behaghel, L., B. Crépon, and T. Le Barbanchon (2015). Unintended effects of anonymous resumes. *American Economic Journal: Applied Economics 7*(3), 1–27.

Bertrand, M. and E. Duflo (2017). Field experiments on discrimination. In *Handbook of economic field experiments*, Volume 1, pp. 309–393. Elsevier.

Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review 94*(4), 991–1013.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *American Economic Review*, 1041–1067. ISBN: 0002-8282 Publisher: JSTOR.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2019). Inaccurate statistical discrimination: An identification problem. Technical report, National Bureau of Economic Research.

Bohren, J. A., A. Imas, and M. Rosenberg (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review 109*(10), 3395–3436.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–73.

Bowles, H. R., L. Babcock, and K. L. McGinn (2005). Constraints and triggers: Situational mechanics of gender in negotiation. *Journal of personality and social psychology 89*(6), 951.

Breda, T. and S. T. Ly (2015). Professors in core science fields are not always biased against women: Evidence from france. *American Economic Journal: Applied Economics 7*(4), 53–75.

Buckles, K. (2019). Fixing the leaky pipeline: Strategies for making economics work for women at every stage. *Journal of Economic Perspectives 33*(1), 43–60.

Budden, A. E., T. Tregenza, L. W. Aarssen, J. Koricheva, R. Leimu, and C. J. Lortie (2008). Double-blind review favours increased representation of female authors. *Trends in ecology & evolution 23*(1), 4–6. ISBN: 0169-5347 Publisher: Elsevier.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal 17*(2), 372–404.

Canay, I. A., M. Mogstad, and J. Mountjoy (2020). On the use of outcome tests for detecting bias in decision making. *NBER Working Paper*.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics 135*(1), 269–327.

Charness, G., A. Dreber, D. Evans, A. Gill, and S. Toussaert (2022). Improving peer review in economics: Stocktaking and proposals.

Crane, D. (1967). The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 195–201.

Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2021). Measuring bias in consumer lending. *The Review of Economic Studies 88*(6), 2799–2832.

Ferber, M. A. and M. Teiman (1980). Are women economists at a disadvantage in publishing journal articles? *Eastern Economic Journal 6*(3/4), 189–193. ISBN: 0094-5056 Publisher: JSTOR.

Goldberg, P. K. (2012). Report of the editor: American economic review. *American Economic Review 102*(3), 653–65.

Goldin, C. and C. Rouse (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review 90*(4), 715–741.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic literature 42*(4), 1009–1055.

Huber, J., S. Inoua, R. Kerschbamer, C. König-Kersting, S. Palan, and V. L. Smith (2022). Nobel and novice: Author prominence affects peer review. *Proceedings of the National Academy of Sciences 119*(41), e2205779119.

Huber, K., V. Lindenthal, and F. Waldinger (2021). Discrimination, managers, and firm performance: Evidence from "aryanizations" in nazi germany. *Journal of Political Economy 129*(9), 2455–2503.

Jin, G. Z., B. Jones, S. F. Lu, and B. Uzzi (2019). The reverse matthew effect: Consequences of retraction in scientific teams. *Review of Economics and Statistics 101*(3), 492–506.

Kessler, J. B., C. Low, and C. D. Sullivan (2019). Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review 109*(11), 3713–44.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy 109*(1), 203–229.

Koffi, M. (2021). Innovative ideas and gender inequality. Technical report, Working Paper Series.

Krause, A., U. Rinne, and K. F. Zimmermann (2012). Anonymous job applications of fresh Ph.D. economists. *Economics Letters 117*(2), 441–444.

Laband, D. N. and M. J. Piette (1994). Does the" blindness" of peer review influence manuscript selection efficiency? *Southern Economic Journal*, 896–906.

Leibbrandt, A. and J. A. List (2018). Do equal employment opportunity statements backfire? evidence from a natural field experiment on job-entry decisions. Technical report, National Bureau of Economic Research.

Levitt, S. D. and J. A. List (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives 21*(2), 153–174.

Li, D. (2017). Expertise versus bias in evaluation: Evidence from the nih. *American Economic Journal: Applied Economics 9*(2), 60–92.

List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics 119*(1), 49–89.

List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. *NBER Working Paper*.

List, J. A., A. M. Shaikh, and Y. Xu (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics 22*(4), 773–793.

Niederle, M., C. Segal, and L. Vesterlund (2013). How costly is diversity? affirmative action in light of gender differences in competitiveness. *Management Science 59*(1), 1–16.

Petersen, T. and I. Saporta (2004). The opportunity structure for discrimination. *American Journal of Sociology 109*(4), 852–901. Place: US Publisher: Univ of Chicago Press.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review 62*(4), 659–661.

Pleskac, T., E. Kyung, G. Chapman, and O. Urminsky (2024). Blinded versus unblinded review: A field study comparing the equity of peer-review. *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals.

Smart, S. B. and J. Waldfogel (1996). A citation-based test for discrimination at economics and finance journals.

Tomkins, A., M. Zhang, and W. D. Heavlin (2017). Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences 114*(48), 12708–12713. ISBN: 0027-8424 Publisher: National Acad Sciences.

Welch, I. (2014). Referee recommendations. *The Review of Financial Studies 27*(9), 2773–2804.

Appendix For Online Publication

What Do Names Reveal?
Impacts of Blind Evaluations on Composition and Quality

Haruka Uchida

A Collecting Paper Measures

A.1 Author Traits

Applicant and PI information was either received directly from the applicant or collected. Submission forms included fields for the applicant's full name, gender, institution, student-status, and the PI's gender. These self-reported traits have been asked for in previous years, and were never given to reviewers. For the purposes of analysis for this paper, traits that were not self-reported were filled in (about 7%) through online searches done by a team of research assistants after the applications were submitted. This is because the goal of the study is to measure disparities on the basis of actual traits that are perceived by the reviewer, rather than self-reported ones or the propensity to self-report.

Each trait for each paper was collected separately by at least two research assistants, and compared. If there were discrepancies, they were re-collected. There were two PIs and one applicant whose genders could not be discerned. They were coded as "Gender: Unknown". Similarly, the affiliated institutions for four of the PIs could not be identified and they were coded as "Institution Unknown". Historical citation counts for applicants and PIs were collected using Google Scholar, taking the cumulative number of citations associated with an author until 2019 (since the experiment was conducted in the end of 2019). Applicants and PIs who did not have a Google Scholar page (39 and 18%, respectively) were coded as "Citations: Unknown". Institution ranks were collected using the 2020 US News Global Universities Rankings List, which ranks 1,500 universities around the world. This ranking list uses a number of indicators that quantify metrics related to research production and collaboration. For each applicant's and PI's institution, at least two research assistants first verified whether the institution was a university or not. If not, then the institution was coded as "Not University". If it was a university and did appear on the US News rankings list, the rank was recorded. If it did not appear on the US News rankings list, the institution was coded as "Unranked".

A.2 Paper-level Measures

Papers were searched during the experimental review process, and four years after the experiment. Paper were searched online in the same manner in both. Research assistants searched for each paper online, using the title, authors, and abstract together, which included examining author webpages when available. At least two research assistants searched for the same paper, and inconsistencies were resolved by a third. These matchings considered articles that did not necessarily have the same titles, so that the submission title and final paper title do not necessarily have to be the same but the study content did. During the experiment review process, I collect whether the paper was available online. Two and four years later, I collect whether the paper is available online, is published, journal of publication, and the number of citations it is associated with. This used the same search process as before, meaning that titles did not have to remain the same as submission titles but the study content did.

In the analysis of acceptance outcomes, I first convert scores to percentiles. To construct Blind score percentiles, I first residualize out reviewer fixed effects from scores assigned by Blind reviewers, then take the average Blind score for each paper and rank them. In constructing score percentiles, ties were broken by a random number generator, and the results are not dependent on the randomness of tie-breaking. These percentiles were then used to construct the predicted acceptance outcomes. I repeat the same process for Non-Blind scores.

B Additional Tables and Figures for Experimental Results

B.1 Summary Statistics

Table A1: Correlations in Author Traits

	Applicant Trait								**PI Trait**			**Paper Trait**	
	A.F	A.S	A.20	A.21	A.Non	A.Miss	A.C	A.Lo	PI.F	PI.C	PI.Lo	N.Auth	N.Solo
Female	1.00												
Student	0.02	1.00											
Inst Rank: 1-20	0.02	-0.02	1.00										
Inst Rank: 21+	-0.01	0.12**	-0.72***	1.00									
Inst Rank: Not University	-0.04	-0.13**	-0.34***	-0.37***	1.00								
Inst Rank: Missing	0.06	-0.04	-0.10*	-0.10**	-0.05	1.00							
Citations: N	-0.06	-0.16***	-0.01	0.03	-0.03	-0.00	1.00						
Citations: Below Median	-0.01	0.11**	0.03	-0.00	-0.01	-0.08*	-0.13***	1.00					
PI Female	0.03	0.04	0.02	-0.04	0.04	-0.05	-0.05	0.06	1.00				
PI Citations: N	0.06	0.03	0.03	0.01	-0.04	-0.02	0.03	-0.01	-0.09*	1.00			
PI Citations: Below Median	-0.06	0.01	-0.11**	0.05	0.07	0.03	-0.06	0.08*	0.08*	-0.27***	1.00		
Number of Authors	0.06	-0.02	0.07	-0.08*	0.04	-0.05	0.01	-0.04	0.03	0.04	-0.05	1.00	
Solo Author	-0.05	-0.18***	-0.05	0.03	0.03	-0.02	0.06	-0.07	-0.03	-0.05	0.04	-0.22***	1.00

Notes. This table shows the pair-wise correlations between each trait. Observations are at the paper-level.

