What Do Blind Evaluations Reveal?
How Discrimination Shapes Representation and Quality

Haruka Uchida*

October 21, 2025

Please click here for the most recent version.

Abstract

Concealing candidate identities during evaluations ("blinding") is often proposed to combat discrimination. I study how blinding affects the composition and quality of selected candidates, and the forms of discrimination driving effects. I implement a field experiment with an academic conference, running each submitted paper through blind and non-blind review, and collecting proxies of paper quality—citations and publication statuses five years later. I find that blinding significantly reduces gaps in reviewer scores and acceptances by student status and institution rank, with no significant changes by gender. Despite changes in representation, blinding does not lead to selecting lower-quality papers. To understand mechanisms, I elicit reviewer predictions of future submission outcomes in a second experiment and estimate a model of reviewer scores that decomposes non-blind disparities into distinct forms of discrimination: accurate and inaccurate statistical discrimination, considerations of alternative objectives (such as favoring authors who may benefit more from acceptance, or benefit others), and bias. I find that the nature of discrimination differs by trait: student score gaps are attributable to misperceptions of paper quality and alternative objectives, while institution rank gaps are consistent with bias.

*University of Chicago. uchida@uchiago.edu. I am grateful to John List, Michael Dinerstein, and Evan Rose for guidance and encouragement. I thank Tomas Hromadka, Srdjan Ostojic, Anne-Marie Oswald, and the committee members of the Cosyne conference. These parties are not responsible for the analyses and interpretations presented in this paper. I thank Jesse Backstrom, Alec Brandon, Stephane Bonhomme, Christina Brown, Leonardo Bursztyn, Katie Coffman, Michael Guy Cuna, Michael Greenstone, Justin Holz, Kilian Huber, Alex Imas, Juanna Joensen, Xianglong Kong, Clara Kyung, Ariel Listo, Claire Mackevicius, Magne Mogstad, Jack Mountjoy, Matthew Notowidigdo, David Novgorodsky, Devin Pope, Sally Sadoff, Heather Sarsons, Andrew Simon, Dan Svirsky, Min Sok Lee, Germán Villegas Bauer, Rebecca Wu, Karen Ye, 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 Conference, and Advances with Field Experiments at the University of Chicago for comments and discussion. Izzy Allum, Dushan Arsov, Faith Fatchen, Zack Fattore, Ashley Folts, Sonoe Fitzsimonds, Owen Humphries, John Kim, Yujin Lee, Zack Olson, Riley Osborn, Francesca Pagnotta, Victor Sengpiel, Steven Shi, and Kaixin Wang provided outstanding research assistance. This research is funded by the Becker Friedman Institute at the University of Chicago. 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

Disparities in evaluation outcomes are pervasive, arising in hiring decisions, loan approvals, and even criminal justice proceedings (e.g. Bertrand and Duflo, 2017; Lang and Spitzer, 2020). Academia is no exception, with consequential outcomes—including opportunities to present work, journal publications, tenure decisions—differing by gender, institution prestige, and career stage (e.g. Blank, 1991; Sarsons, 2017b; Doleac et al., 2021). Consequently, "blinding" procedures that conceal candidate identities have emerged as a popular policy intervention: if gaps arise because evaluators discriminate against particular traits, then removing this information should improve outcomes for traditionally disadvantaged groups.

Yet it remains unclear how blinding affects evaluator decisions, and its impact on representation and quality. The answers to these questions depend on whether reviewers use candidate information, and how. If evaluators use candidate identities to cater to biased preferences (Becker, 1957), then blinding can simultaneously increase representation and enhance the identification of high-quality candidates. In contrast, if candidate identities serve as informative signals of underlying quality (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977), blinding may obscure useful information and improve representation at the cost of quality. Evaluators may also act on inaccurate beliefs (Bordalo et al., 2019; Bohren et al., 2019; Coffman et al., 2021), or pursue alternative objectives, such as favoring candidates who would benefit most from acceptance or generate greater benefits to others.

Answering these questions is empirically challenging because (1) isolating effects on reviewer decisions requires holding all else constant, including evaluator and applicant traits, but the decision to adopt or enter a blind evaluation is typically endogenous. Moreover, identifying distinct forms of discrimination requires data on candidate outcomes: (2) underlying candidate quality for accurate statistical discrimination and (3) evaluator beliefs for inaccurate statistical discrimination and alternative objectives. Lastly, using group differences in these outcomes to identify channels of discrimination (4) requires holding constant all other features of the submission that the evaluator observes but the researcher may not:[1] otherwise, differences in outcomes may reflect differences in application content rather than discrimination.