B.2 Blinding Effect on Score Gaps

Table A2: Mean Reviewer Scores

	Non-Blind		Blind		Diff (Blind - Non-Blind)	
Entire Sample	5.99	(1.99)	5.83	(2.01)	-0.16**	[0.08]
			By Applicant Student Status			
Student	5.75	(1.95)	5.76	(2.01)	0.01	[0.11]
Not Student	6.24	(2.00)	5.90	(2.02)	-0.34***	[0.11]
Difference	-0.49***	[0.11]	-0.14	[0.11]		
			By Applicant Institution Rank			
Lower Ranked	5.63	(2.03)	5.60	(2.01)	-0.02	[0.12]
Top 20	6.40	(1.91)	6.12	(2.01)	-0.28**	[0.12]
Difference	-0.77***	[0.12]	-0.51***	[0.12]		
			By Applicant Gender			
Female	5.84	(2.08)	5.76	(2.03)	-0.08	[0.17]
Male	6.04	(1.95)	5.86	(2.01)	-0.19**	[0.09]
Difference	-0.19	[0.13]	-0.09	[0.13]		
			By PI Gender			
Female	5.77	(1.99)	5.57	(1.93)	-0.19	[0.19]
Male	6.04	(1.98)	5.89	(2.02)	-0.15*	[0.09]
Difference	-0.27*	[0.15]	-0.31**	[0.15]		

Notes. This table shows the average scores given by Blind and Non-Blind reviewers. The first row pools the entire sample of papers, and the following rows divide the sample by author traits: applicant student status, applicant institution rank, applicant gender, and principal investigator (PI) gender. “Lower ranked” institution corresponds to those below a rank of 20, which also corresponds to the median rank. Score means are taken at the paper-reviewer level. Standard deviations in parentheses and standard errors in brackets.

Table A3: Blind and Non-Blind Score Gaps

	Non-Blind Scores					Blind Scores				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student	-0.51***				-0.46***	-0.15				-0.15
	(0.11)				(0.11)	(0.12)				(0.12)
Lower Rank Inst.		-0.77***			-0.74***		-0.56***			-0.56***
		(0.12)			(0.12)		(0.12)			(0.13)
Female			-0.25		-0.25			-0.05		-0.06
			(0.16)		(0.16)			(0.13)		(0.14)
Has Female PI				-0.25	-0.28*				-0.31**	-0.33**
				(0.15)	(0.15)				(0.15)	(0.15)
Reviewer FE	×	×	×	×	×	×	×	×	×	×
N	1289	1289	1289	1289	1289	1302	1302	1302	1302	1302
N Clusters	119	119	119	119	119	126	126	126	126	126
N Papers	657	657	657	657	657	657	657	657	657	657
R^2	0.14	0.16	0.14	0.13	0.18	0.11	0.13	0.12	0.12	0.14

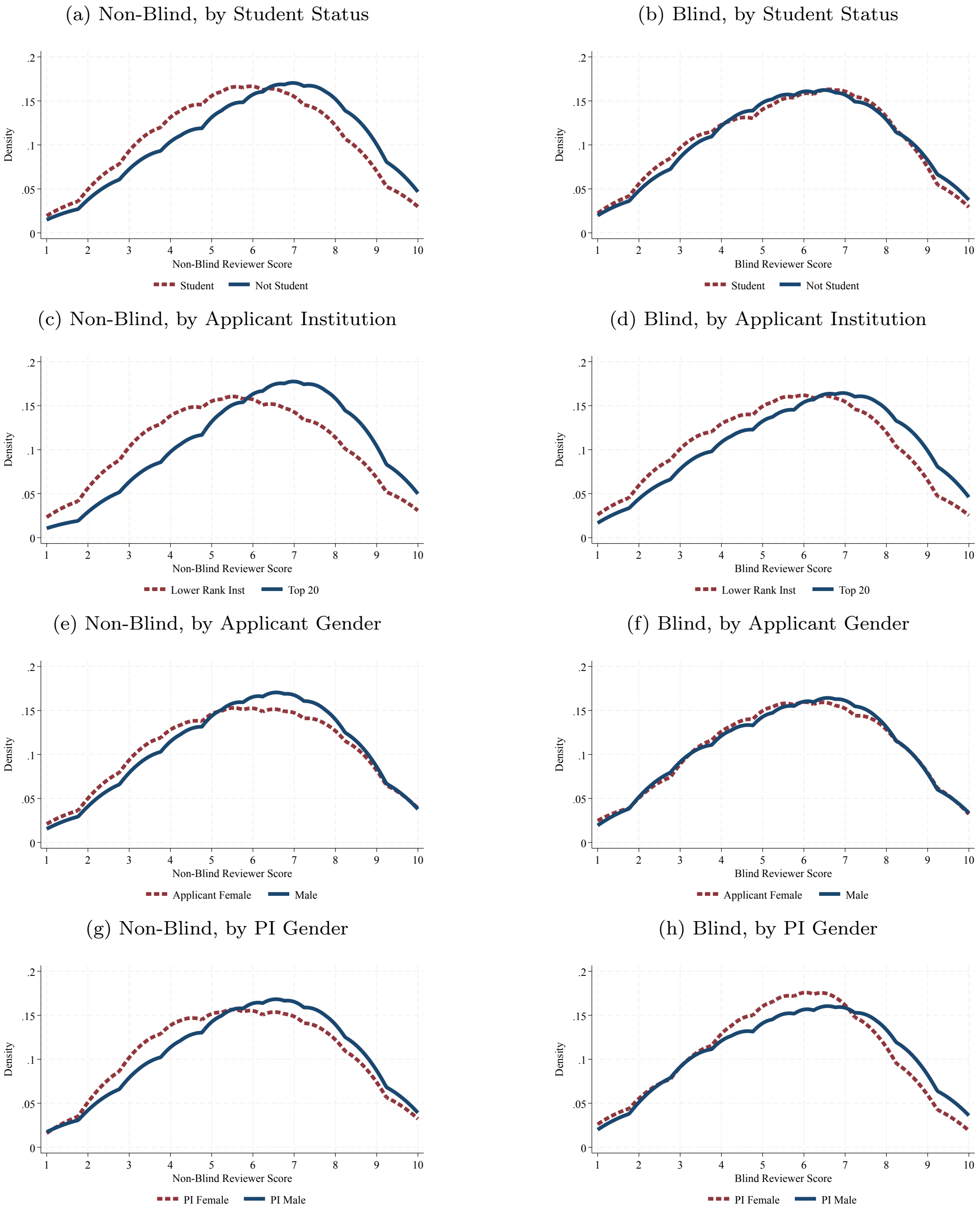
Notes. Dependent variable is the score that a reviewer gave to a paper. Student is an indicator for whether the applicant is a student. "Lower ranked" institution corresponds to those below a rank of 20, which also corresponds to the median rank. Female is an indicator for whether the applicant, the individual who submits the paper and would present it if selected, is a female. Has Female PI is an indicator for whether the principal investigator associated with the paper is a female. Observations are at the paper-reviewer level. Standard errors in parentheses, clustered at the reviewer level.

Table A4: Effects of Blinding on Reviewer Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Blind	-0.16*	-0.17*	-0.34***	-0.29**	-0.19**	-0.16*	-0.44***	
	(0.09)	(0.09)	(0.11)	(0.12)	(0.09)	(0.10)	(0.15)	
Student × Blind			0.32**				0.30**	0.33**
			(0.13)				(0.13)	(0.13)
Lower Rank Inst. × Blind				0.26*			0.22	0.28*
				(0.15)			(0.16)	(0.15)
Female × Blind					0.08		0.08	0.23
					(0.16)		(0.17)	(0.16)
Has Female PI × Blind						-0.06	-0.07	-0.06
						(0.18)	(0.18)	(0.20)
Paper FE		×	×	×	×	×	×	×
Reviewer FE								×
N	2591	2591	2591	2591	2591	2591	2591	2591
N Clusters	245	245	245	245	245	245	245	245
N Papers	657	657	657	657	657	657	657	657
R^2	0.00	0.44	0.44	0.44	0.44	0.44	0.44	0.57

Notes. Dependent variable is the score that a reviewer gave to a paper. Observations are at the paper-reviewer level. Blind is an indicator for whether the reviewer was randomly assigned to score papers without author lists. Student is an indicator for whether the applicant is a student. "Lower ranked" institution corresponds to those below a rank of 20, which also corresponds to the median rank. Female is an indicator for whether the applicant, the individual who submits the paper and would present it if selected, is a female. Has Female PI is an indicator for whether the principal investigator associated with the paper is a female. Standard errors in parentheses, clustered at the reviewer level. P-values adjusted for multiple hypothesis testing, using Theorem 3.1 of List et al. (2019), for the coefficients in the last column are: 0.05, 0.22, 0.29, 0.79, respectively.

Figure A1: Distributions of Blind and Non-Blind Scores by Trait



(a) Non-Blind, by Student Status

(b) Blind, by Student Status

(c) Non-Blind, by Applicant Institution

(d) Blind, by Applicant Institution

(e) Non-Blind, by Applicant Gender

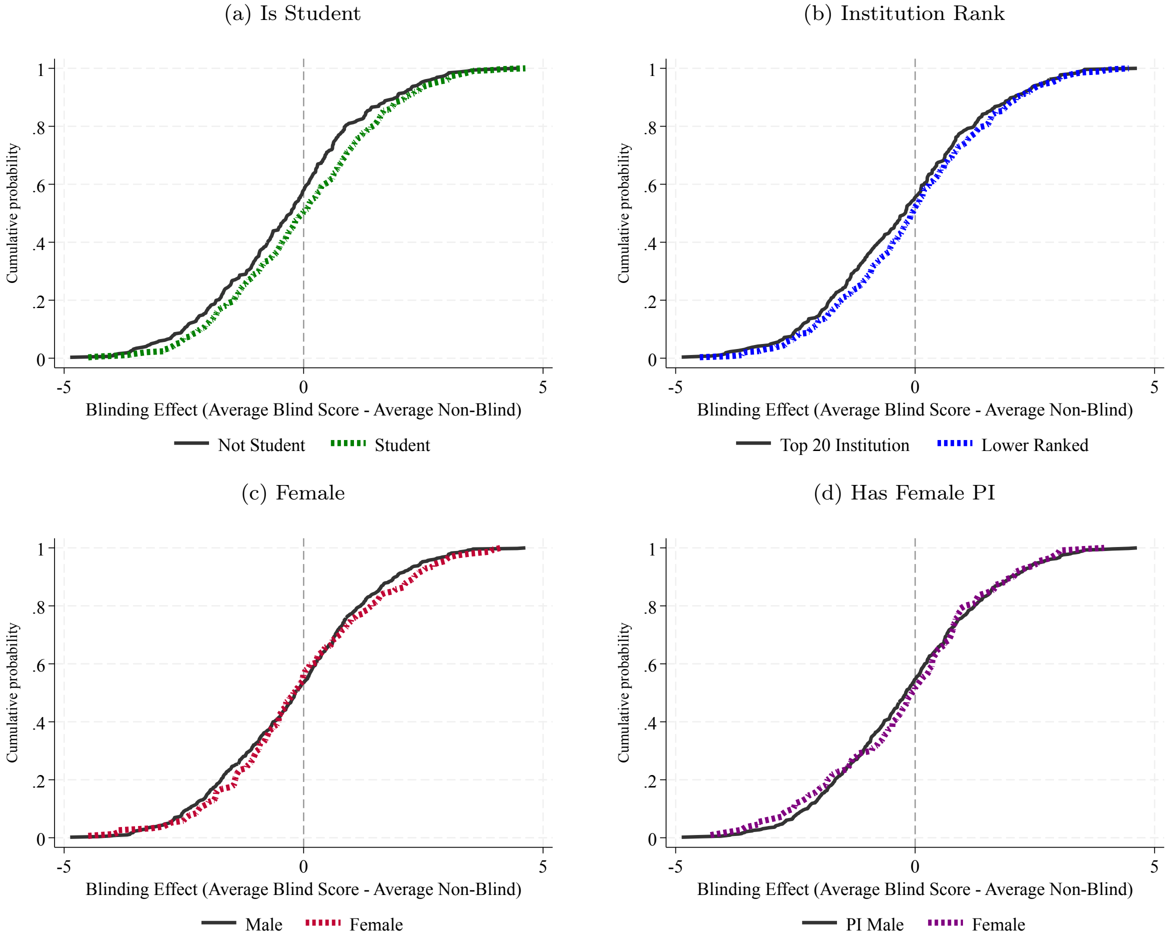
(f) Blind, by Applicant Gender

(g) Non-Blind, by PI Gender

(h) Blind, by PI Gender

Notes. These figures show the kernel density distribution of Blind and Non-Blind scores by each author trait. Observations are at the paper-reviewer level.

Figure A2: Distribution of Paper's Difference Between Blind and Non-Blind Scores

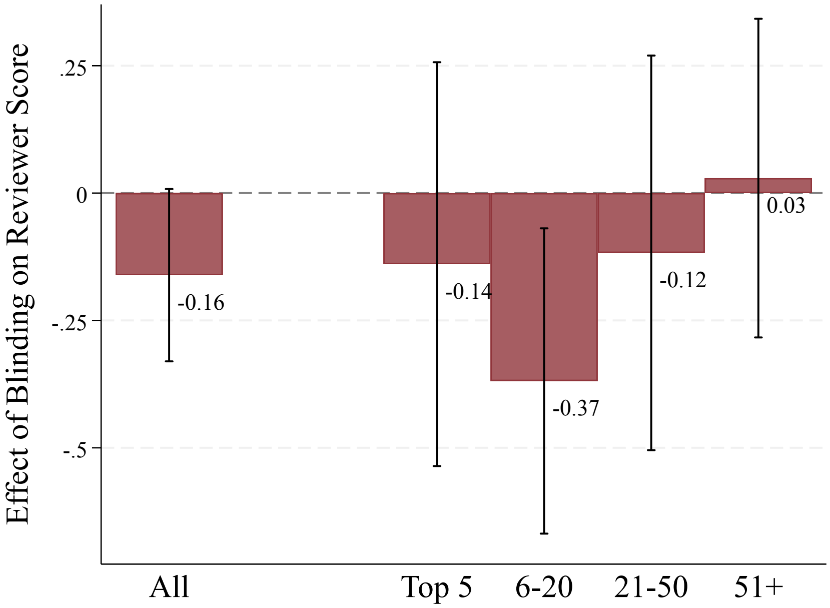


Notes. These figures show the kernel densities for the difference in a paper's average Blind score and its average Non-Blind score, after residualizing out reviewer fixed effects. Observation are at the paper level.

B.3 Heterogeneity By Institution Rank

With respect to applicants' institution ranks, the main text focused differentiated between whether an applicant's affiliated institution rank was better than the median ranking or not. Figure A3 shows the change in scores induced by blinding, binning by finer categories of institution rank.

Figure A3: Effect of Blinding on Scores, by Applicant Institution Rank



Notes. This figure shows the effect of blinding on paper scores. The leftmost bar shows the difference in Blind and Non-Blind scores, using the full sample of papers. The right four bars then repeat the same exercise, but split the full sample into mutually exclusive groups that are defined by the institution rank associated with the paper's applicant. Bar ticks correspond to the 95% confidence intervals.

B.4 Heterogeneity By Coauthor Traits

Table A5: Blinding Effect by Applicant, Coauthor, and PI Traits

	(1)	(2)
Student × Blind	0.32**	0.33**
	(0.13)	(0.13)
Lower Rank Inst. × Blind	0.36*	0.31
	(0.21)	(0.22)
Female × Blind	0.25	0.26
	(0.16)	(0.16)
PI Female × Blind	-0.08	-0.09
	(0.20)	(0.20)
N Coauthors Student × Blind	-0.10	-0.10
	(0.12)	(0.12)
N Coauthors Lower Ranked Inst × Blind	-0.06	-0.05
	(0.08)	(0.09)
N Coauthors Female × Blind	0.04	0.03
	(0.12)	(0.12)
N Coauthors × Blind	0.02	0.01
	(0.08)	(0.08)
Sample	All	MultiAuthor
Reviewer FE	×	×
Paper FE	×	×
N	2591	2516
N Clusters	245	245
N Papers	657	638
R^2	0.57	0.56

Notes. Observations are at the paper-reviewer level. Dependent variable is the score that a paper gets from a reviewer. The first column shows for all papers in the experiment, and the second column shows for papers that had more than one author. Standard errors in parentheses, clustered at the reviewer level.

B.5 Heterogeneity By Applicant and PI's Past Citations

Table A6: Blind and Non-Blind Score Gaps, with Author's Past Citations

	Non-Blind Scores				Blind Scores			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
App. Citations: Above Median	0.25*	0.05			0.07	0.03		
	(0.14)	(0.14)			(0.16)	(0.16)		
PI Citations: Above Median	0.09	0.02			0.22	0.11		
	(0.12)	(0.13)			(0.14)	(0.15)		
Student		-0.44***		-0.42***		-0.08		-0.06
		(0.11)		(0.12)		(0.12)		(0.12)
Lower Rank Inst.		-0.73***		-0.73***		-0.53***		-0.48***
		(0.12)		(0.13)		(0.13)		(0.14)
Female		-0.24		-0.23		-0.04		-0.07
		(0.15)		(0.15)		(0.14)		(0.14)
Has Female PI		-0.27*		-0.25		-0.29*		-0.31*
		(0.16)		(0.15)		(0.16)		(0.16)
App. Citations: Q2			0.32	-0.01			0.39*	0.27
			(0.21)	(0.22)			(0.20)	(0.21)
App. Citations: Q3			0.29	-0.07			0.51***	0.39*
			(0.24)	(0.24)			(0.19)	(0.20)
App. Citations: Q4			0.55***	0.17			0.05	-0.04
			(0.20)	(0.22)			(0.23)	(0.24)
PI Citations: Q2			0.33*	0.13			0.11	0.02
			(0.18)	(0.18)			(0.16)	(0.17)
PI Citations: Q3			0.18	0.05			0.28	0.14
			(0.17)	(0.17)			(0.17)	(0.18)
PI Citations: Q4			0.33*	0.13			0.20	0.05
			(0.18)	(0.17)			(0.20)	(0.20)
Reviewer FE	×	×	×	×	×	×	×	×
N	1289	1289	1289	1289	1302	1302	1302	1302
N Clusters	119	119	119	119	126	126	126	126
R^2	0.13	0.18	0.14	0.18	0.12	0.14	0.13	0.15

Notes. Dependent variable is the score that a reviewer gave to a paper. Student is an indicator for whether the applicant is a student. "Lower ranked" institution corresponds to those below a rank of 20, which also corresponds to the median rank. Female is an indicator for whether the applicant, the individual who submits the paper and would present it if selected, is a female. Has Female PI is an indicator for whether the principal investigator associated with the paper is a female. Observations are at the paper-reviewer level. Standard errors in parentheses, clustered at the reviewer level.