In this paper, I conduct a field experiment at an international academic conference to study whether blinding affects the composition and quality of selected candidates, and the forms of dis-

[1]Examples include the quality of a cover letter during hiring, or the novelty of research described in a journal or conference submission.

crimination that arise in the absence of blinding. In prior years, reviewers for the conference scored submissions using noisy signals of underlying paper quality: author names, an abstract, and a two-page summary of the paper—but never the full paper. In the experiment, I randomly assign all 245 reviewers to either blind or non-blind review, and then randomly assign each of the 657 submissions to both two blind and two non-blind reviewers. Five years later, I collect proxies of paper quality—citations and publication statuses—for each submission. Finally, I run a second experiment with the same conference to directly elicit elicit reviewers' beliefs about submissions' future outcomes and again implement both blind and non-blind review.

My experimental design addresses the four main identification challenges. First, randomly assigning blind status to reviewers and assigning every submission to both blind and non-blind reviewers shuts down endogenous selection by reviewers and applicants into blinding.[2] This isolates the impact of blinding on evaluator decisions. Second, linking reviewer scores to proxies of paper quality allows for directly testing claims that blinding alters representation at the cost of selecting lower-quality submissions, and for quantifying the role of accurate statistical discrimination. Third, eliciting reviewer beliefs of submission outcomes in my second experiment allows me to separate accurate statistical discrimination from inaccurate reviewer beliefs, and considerations of alternative objectives. Finally, I combine the rich data from both experiments with a model of reviewer scores that interprets blind scores as proxies of submission content aside from author identity. I estimate the model to decompose disparities from non-blind evaluations into distinct forms of discrimination. My use of blind scores addresses comparability issues through an identification strategy that avoids several restrictive assumptions common in prior work. In this way, these novel data linkages allow for understanding not just whether blinding changes outcomes, but how evaluators use identity information when it is available.

I begin with testing how blinding changes the allocation of scores across observable author traits. Non-blind reviewers award significantly higher scores to applicants who are more senior, from top 20 ranked institutions, and male, relative to their counterparts. Blinding significantly reduces scores for traditionally high-scoring groups. Most stark is by student status: while non-blind reviewers score non-student papers nearly 0.25 standard deviations higher than student papers, this gap shrinks by over 60 percent among blind reviewers. This implies that the non-blind score disparity between these groups is not fully explained by differences in submission content. Differences by gender are

[2]For instance, Boring et al. (2025) document differential selection into blind evaluations.

noisier and I fail to reject that gender score gaps are unaffected by blinding. 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 and find similar conclusions.

I then examine whether impacts on scores translate to changes in acceptances. The relationship is not mechanical, because changes in scores could be driven by infra-marginal candidates who are not on the margin of acceptance. The conference, both in past years and in the experimental year, determines acceptances by choosing an overall acceptance rate given venue capacity constraints, and selecting the papers with the highest scores.[3] Comparing papers that would be accepted if only blind review were used with those that would be accepted if only non-blind review were used, I find that blind review significantly crowds in students. In fact, using only blind review would nearly eliminate the student acceptance gap. Simulating acceptance outcomes for other acceptance thresholds shows that this effect persisted across overall acceptance rates, implying that blinding benefits students at essentially every margin. Blinding also reduces the applicant institution rank acceptance gap, but this change is only statistically significant at higher acceptance thresholds because changes in scores are concentrated among papers higher in the distribution relative to the 60% cutoff.

Given that blinding changes the composition of accepted authors, I next test whether blinding affects the quality of selected submissions, directly examining claims that blinding helps representation at the expense of reviewers' abilities to identify high-quality submissions. I collect each paper's number of citations and publication status, including the journal of publication, up to five years after the experiment. Nearly 70% of papers have at least one citation and around half are published. Reviewer scores are generally informative of paper quality: both blind and non-blind score rankings positively correlate with paper quality rankings, although the correlations are far from one. Moreover, blind scores are not significantly worse predictors of paper quality than non-blind scores. Consequently, papers that would be admitted to the conference under blind review are not significantly lower-quality than those that would be admitted under non-blind review: I rule out differences larger than 5 citations on average. These conclusions hold when using other measures of quality, such as a binary for having at least one citation, publication outcomes, or citations 2 years later instead of 5. This suggests that even entities that prioritize selecting the most qualified

[3]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.

candidates over equalizing representation may consider adopting blinding.[4]

The result that blinding affects composition without significantly worsening quality suggests that accurate statistical discrimination alone cannot explain non-blind disparities. To quantify and compare the role of each mechanism, I formalize a model of reviewer scores and use the additional model structure to decompose disparities under non-blind review into distinct forms of reviewer discrimination. Let submissions be characterized by author traits, and submission content aside from author traits (e.g. the research idea explained in the submission). Non-blind scores—unlike blind ones—may depend on author identities due to (1) accurate statistical discrimination based on underlying subgroup differences in paper quality (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977), (2) discrimination based on inaccurate beliefs of paper quality (Bordalo et al., 2019; Bohren et al., 2019; Coffman et al., 2021), (3) discrimination based on alternative objectives beyond paper quality, such as favoring authors who would benefit most from acceptance, or those who would benefit the conference the most, and (4) all other sources of disparities including animus (Becker, 1957). In contrast, blind reviewers cannot use author identities, so blind scores depend only on submission content aside from author names.

Identifying the model requires identifying reviewer beliefs and their link to scoring behavior. I partner with the same conference to run a second experiment that directly elicits reviewer beliefs during the review process, and again implement blind and non-blind review. In addition to providing the usual reviewer score, reviewers are asked to predict paper quality outcomes (citation count and publication status 5 years later) and beliefs over three potential alternative objectives (talk quality, how much the author benefits from acceptance, how much the conference benefits from acceptance). I find that reviewers who receive author names predict that student submissions will have significantly fewer citations in 5 years, be a significantly less engaging talk, and benefit significantly less from acceptance than submissions by non-students.

I bring the rich data generated by my two experiments to the model, and decompose non-blind disparities in the main experiment. My approach leverages blind scores as a proxy for submission content that is unobserved to the researcher. I address potential biases arising from reviewer-specific

[4]To address concerns that my measures of paper quality are endogenous to acceptance to the conference, I repeat the analyses (1) subsetting to papers whose conference acceptance statuses would have remained the same regardless of blind regime, (2) deflating accepted papers' citations using my estimates for the impact of acceptance on citations from a regression discontinuity that leverages the conference's threshold rule, and (3) residualizing out acceptance status, author observables, and their interactions, from citations. Each one leads to the conclusion that blinding does not perform significantly worse in accepting high-quality papers relative to non-blind (see Section 3.3 for details).

idiosyncrasies in reviewer scores[5] by leveraging the fact that each paper was randomly assigned to two blind reviewers: for a given paper, one blind reviewer's score can be instrumented with its other blind reviewer's. To supplement my use of blind scores as proxies of submission content, I apply modern natural language processing methods to the text of each submission to generate additional controls for submission content: using a neural network-based sentence embedding model designed for scientific papers (Singh et al., 2022), I find that accounting for these observables increases explanatory power but does not alter the decomposition conclusions, further corroborating the value of using blind scores.

Model estimates imply that the form of discrimination driving disparities varies by characteristic. The majority of the difference in non-blind scores by student status can be attributed to disparities in reviewers' expectations of alternative objectives, particularly talk quality, and some to reviewer misbeliefs about underlying paper quality. In contrast, the institution rank score gap cannot be explained by paper quality beliefs nor alternative objectives: a large portion of the gap remains unexplained, consistent with the presence of institutional bias. These results help rationalize why blinding differentially affects demographic groups without significantly affecting admit quality: removing reviewers' abilities to accurately statistically discriminate can worsen quality, but reducing the ability for reviewing to act on inaccurate beliefs, engage in bias, or pursue alternative objectives can improve quality. In my context, they offset.

Taken together, this paper contributes to past work on blinding[6] (e.g. Blank, 1991) which has primarily focused on documenting its effects on disparities in evaluation outcomes but not necessarily discrimination. I contribute to these works by (1) collecting proxies of quality to test whether a policy often aimed to equalize representation comes at the cost of quality, (2) eliciting reviewer beliefs to disentangle inaccurate beliefs, alternative objectives, and other drivers including animus, (3) formulating and estimating a model of reviewer scores to decompose disparities into distinct forms of discrimination. Most similar to this paper is Pleskac et al. (2024), who study blinding at an academic conference and track publication outcomes two years later but do not incorporate reviewer beliefs nor estimate a model of reviewers to quantify mechanisms. More broadly, I build

[5] Gillen et al. (2019) illustrate the importance of addressing these forms of measurement error.