Table A7: Blinding Effect on Scores, with Authors' Past Citations

	(1)	(2)	(3)	(4)
App. Citations: Above Median × Blind	-0.13	0.07		
	(0.16)	(0.17)		
PI Citations: Above Median × Blind	0.15	0.12		
	(0.13)	(0.13)		
Applicant Student × Blind		0.35***		0.37**
		(0.14)		(0.15)
Lower Rank Inst. × Blind		0.29*		0.34**
		(0.15)		(0.16)
Applicant Female × Blind		0.23		0.20
		(0.16)		(0.16)
PI Female × Blind		-0.03		-0.05
		(0.20)		(0.20)
App. Citations: Quartile 2 × Blind			0.02	0.24
			(0.18)	(0.19)
App. Citations: Quartile 3 × Blind			0.15	0.40*
			(0.20)	(0.21)
App. Citations: Quartile 4 × Blind			-0.40*	-0.12
			(0.21)	(0.23)
PI Citations: Quartile 2 × Blind			-0.09	0.01
			(0.18)	(0.19)
PI Citations: Quartile 3 × Blind			0.16	0.14
			(0.17)	(0.18)
PI Citations: Quartile 4 × Blind			0.05	0.07
			(0.19)	(0.19)
Reviewer FE	×	×	×	×
Paper FE	×	×	×	×
N	2591	2591	2591	2591
N Clusters	245	245	245	245
R^2	0.56	0.57	0.57	0.57

Dependent variable is the score that a reviewer gave to a paper. Observations are at the paper-reviewer level. Blind is an indicator for whether the reviewer was randomly assigned to score papers without author lists. Student is an indicator for whether the applicant is a student. "Lower ranked" institution corresponds to those below a rank of 20, which also corresponds to the median rank. Female is an indicator for whether the applicant, the individual who submits the paper and would present it if selected, is a female. Has Female PI is an indicator for whether the principal investigator associated with the paper is a female. "Traits" indicates whether the regression controls for each trait interacted with reviewer blind status. Standard errors in parentheses, clustered at the reviewer level.

B.6 Heterogeneity by Reviewer Traits

Table A8: Blinding Effect Heterogeneity by Reviewer Traits

	(1)	(2)	(3)
Student × Reviewer Female	-0.19		-0.19
	(0.24)		(0.23)
Student × Reviewer Female × Blind	0.28		0.39
	(0.33)		(0.33)
Lower Ranked Inst × Reviewer Female	0.37		0.40
	(0.27)		(0.26)
Lower Ranked Inst × Reviewer Female × Blind	-0.21		-0.31
	(0.34)		(0.35)
Applicant Female × Reviewer Female	0.22		0.20
	(0.29)		(0.28)
Female × Reviewer Female × Blind	-0.26		-0.62*
	(0.36)		(0.36)
PI Female × Reviewer Female	0.39		0.39
	(0.31)		(0.32)
PI Female × Reviewer Female × Blind	-0.75		-0.81*
	(0.46)		(0.48)
Female × Reviewer Lower Inst		-0.37	-0.32
		(0.28)	(0.29)
Female × Reviewer Lower Inst × Blind		-0.66*	-0.82**
		(0.38)	(0.38)
Lower Ranked Inst × Reviewer Lower Inst		-0.29	-0.27
		(0.27)	(0.26)
Lower Ranked Inst × Reviewer Lower Inst × Blind		-0.11	-0.12
		(0.37)	(0.38)
Student × Reviewer Lower Inst		0.08	0.09
		(0.23)	(0.24)
Student × Reviewer Lower Inst × Blind		0.01	0.08
		(0.33)	(0.32)
PI Female Inst × Reviewer Lower Inst		0.26	0.31
		(0.33)	(0.33)
PI Female × Reviewer Lower Inst × Blind		-0.49	-0.65
		(0.51)	(0.53)
Reviewer FE	×	×	×
Paper FE	×	×	×
N	2591	2591	2591
N Clusters	245	245	245
R^2	0.57	0.58	0.58

Dependent variable is the score that a reviewer gave to a paper. Observations are at the paper-reviewer level. Blind is an indicator for whether the reviewer was randomly assigned to score papers without author lists. Student is an indicator for whether the applicant is a student. "Lower ranked" institution corresponds to those below a rank of 20, which also corresponds to the median rank. Female is an indicator for whether the applicant, the individual who submits the paper and would present it if selected, is a female. Has Female PI is an indicator for whether the principal investigator associated with the paper is a female. "Traits" indicates whether the regression controls for each trait interacted with reviewer blind status. Standard errors in parentheses, clustered at the reviewer level.

B.7 By Applicant and PI's Prior Submission Behaviors

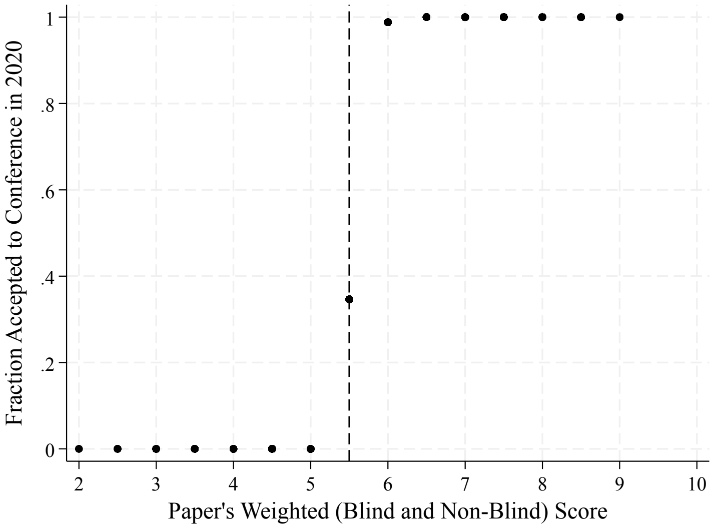
Table A9: Blinding Effects by Whether Applicant or PI Is New

	(1)	(2)	(3)
New applicant × Applicant Student × Blind	-0.24		-0.15
	(0.28)		(0.30)
New applicant × Lower Rank Inst. × Blind	-0.30		-0.28
	(0.34)		(0.37)
New applicant × Applicant Female × Blind	0.13		0.24
	(0.32)		(0.38)
New applicant × PI Female × Blind	0.41		0.27
	(0.39)		(0.44)
New PI × Applicant Student × Blind		-0.37	-0.31
		(0.29)	(0.32)
New PI × Lower Rank Inst. × Blind		-0.13	-0.09
		(0.33)	(0.35)
New PI × Applicant Female × Blind		-0.17	-0.29
		(0.33)	(0.39)
New PI × PI Female × Blind		0.53	0.46
		(0.36)	(0.41)
Paper FE	×	×	×
Reviewer FE	×	×	×
Traits	×	×	×
N	2591	2591	2591
N Clusters	245	245	245
N Papers	657	657	657
R^2	0.57	0.57	0.57

Notes. This table shows average blinding effects on score gaps by whether the applicant or PI was a repeat. I define an applicant or PI as "new" if they have never applied (before the experiment) at least the two years prior to the experiment. Dependent variable is the score that a reviewer gave to a paper. Observations are at the paper-reviewer level. Blind is an indicator for whether the reviewer was randomly assigned to score papers without author lists. Student is an indicator for whether the applicant is a student. "Lower ranked" institution corresponds to those below a rank of 20, which also corresponds to the median rank. Female is an indicator for whether the applicant, the individual who submits the paper and would present it if selected, is a female. Has Female PI is an indicator for whether the principal investigator associated with the paper is a female. "Traits" indicates whether the regression controls for each trait interacted with reviewer blind status. Standard errors in parentheses, clustered at the reviewer level.