[6] Past studies have examined the effects of blinding job applications (Åslund and Skans, 2012; Krause et al., 2012a,b; Behaghel et al., 2015), orchestra auditions (Goldin and Rouse, 2000), high school tests (Lavy, 2008; Hinnerich et al., 2011; Breda and Ly, 2015; Terrier, 2020), academic conferences (Tomkins et al., 2017; Madden and DeWitt, 2006; Ferber and Teiman, 1980) and journals (Blank, 1991; Budden et al., 2008; Roberts and Verhoef, 2016; Huber et al., 2022). Grogger and Ridgeway (2006) apply a similar logic to compare racial differences in traffic stops during daylight, when police are more likely to observe driver race, with differences during darkness.

on the large literature on discrimination (e.g. Bertrand and Mullainathan, 2004; Edelman et al., 2017; Sarsons, 2017a; Kline and Walters, 2021; Kline et al., 2022, 2024) which generally identifies the presence of discrimination but does not distinguish its forms.[7]

Additionally, my within-paper design identifies the impact of blinding on reviewer behavior, while holding all else including submission content, fixed. This builds on the blinding literature which has largely focused on across-candidate comparisons (e.g. Blank, 1991; Krause et al., 2012b). Tomkins et al. (2017) and Huber et al. (2022) also collect blind and non-blind evaluation outcomes for a given paper, but the former allows for endogenous sorting by reviewers to papers and the latter compares blind and non-blind outcomes for a single paper.

Methodologically, this paper contributes to the large literature on disentangling forms of discrimination (List, 2004), particularly work using "outcomes tests" (Becker, 1957) which aims to separate biased decision-making from accurate statistical discrimination by comparing post-evaluation outcomes across subgroups.[8] Closest to my setting are Laband and Piette (1994), Smart and Waldfogel (1996), Card et al. (2020) and Carrell et al. (2024), who test for biases in the journal review process using subgroup differences in papers' citation counts conditional on referee and editor decisions, as well as Li (2017) who similarly tests for bias in peer review for grant funding. A central challenge with outcome tests is ensuring comparability, because of "infra-marginality bias" wherein average differences are not informative of marginal differences (Heckman, 1998; Ayres, 2002), and the presence of unobserved submission characteristics requires strong assumptions on the evaluator decision model (Canay et al., 2024). Moreover, these approaches do not distinguish traditional notions of animus from other forms of discrimination, such as inaccurate statistical discrimination and consideration of alternative objectives, which hold distinct policy implications. I tackle these challenges by (1) incorporating data from blind evaluations to proxy for submission content and side-step some of the usual assumptions in the literature, and (2) eliciting reviewer beliefs directly to unpack "bias" and disentangle misbeliefs, alternative objectives, and other drivers such as animus. I build on a framework that can be applied to investigate discrimination in contexts beyond

[7]Related literatures have tested the effects of varying specific types of candidate information such as removing criminal backgrounds (Agan and Starr, 2018; Doleac and Hansen, 2020; Rose, 2021), or adding productivity signals including test performance (Autor and Scarborough, 2008; Wozniak, 2015) and ratings of past quality (Pallais, 2014; Cui et al., 2020; Chan, 2022). In a similar spirit, Aneja et al. (2025) study the effects of a policy that made it easier for consumers to search for Black-owned businesses.

[8]Other approaches include work that examines the impacts of policies and movements that alter the permissible extent of bias on productivity, such as Huber et al. (2021) who find that biased expulsions of Jewish managers during the rise of Nazi Germany worsened firm outcomes.

academia (Canay et al., 2024).[9] Taken together with the fact that blinding has been implemented in numerous settings, my methodology can be leveraged to disentangle forms of discrimination in various other settings.

Finally, I build on past work on evaluator beliefs (Bordalo et al., 2019; Bohren et al., 2019; Coffman et al., 2021) by directly eliciting reviewer beliefs and linking them to blind and non-blind scores. This approach enables me to disentangle inaccurate statistical discrimination from other hypothesized "omitted payoffs" (Kleinberg et al., 2018) that reviewers may consider, and to separate both from additional sources of disparities such as animus. Moreover, I show how combining these belief and scoring data can identify the weights that reviewers place to different objectives, thereby quantifying the relative importance of each in shaping evaluation outcomes.

This paper proceeds as follows. Section 2 explains the setting and experimental design. Section 3 examines the effects of blinding on evaluation outcomes (reviewer scores and acceptances) and on quality. Section 4 presents the model of reviewer scores that decomposes disparities in non-blind evaluation outcomes into distinct forms of discrimination, and Section 5 its results. Section 6 considers additional mechanisms, and Section 7 concludes.

2 Experimental Design

2.1 Main experiment

The experiment was conducted from October through early December of 2019, during the review process for the the 2020 Computational and Systems Neuroscience ("Cosyne"), an annual international academic computational neuroscience conference.[10] It is one of the landmark conferences of its field, and draws a little over 1,000 attendees each year. Conference submissions are generally early works. Applicants submit information on the paper that they would present if accepted. Specifically, applicants send a title, a 300-word description, a two-page detailed explanation which may include figures and tables, a list of relevant subfields, and a list of all authors (see Appendix A for images of the review portal and an example submission). Importantly, reviewers are not given the entirety of the paper associated with the submission.

[9] Outcome tests have been implemented in various settings, including bail bond dealers (Ayres and Waldfogel, 1994), bail decisions (Arnold et al., 2018; Kleinberg et al., 2018; Arnold et al., 2022), motor vehicle searches (Knowles et al., 2001; Anwar and Fang, 2006; Persico and Todd, 2006), capital sentencing Alesina and La Ferrara (2014), mortgage lending (Ferguson and Peters, 1995), consumer lending (Dobbie et al., 2021), and social insurance receipt (Low and Pistaferri, 2025).

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

Each year, reviewers are on average assigned 10 papers, and score their assigned papers on an integer scale of 1 through 10, inclusive. Reviewers give one score per paper. Reviewers are 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 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 types of authors.[11]

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

Before the submissions were due in the experimental year, the conference committee (truthfully) told the public that the conference would implement both blind and non-blind review, 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). Applicants were instructed to anonymize their submission documents such that aside from the author list, materials 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.

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: 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. Reviewer decisions had real stakes: all reviewer scores, both blind and non-blind, were used to determine acceptance outcomes, and reviewers were told this upfront.

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

Second, every application was randomly assigned to four reviewers from its relevant subfields: two "non-blind" reviewers and two "blind" ones (Table A1 shows the distribution of subfield).[12] In the non-blind group, consistent with past years, reviewers were directly presented the author names with the paper information. All reviewers received 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—traits such as affiliated institution were left to be inferred from names.

As stated in my pre-analysis plan, 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). Non-blind reviewers are informed of the applicant's identity, so that reviewers know which coauthor would present at the conference if accepted. Author information was collected through self-reported forms and later verified.[13] Specifically, I collected 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 the 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 race and the number of historical cumulative citations that each applicant and PI is associated with at the time of submission.[14] Finally, five years after the experiment, I collect outcomes for each submission: whether it is available online, is published, journal of publication, and the number of citations it is associated with. Appendix B elaborates on the data collection process.

[12] In prior years, each paper was assigned to three non-blind reviewers of its relevant subfields.

[13] 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.

[14] I generally find that the main results presented do not change whether I control for citation counts or not (Appendix C.5).

2.2 Summary Statistics

The main experimental sample consisted of 657 paper submissions and 245 reviewers, translating to 2591 unique paper-reviewer observations.[15] Table 1 summarizes author traits at the paper level. Note that there is not a balance table because each paper was scored both blind and non-blind, such that balance is achieved by construction. Around half of applicants are students. Among applicants who are affiliated with an institution rank (553 out of 657 papers), the median rank was 21 (see Appendix C.1). Going forward, I bin institution rank by whether it is top 20 or not ("lower rank").

Table 1: Submission Traits

(a) Applicant Traits

	Mean	SD
Applicant Traits		
Student: Yes (%)	51.3	50.0
Student: No (%)	48.7	50.0
Has Institution Rank (%)	84.0	36.7
Institution Rank | Have Rank	72.9	165.8
Institution Rank: Is Top 20 (%)	40.6	49.2
Institution Rank: Is 20+ (%)	43.4	49.6
Institution Rank: Not University (%)	14.9	35.7
Institution Unknown (%)	1.1	10.3
Gender: Female (%)	23.4	42.4
Gender: Male (%)	76.6	42.4
Observations (Papers)	657	

(b) PI and Coauthor Traits

	Mean	SD
PI Traits		
Gender: Female (%)	16.7	37.4
Gender: Male (%)	83.1	37.5
Gender: Unknown (%)	0.2	3.9
Other		
Number of Authors	3.9	2.2
Solo Author (%)	2.9	16.8
All student authors (%)	0.5	6.7
No student authors (%)	31.1	46.3
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 main 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.