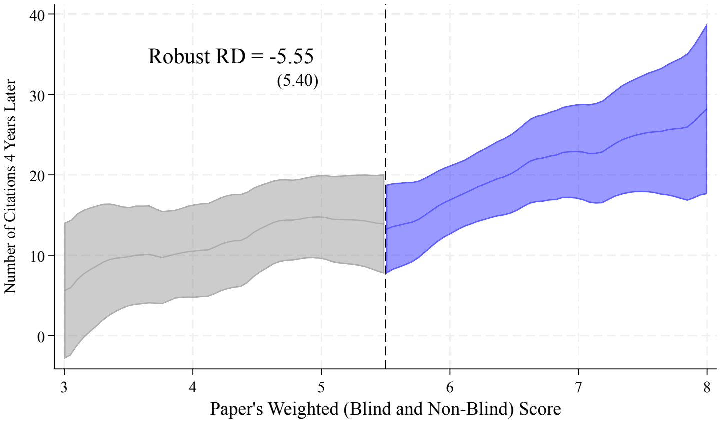
B.8 Impacts on Acceptance Outcomes

Figure A4: Realized Acceptance Status



Notes. This figure shows the fraction of submissions that were accepted to the conference in 2020, given the paper's weighted-average reviewer score (constructed by summing 2/3 of the paper's average blind score and 1/3 of the paper's average non-blind score). The vertical line represents the score of 5.5, which corresponds to the overall acceptance rate, 61% (402 out of 657 papers accepted).

Figure A5: Effects of Conference Acceptance on Citations 4 Years Later



Notes. This figure shows the a paper's citations given its weighted-average reviewer score (constructed by summing 2/3 of the paper's average blind score and 1/3 of the paper's average non-blind score). The vertical line represents the score of 5.5, which corresponds to the overall acceptance rate, 61% (402 out of 657 papers accepted). Those to the right of the cutoff were accepted to the conference and those to the left were not. The regression discontinuity coefficient is estimated following the bias-corrected approach of Calonico et al. (2017).

Table A10: Subgroup Acceptance Rates by Overall Acceptance Rate

	Overall Acceptance Rate																	
	10		20		30		40		50		60		70		80		90	
	Non-Blind	Blind	Non-Blind	Blind	Non-Blind	Blind	Non-Blind	Blind	Non-Blind	Blind	Non-Blind	Blind	Non-Blind	Blind	Non-Blind	Blind	Non-Blind	Blind
									By Applicant Student Status									
Student	0.07	0.10	0.17	0.21	0.24	0.29	0.32	0.37	0.42	0.46	0.52	0.57	0.65	0.69	0.75	0.78	0.89	0.89
	(0.26)	(0.30)	(0.38)	(0.40)	(0.43)	(0.45)	(0.47)	(0.48)	(0.49)	(0.50)	(0.50)	(0.50)	(0.48)	(0.46)	(0.43)	(0.41)	(0.31)	(0.31)
Not Student	0.13	0.10	0.23	0.20	0.37	0.32	0.49	0.43	0.59	0.54	0.68	0.63	0.75	0.71	0.85	0.82	0.91	0.91
	(0.33)	(0.30)	(0.42)	(0.40)	(0.48)	(0.47)	(0.50)	(0.50)	(0.49)	(0.50)	(0.47)	(0.48)	(0.44)	(0.45)	(0.36)	(0.39)	(0.28)	(0.29)
Diff.	-0.05**	0.00	-0.06*	0.01	-0.12***	-0.03	-0.17***	-0.06	-0.17***	-0.08**	-0.16***	-0.05	-0.09**	-0.03	-0.10***	-0.04	-0.02	-0.02
	[0.02]	[0.02]	[0.03]	[0.03]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.03]	[0.03]	[0.02]	[0.02]
Diff. in Diff.	0.05*		0.07*		0.09**		0.11**		0.09*		0.10**		0.07		0.06		0.01	
	[0.03]		[0.04]		[0.04]		[0.05]		[0.05]		[0.05]		[0.04]		[0.04]		[0.03]	
									By Applicant Institution Rank									
Lower Rank Inst.	0.07	0.08	0.15	0.17	0.22	0.28	0.29	0.35	0.40	0.43	0.51	0.53	0.63	0.63	0.72	0.76	0.85	0.89
	(0.26)	(0.28)	(0.35)	(0.38)	(0.42)	(0.45)	(0.45)	(0.48)	(0.49)	(0.50)	(0.50)	(0.50)	(0.48)	(0.48)	(0.45)	(0.43)	(0.36)	(0.32)
Top 20 Inst.	0.15	0.13	0.26	0.25	0.39	0.35	0.52	0.46	0.62	0.59	0.70	0.69	0.78	0.77	0.86	0.86	0.94	0.94
	(0.35)	(0.34)	(0.44)	(0.43)	(0.49)	(0.48)	(0.50)	(0.50)	(0.49)	(0.49)	(0.46)	(0.46)	(0.41)	(0.42)	(0.34)	(0.35)	(0.23)	(0.25)
Diff.	-0.07***	-0.04*	-0.11***	-0.07**	-0.17***	-0.07*	-0.23***	-0.12***	-0.22***	-0.16***	-0.19***	-0.16***	-0.16***	-0.14***	-0.14***	-0.10***	-0.09***	-0.05**
	[0.03]	[0.03]	[0.03]	[0.03]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.03]	[0.03]	[0.03]	[0.02]
Diff. in Diff.	0.03		0.04		0.10**		0.11**		0.07		0.03		0.02		0.04		0.05	
	[0.03]		[0.04]		[0.05]		[0.05]		[0.05]		[0.05]		[0.05]		[0.04]		[0.03]	
									By Applicant Gender									
Female	0.08	0.10	0.17	0.20	0.30	0.31	0.39	0.38	0.46	0.48	0.55	0.58	0.67	0.69	0.75	0.79	0.87	0.88
	(0.28)	(0.31)	(0.38)	(0.40)	(0.46)	(0.46)	(0.49)	(0.49)	(0.50)	(0.50)	(0.50)	(0.50)	(0.47)	(0.46)	(0.43)	(0.41)	(0.34)	(0.32)
Male	0.11	0.10	0.21	0.20	0.30	0.30	0.40	0.41	0.51	0.51	0.62	0.61	0.71	0.71	0.82	0.80	0.91	0.91
	(0.31)	(0.30)	(0.41)	(0.40)	(0.46)	(0.46)	(0.49)	(0.49)	(0.50)	(0.50)	(0.49)	(0.49)	(0.45)	(0.46)	(0.39)	(0.40)	(0.28)	(0.29)
Diff.	-0.02	0.00	-0.04	0.00	0.00	0.00	-0.01	-0.02	-0.05	-0.03	-0.07	-0.03	-0.04	-0.02	-0.06*	-0.01	-0.04	-0.03
	[0.03]	[0.03]	[0.04]	[0.04]	[0.04]	[0.04]	[0.05]	[0.05]	[0.05]	[0.05]	[0.05]	[0.05]	[0.04]	[0.04]	[0.04]	[0.04]	[0.03]	[0.03]
Diff. in Diff.	0.03		0.04		0.01		-0.01		0.03		0.04		0.03		0.05		0.02	
	[0.04]		[0.05]		[0.05]		[0.06]		[0.06]		[0.05]		[0.05]		[0.04]		[0.04]	
									By PI Gender									
Has Female PI	0.10	0.07	0.18	0.12	0.25	0.21	0.32	0.31	0.44	0.44	0.56	0.52	0.65	0.65	0.76	0.80	0.85	0.89
	(0.30)	(0.26)	(0.39)	(0.32)	(0.43)	(0.41)	(0.47)	(0.46)	(0.50)	(0.50)	(0.50)	(0.50)	(0.48)	(0.48)	(0.43)	(0.40)	(0.35)	(0.31)
Has Male PI	0.10	0.11	0.20	0.22	0.31	0.32	0.42	0.42	0.51	0.51	0.61	0.62	0.71	0.71	0.81	0.80	0.91	0.91
	(0.30)	(0.31)	(0.40)	(0.41)	(0.46)	(0.47)	(0.49)	(0.49)	(0.50)	(0.50)	(0.49)	(0.49)	(0.45)	(0.45)	(0.39)	(0.40)	(0.28)	(0.29)
Diff.	0.00	-0.03	-0.02	-0.10**	-0.07	-0.11**	-0.10*	-0.11**	-0.08	-0.08	-0.05	-0.10**	-0.06	-0.07	-0.05	0.00	-0.06*	-0.01
	[0.03]	[0.03]	[0.04]	[0.04]	[0.05]	[0.05]	[0.05]	[0.05]	[0.05]	[0.05]	[0.05]	[0.05]	[0.05]	[0.05]	[0.04]	[0.04]	[0.03]	[0.03]
Diff. in Diff.	-0.04		-0.08		-0.05		-0.01		0.00		-0.06		-0.01		0.04		0.04	
	[0.04]		[0.05]		[0.06]		[0.06]		[0.06]		[0.06]		[0.06]		[0.05]		[0.04]	

Notes. This table summarize simulated disparities in acceptance rates, across overall acceptance rates ranging from 10 through 90 percent. 10% overall acceptance rate means that 10% of submissions are accepted. Headers give the overall acceptance rate. The "Non-Blind" columns show the disparities in acceptance rates when acceptances are simulated using only Non-Blind scores. The "Blind" columns show for when acceptances are simulated using only Blind scores. "Diff" rows show the difference between the two preceding author traits (which are mutually exclusive), using a t-test comparison of means. "Diff. in Diff." rows show the effects of blinding: the difference between the two preceding author traits in the change in acceptance outcome due to blinding, using a t-test comparison of means. All acceptance outcomes are simulated by assuming that papers that scored in the top $X\%$ of Blind scores are accepted under blinding for an $X\%$ overall acceptance rate, and similarly for Non-Blind. Acceptances are simulated using a paper's reviewer-residualized average Blind and average Non-Blind score. Observations are at the paper-level.