Applicants are majority male (77%) and PIs even more (83%). Applicant gender is not significantly correlated with student status nor institution rank (Table A2). Applicant and PI genders are not significantly correlated. Papers on average have 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 the number of coauthors.

As expected given random assignment, blind reviewers do not systematically differ in observed traits from non-blind reviewers (Table 2).

[15]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 C.9.

Table 2: Reviewer Balance Table

	Treatment			
	Non-Blind	**Blind**	Difference	p-value
Gender: Female	0.36 (0.48)	0.33 (0.47)	-0.03	0.65
Gender: Male	0.64 (0.48)	0.67 (0.47)	0.03	0.65
Years Since PhD	6.51 (4.79)	6.61 (4.19)	0.10	0.88
Student	0.02 (0.13)	0.03 (0.18)	0.01	0.45
Inst: Top 20	0.39 (0.49)	0.35 (0.48)	-0.05	0.46
Inst: 20+	0.44 (0.50)	0.42 (0.50)	-0.02	0.80
Inst: Not Uni	0.14 (0.35)	0.21 (0.41)	0.07	0.15
Inst: Rank Unknown	0.03 (0.16)	0.02 (0.13)	-0.01	0.61
Has Reviewed Pre-2020	0.92 (0.28)	0.92 (0.27)	0.00	0.89
N Reviewed Pre-2020	1.72 (1.21)	1.92 (1.61)	0.20	0.39
N	119	126		

Notes. This table shows summary statistics for blind and non-blind reviewers for the 2020 conference. 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" is self-reported and corresponds to whether the reviewer has reviewed for the 2019 (one year before the experiment) or 2018 (two years before the experiment) conference, and the number of times if so. Standard deviations in parentheses.

2.3 Second Experiment to Directly Elicit Reviewer Beliefs

To investigate the role of reviewer beliefs and their link to score disparities, I conduct a second experiment with the same conference from the review process during October to December of 2024, 5 years after the main experiment.[16] The conference switched entirely to blind review since the 2020 experiment. Given this, the main changes for the second experiment relative to the new (blind) status quo were to (1) directly elicit reviewer beliefs about submissions' future citations and publication statuses during the review process and (2) add a non-blind review component where the committee requested reviewers to complete two additional reviews after the blind component was finished. All questions were asked on the reviewer portal that the conference uses each year.

During non-blind review, I elicit reviewer beliefs of three alternative objectives that reviewers may consider when assigning scores: how engaging the talk would be if the submission were selected, the extent to which the author would benefit from presenting a talk at the conference, the extent to which the conference would benefit from having the authors attend. Appendix E provides details.

3 Effects of Blinding on Evaluation Disparities and Quality

I begin with presenting the results from the main experiment. I first document disparities in evaluation outcomes that arise in the absence of blinding, by focusing on reviewers who were randomly

[16]This experiment was pre-registered as an amendment to the main experiment, AEARCTR-0005139.

assigned to be non-blind. I then examine whether these non-blind disparities reflect underlying differences in submission content or reviewer use of author information, by using the two levels of random assignment in my experiment that isolate the impacts of blinding on reviewer decisions: Section 3.1 shows for evaluation scores, and Section 3.2 for acceptances. Section 3.3 then tests whether blinding leads to accepting lower-quality submissions.

3.1 Effects of Blinding on Reviewer Scores

What disparities would arise if the conference ran "business as usual", without blinding? Figure 1 reports scores by each trait. On average, when reviewers receive author identities, student applicants score 0.49 points ($\approx$ 0.25 SD) worse than their senior counterparts. Applicants from top 20 ranked institutions (better than the median ranking) score 0.77 points ($\approx$ 0.40 SD) higher than those from lower ranked institutions. Female applicants and PIs receive lower scores than their male peers. The conditional score gaps—when I consider each trait while controlling for the rest—exhibit similar patterns (Table A3).

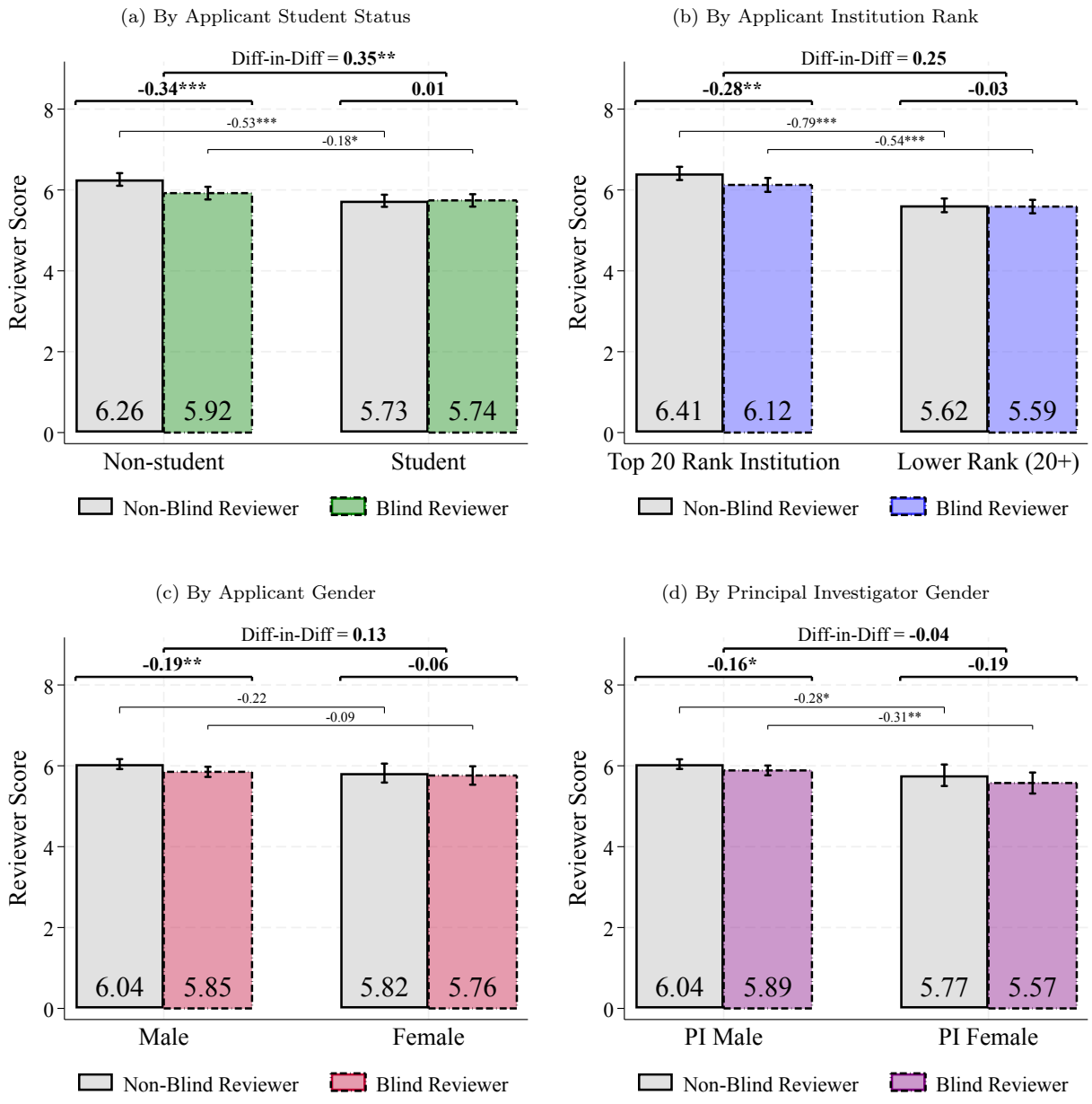
The notion that some subgroups score worse than others under non-blind review alone does not reveal differential treatment, since it could be driven by differences in submission content. For this, I compare disparities between blind and non-blind scores for the same set of papers. Recall that the conference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

Blind reviewers on average give lower scores overall than non-blind ones, which is consistent with past work (Blank, 1991), but blinding does not significantly change variation in reviewer scores ($p = 0.652$ for a variance ratio test) nor the disagreement in scores between two reviewers for a given paper (Figure A4).[17]

Turning to disparate impacts, I find that blinding significantly reduces scores for traditionally better-scoring applicants: applicants who are senior (non-students) and from top 20 institutions (Figure 1). The estimates for gender are noisy, and in terms of magnitude, female PIs scores decrease by a greater magnitude than male PIs. To consider conditional disparities, I estimate, for

[17]The correlation between the scores given by a paper's two same-treatment reviewers are roughly similar: non-blind correlation of 0.26, blind 0.23 (see Figure A3 for correlation across). Interestingly these within-paper score correlations are similar in magnitude to those in Blank (1991).