Table A11: Selection into Robustness Subsamples

	Available Online						Missing Review					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	-0.01				-0.01	-0.02	-0.03				-0.03	-0.04
	(0.03)				(0.03)	(0.03)	(0.02)				(0.02)	(0.02)
Student		-0.06**			-0.06**	-0.06*		-0.04**			-0.04**	-0.04**
		(0.03)			(0.03)	(0.03)		(0.02)			(0.02)	(0.02)
Lower Rank Inst.			-0.02		-0.01	-0.02			-0.03		-0.02	-0.03
			(0.03)		(0.03)	(0.03)			(0.02)		(0.02)	(0.02)
Female PI				-0.08***	-0.07**	-0.07**				0.00	0.00	0.01
				(0.03)	(0.03)	(0.03)				(0.02)	(0.02)	(0.02)
Subfield FE						×						×
N	657	657	657	657	657	657	657	657	657	657	657	657
R^2	0.00	0.01	0.01	0.01	0.03	0.04	0.00	0.01	0.00	0.00	0.02	0.04

Notes. This table shows which papers were removed in the robustness checks in Section 3.2. The dependent variable for the first six columns are whether a paper was available online during the experimental period, and the dependent variable for the subsequent columns is whether a paper was missing a review (i.e. have three reviews instead of four). Observations are at the paper level. Heteroskedastic robust standard errors in parentheses.

C Citations and Publication Status Four Years Later

Table A12: Reviewer Scores and Paper Quality

	Panel A: Percentile in Citations						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Trait-Predicted Percentile	0.12***			0.06	0.08**		
	(0.04)			(0.04)	(0.04)		
Non-Blind Percentile		0.25***		0.24***		0.19***	0.18***
		(0.04)		(0.04)		(0.04)	(0.04)
Blind Percentile			0.24***		0.23***	0.17***	0.17***
			(0.04)		(0.04)	(0.04)	(0.04)
p-val H0: Reviewer = Predicted				0.00	0.01		
p-val H0: Blind = Non-Blind						0.80	0.82
Subfield FE							×
N	657	657	657	657	657	657	657
R^2	0.01	0.07	0.06	0.07	0.07	0.09	0.12
	Panel B: Percentile in Weighted Publication						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Trait-Predicted Percentile	0.08**			0.04	0.05		
	(0.04)			(0.04)	(0.04)		
Non-Blind Percentile		0.18***		0.17***		0.12***	0.11***
		(0.04)		(0.04)		(0.04)	(0.04)
Blind Percentile			0.20***		0.20***	0.16***	0.15***
			(0.04)		(0.04)	(0.04)	(0.04)
p-val H0: Reviewer = Predicted				0.02	0.01		
p-val H0: Blind = Non-Blind						0.52	0.55
Subfield FE							×
N	657	657	657	657	657	657	657
R^2	0.01	0.04	0.05	0.04	0.05	0.06	0.09

Notes. This table shows the predictive power of author traits, Blind score percentile ranks, and Non-Blind score percentile ranks for paper quality, measured by (a) citations and (b) journal-weighted publication status. Trait-predicted percentiles are created by the predicted values from regressing paper citations on author traits (applicant student status, institution rank, gender, PI gender). Percentile ranks take on a value between 0 and 100, and can have ties. Papers with no citations are coded as having zero citations before calculating rankings. Journal-weighted publication status uses the impact factor associated with the journal that a paper was published in, and takes value zero if the paper is unpublished. "p-val H0: Reviewer = Predicted" shows the p-value from testing the null hypothesis that the coefficient on the reviewer score percentile is equal to the coefficient on the trait-predicted percentile. "p-val H0: Blind = Non-Blind" shows the p-value from testing the null hypothesis that the coefficient on the Blind reviewer score percentile is equal to the coefficient on the Non-blind reviewer percentile. Observations are at the paper level. Heteroskedastic robust standard errors in parentheses.

Table A13: Within-Reviewer Score and Quality Rankings

	Percentile in Citations			Percentile in Weighted Publication Status		
	Non-Blind	Blind	All	Non-Blind	Blind	All
	(1)	(2)	(3)	(4)	(5)	(6)
Within-Reviewer Percentile	0.20***	0.21***	0.20***	0.16***	0.17***	0.16***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Within-Reviewer Percentile × Blind			0.01			0.01
			(0.04)			(0.04)
N	1289	1302	2591	1289	1302	2591
R^2	0.04	0.05	0.04	0.03	0.03	0.03

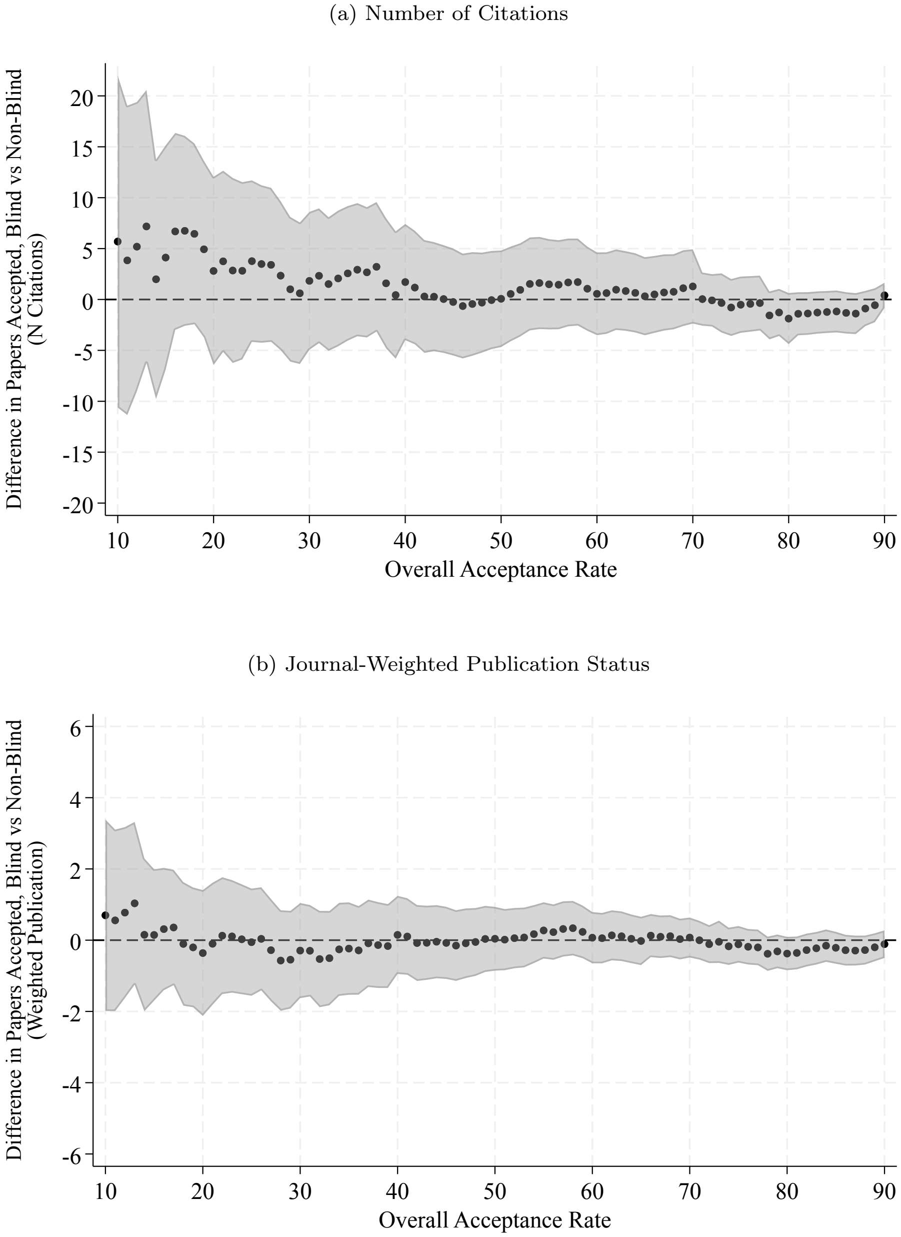
Notes. This table shows the predictive power of within-reviewer Blind score percentile ranks and within-reviewer Non-Blind score percentile ranks for within-reviewer paper quality rankings, measured by citations (columns 1-4) or journal-weighted publication status (columns 5-8). Percentile ranks take on a value between 0 and 100, and can have ties. Papers with no citations are coded as having zero citations before calculating rankings. Journal-weighted publication status uses the impact factor associated with the journal that a paper was published in, and takes value zero if the paper is unpublished. Observations are at the paper level. Heteroskedastic robust standard errors in parentheses.

Table A14: Score and Paper Quality Rankings (Robustness Subsample)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Trait-Predicted Percentile	0.17***			0.09*	0.11**		
	(0.05)			(0.05)	(0.05)		
Non-Blind Percentile		0.31***		0.28***		0.20***	0.20***
		(0.04)		(0.04)		(0.06)	(0.06)
Blind Percentile			0.30***		0.27***	0.15**	0.14**
			(0.04)		(0.04)	(0.06)	(0.06)
p-val H0: Reviewer = Predicted				0.01	0.02		
p-val H0: Blind = Non-Blind						0.67	0.59
Subfield FE							×
N	417	417	417	417	417	417	417
R^2	0.03	0.11	0.10	0.12	0.11	0.12	0.16

Notes. This table shows the predictive power of author traits, Blind score percentile ranks, and Non-Blind score percentile ranks for paper quality, measured by citations. Trait-predicted percentiles are created by the predicted values from regressing paper citations on author traits (applicant student status, institution rank, gender, PI gender). Percentile ranks take on a value between 0 and 100, and can have ties. Papers with no citations are coded as having zero citations before calculating rankings. Journal-weighted publication status uses the impact factor associated with the journal that a paper was published in, and takes value zero if the paper is unpublished. "p-val H0: Reviewer = Predicted" shows the p-value from testing the null hypothesis that the coefficient on the reviewer score percentile is equal to the coefficient on the trait-predicted percentile. "p-val H0: Blind = Non-Blind" shows the p-value from testing the null hypothesis that the coefficient on the Blind reviewer score percentile is equal to the coefficient on the Non-blind reviewer percentile. Observations are at the paper level. Heteroskedastic robust standard errors in parentheses.

Figure A6: Effects of Blinding on Quality



(a) Number of Citations

(b) Journal-Weighted Publication Status

Notes. These figures present simulations of the impacts of blinding on the average quality of papers that the conference selects, where paper quality is measured by (a) number of citations (b) journal-weighted publication statuses. Each dot in a figure reflects the coefficient from regressing paper quality on indicators for whether (1) the paper would only be accepted to the conference under blinding, (2) would be accepted under either regime, and (3) would be accepted under neither regime: the omitted category is (4) papers that would only be accepted under non-blind review. The coefficient on whether the paper would only be accepted to the conference under blinding (1) is then scaled by the fraction of papers that change acceptance status (category 1 and 4) among those who are ever accepted (1, 3, 4) to reflect the change in the average citations associated in accepted papers by blind status. Observations are at the paper level. All acceptance outcomes are simulated by assuming that papers that scored in the top $X\%$ of Blind scores are accepted under blinding for an $X\%$ overall acceptance rate, and similarly for Non-Blind. Acceptances are simulated using a paper's reviewer-residualized average Blind and average Non-Blind score. Shaded areas correspond to 95% confidence intervals, from heteroskedastic robust standard errors.

D Model Details

D.1 Reviewer's Decision Model

In this section, I explain the connection between the model of an individual reviewer's decision-making and the definition of average bias presented in the main text.

Following the notation of the main text, first, consider Non-Blind reviewers, who have the possibility of using information on author identities to engage in bias. Let $\tau_z(x_p, v_p)$ capture the magnitude bias that reviewer z inflicts onto papers of author traits x_p and submission content v_p. Reviewer z is *biased* in scores against subgroup $x \in \mathcal{X}$ (and in favor of subgroup $x' \in \mathcal{X}$) at $v \in \mathcal{V}$ if

$$\tau_z(x, v) \geq \tau_z(x', v)$$

On the other hand, Reviewer z is *unbiased* if $\tau_z(x, v) = \tau_z(v) \quad \forall v \in \mathcal{V}$. Under an unbiased reviewer, the reviewer-specific paper cost function is independent of author traits, so that among papers with the same submission content v, variation in Non-Blind scores that the reviewer gives is driven only by subgroup differences in the conditional expectations of paper quality. Note that the τ function depends also on v_p, not just x_p, allowing for bias to be heterogeneous in submission content. Then, paper p's score from reviewer z under Non-Blind review is given by:

$$S_{p,z}^{NB} = \mathbb{E}[Q_p | x_p, v_p] - \tau_z(x_p, v_p) \tag{A1}$$

In this way, consistent with the definition of bias in the main text, reviewer bias in scores drives deviations between latent paper quality and reviewer scores. Let $\tau(x, v) \equiv \mathbb{E}[\tau_z(x, v) | x, v]$, which represents behavior of the average reviewer. Note that even when the average reviewer is unbiased, this can be driven by various underlying distributions of reviewer-level bias.

Subgroup $x \in \mathcal{X}$ faces *bias on average* in scores, relative to subgroup $x' \in \mathcal{X}$, at $v \in \mathcal{V}$ if $\tau(x, v) > \tau(x', v)$.

The absence of the z subscript on the bias terms reiterate that this definition of bias describes the behavior of the average reviewer. A paper's expected Non-Blind score, averaged over the population of potential reviewers, is given by:

$$S_p^{NB} = \mathbb{E}[Q_p | x_p, v_p] - \tau(x_p, v_p) \tag{A2}$$

Now, consider Blind reviewers, who face the same decision process as Non-Blind ones, but evaluate papers without author information, x_p. Assume that for a given paper, Blind reviewers observe the same submission content, v_p, as Non-Blind reviewers.[31] Let $\tau(v) \equiv \mathbb{E}[\tau_z(x, v) | v]$, where the expectation is taken over both x and reviewers. In this case, $\tau(\cdot)$ is no longer a function of author traits since reviewers cannot access this information. By construction, Blind reviewers are therefore by construction unbiased. Paper p's

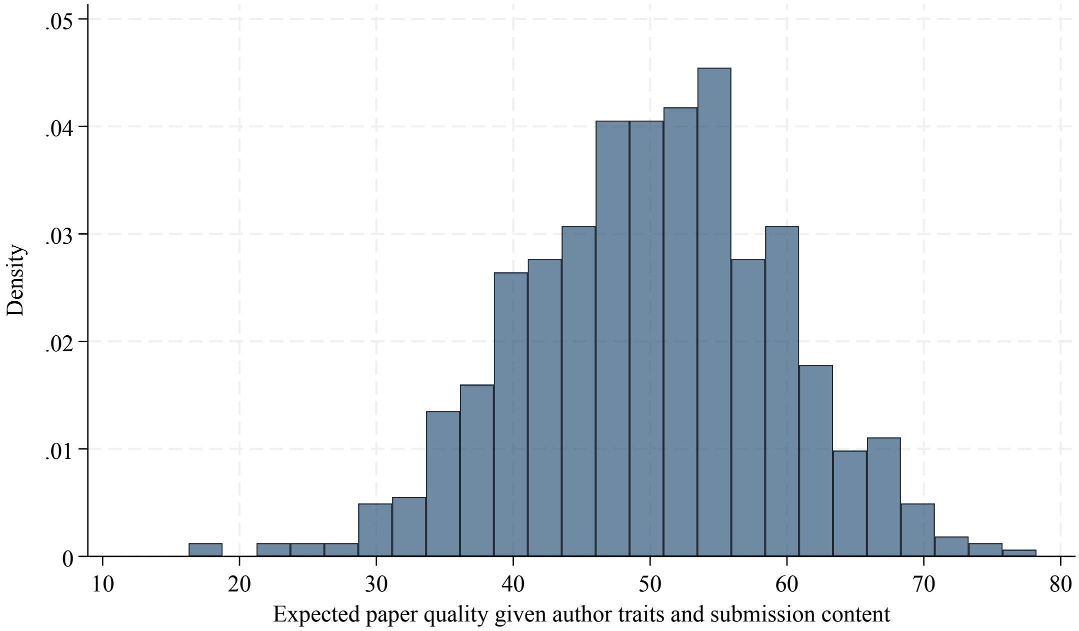
[31] These assumptions are violated if, for instance, blinding not only removes author identities from reviewers' information sets, but also causes reviewers to disengage with the review process and pay less attention to submission content. My previous results showing that Blind scores are as good of predictors of a paper's future quality as Non-Blind scores suggest that this is not the case.

expected Blind score is:

$$S_p^B = \mathbb{E}[Q_p|v_p] - \tau(v_p) \tag{A3}$$

D.2 Model Estimates

Figure A7: Distribution of estimated expectations of paper quality



Notes. This figure shows the distribution of estimated expectations of paper quality given author traits and submission content, $\mathbb{E}[Q|x, R(v)]$. This is estimated by the predicted values of the regression in column 3 of Table 6, but before scaling paper quality to match the possible range of reviewer scores. Paper quality (Q) is measured by a paper's percentile rank (taking on values between 0 and 100) in the number of citations four years after the experiment. Observations are at the paper level.

Table A15: Decomposing the Non-Blind Score Gap: Adjusting Citations

	Student	Inst Rank	Gender	PI Gender
Citations: Rank	-67	61	185	-19
	[-126,7]	[35,129]	[-1115,1706]	[-1769,435]

Notes. This table decomposes disparities in Non-Blind scores, considering the case when citations for traditionally lower-scoring groups (students, lower rank institutions, female applicants and PIs) are deflated by 10 percentage points relative to true paper quality, by inflating the measure of paper quality by 10 percentage points for the relevant demographic. The decomposition shows the contribution in Non-Blind score gaps that is attributable to accurate statistical discrimination, after controlling for submission content (see Equation 5). Accurate statistical discrimination is estimated by the difference in mean paper quality, conditional on submission content and other observables (the rest of the observed author traits and subfield). To reach this estimate, I take the estimates from Table 6, integrate the coefficient on the relevant demographic (e.g. coefficient on Student) and the product of its interaction with submission content (e.g. multiply the coefficient on Student × Submission Content with each paper's submission content) over the full distribution of submission content. Paper quality is measured by a paper's ranking in its number of citations that it has four years after the experiment, re-scaled to match the range of potential reviewer scores (as described in Section 5.3), but inflated by 10 percentage points for the relevant demographic. The displayed percentages divide this magnitude by the average gap in Non-Blind scores for the subgroup. 95% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the reviewer level.

E Generalizability

In this section, I follow the SANS (Selection-Attrition-Naturalness-Scaling) conditions (List, 2020) to discuss the generalizability of this experiment.

With regards to selection, all papers submitted to the conference were part of the experimental sample so that there was no selection conditional on candidate choice to apply. Similarly, all reviewers who were part of the review process were part of the experimental sample so that there was no selection conditional on choosing to be a reviewer. There is likely some selection into the applicant and reviewer pool, and one potential concern is that this selection is different between historical years and the experimental year because while neither applicants nor reviewers were told of the existence of an experiment, though both knew that the conference would use both Blind and Non-Blind review. Generally, a robust evaluation of whether applicants' choices to submit change by knowing that an evaluation is blind or not is out of the scope of this paper. I cannot identify the effect of announcing blind reviewers on submission content as the announcement is correlated with many unobservables (for instance, the location of the conference changes each year so that I cannot separately identify the effect of the location and the effect of announcing blind review). However, I am able to explore whether applicants and PIs seemed to positively select into the experimental sample by testing for heterogeneity by whether an applicant and PI are "new", meaning whether or not they had applied to the conference the two years before the experiment, or if they are a repeat candidate from previous years. For instance, it is possible that applicants most likely to benefit from blinding were more likely to apply in the experimental year, so that my blinding effects are driven by new applicants. I do not find significant evidence of this in the data (Table A9). While the point estimates are noisy, if anything, the impacts on blinding for students and applicants from lower ranked institutions is less positive for "new" applicants than repeat ones.

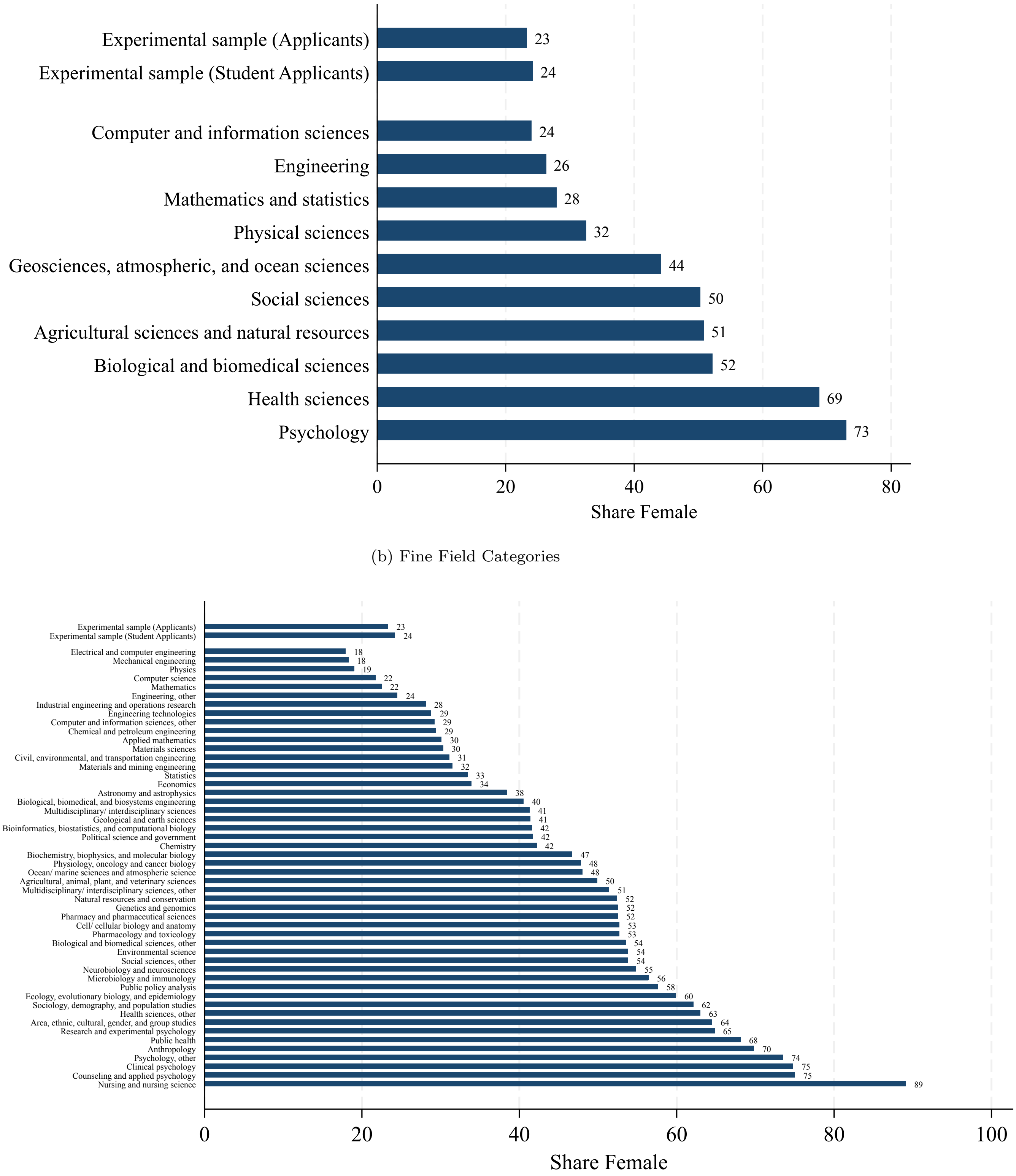
There is likely selection into the conference and into the field of study itself (computational neuroscience). Researchers choosing to enter computational neuroscience likely differ in unobservables from researchers choosing to enter other fields. This does not harm the internal validity of the study but potentially impacts the generalizability of my results when considering how effects may differ in other contexts and fields. For instance, the impacts of blinding may vary with the baseline diversity levels of the field. The direction of this heterogeneity remains unclear, depending on what aspects of the evaluation process and context are considered. For instance, Breda and Ly (2015) find, using differences between a students' scores on blind written tests and non-blind oral tests, that women in male-dominated fields on average face lower levels of gender bias than those in female-dominated fields. They attribute this result to stereotypes. However, Bagues et al. (2017) find that while changes in the number of women in evaluation committees for associate and full professorships in Italy and Spain does not significantly change gender differences in outcomes, male evaluators become more negative towards female candidates when a woman is added to the evaluation committee. In Figure A8 I compare the gender composition of the experimental sample with the gender diversity of various other fields. In the experimental sample, around a quarter of all applicants, as well as among the subsample of student applicants, were women. This is likely a more gender-imbalanced context relative to other fields. Using the share of doctorate recipients from 2021, the fields with the most similar gender shares are computer science and engineering, which are one of the most gender-imbalanced. Neurobiology and neurosciences has a much greater female share (55%), because this incorporates doctorate recipients in computational neuroscience as well as other sub-topics in neuroscience.

In terms of attrition, there was (by construction) 100% compliance among the sample of papers throughout the experiment, even in the collection of measures four years later given that absence of online paper presence itself was an outcome. Some reviewers did not not score all of the papers they were assigned, so that a small fraction of papers (6%) are missing one review. I discuss in the above Section 3.2 why this is likely not a concern that compromises the results.

Concerning naturalness of the choice task, setting, and framing, I use a natural field experiment (Harrison and List, 2004), such that the task at hand for both applicants and reviewers was not artificial. Previous years of the conference used only Non-Blind review, making blinding a potentially novel setting, but the (true) rationale that was told to the public for this was a reasonable one: the conference needed to adjust its usual logistics to accommodate Blind review and wanted an interim year to transition. The framing of this rationale was important, and both reviewers and applicants were not aware of the existence an experiment, minimizing experimenter demand effects (Levitt and List, 2007). Reviewers in the experiment were either always Blind or always Non-Blind, in order to minimize the salience of blinding as a novel setting. Moreover, with respect to setting, it was important that majority of the papers submitted to the conference were early works that were not available online. In other contexts where candidate identity can be discerned by Blind reviewers (e.g. where publicizing early works online is very common), Blind reviewers are likely not truly Blind, and the effects of blinding may differ from the current study. Similarly, contexts that vary in the amount of candidate information given may differ in blinding impacts as well: for instance, conferences that give reviewers the entire paper rather than a 300-word summary and 2 page description as in this study.

In terms of scaling, my results suggest that the efficacy of blinding as a policy depends on exactly who is at the margin of interest, and that blinding may not affect all individuals of a subgroup in a homogeneous manner so that policies aiming to alleviate disparities may have heterogeneous effects. This also illustrates a potential mechanism for why prior studies report seemingly mixed results. Ultimately, the goal of this paper is not to characterize disparities across all contexts but to offer a "WAVE1" insight, in the nomenclature of (List, 2020), to present a novel experimental design that can uncover the intricacies of a policy. Future work can extend this approach to other settings.

Figure A8: Female Representation Among Experimental Sample and 2021 Doctorate Recipients



(a) Broad Field Categories

(b) Fine Field Categories

Notes. This figure shows the share of female doctorate recipients across academic fields in 2021, and the share of females in the experimental sample. Doctorate recipient data and field categorizations are from the National Science Foundation (2021).