Figure 1: Reviewer Scores and Effects of Blinding



Notes. This figure shows the mean scores associated with each trait and blind status. * $p < .1$, ** $p < .05$, *** $p < .01$ P-values adjusted for multiple hypothesis testing (List et al., 2019; Steinmayr, 2020) for the difference-in-differences across each of the characteristics, estimated using equation 1, are 0.07, 0.18, 0.19, 0.77, respectively. Bar ticks correspond to the 95% confidence intervals, with standard errors clustered at the reviewer level.

paper p assigned to reviewer r:

$$S_{p,r} = \delta_0 + \delta X_p * \mathbb{1}\{Blind\}_r + \xi_r + \gamma_p + e_{p,r} \tag{1}$$

where $S_{p,r}$ 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, ξ_r are reviewer fixed effects, γ_p are paper fixed effects, and $e_{p,r}$ is an error term. δ captures the average change in a score gap due to blinding. While the paper and reviewer fixed effects are not necessary 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.[18] Standard errors are clustered at the reviewer-level to reflect the level of randomization.

Table A4 summarizes, and the results are consistent with the unconditional means in Figure 1. Blinding significantly reduces the student score gap by over 60%, and under blind review, student applicants are statistically indistinguishable from non-student applicants. This is not driven by a small number of applicants: the distribution of score changes (difference between a paper's average blind and non-blind score) for students first-order stochastically dominates the distribution for non-students ($p = 0.012$, Figure A6),[19] revealing that students are generally disadvantaged when reviewers know identities.

Blinding reduces the institution rank score gap, with the difference between applicants affiliated with a top 20 ranked institution and those affiliated with a lower ranked institution decreasing by around 25%, although this effect is marginally significant. Moreover, the institution rank score gap persists under blind review, suggesting that a portion of the non-blind disparity can be explained by subgroup differences in submission content.

Interestingly, the notion that those from worse ranked institutions benefit from blinding is largely driven by the fact that applicants from near-top (rank 6 through 20) institutions are worse off (Figure A7), although these differences across finer categories of institution rank are not statistically significant. This finding qualitatively matches those from Blank (1991), who finds that blinding journal referees most affects (negatively) authors from near-top institutions, rather than those at the highest or lowest ranks.

Impacts of blinding on gender score gaps are imprecisely estimated. Point estimates imply that

[18]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.

[19]I implement Barrett and Donald (2003)'s consistent test of first-order stochastic dominance, with the null hypothesis that the distribution of score differences for students does not stochastically dominate the distribution of score differences for non-students. I implement the bootstrap approach by Barrett and Donald (2003) with 1000 draws, using Schaub and Schaub (2024).

blinding reduces the applicant gender score gap by around 50%, while increasing the PI gender score gap, but neither are statistically significant.[20] I find suggestive evidence of interaction effects, with blinding benefitting student and female applicants who have a female PI more than those with a male PI, but these are not statistically significant (Table A5).

The results presented here are likely not driven by subgroup differences in coauthorships beyond applicants and PI traits. The effects reported above persist even after controlling for the number of coauthors who are students, from a lower ranked institution, or female (Appendix C.4). I find similar conclusions with the inclusion of controls for the number of historical citations associated with each applicant and PI (Appendix C.5), or applicant and PI race (Appendix C.6).

Despite possible reviewer heterogeneity (Welch, 2014), I do not find significant evidence of heterogeneity by reviewer gender, reviewer institution rank, nor by reviewer experience as proxied by whether the reviewer was involved with the conference in the prior two years (Appendix C.8). Point estimates suggest that female reviewers are more favorable towards female applicants and PIs than male reviewers, and that blinding reverses this gap, but these patterns are not statistically significant. The directional change in the applicant gender score gap is driven by reviewers affiliated with a top 20 ranked institution rather than those from worse ranked institutions.

A common concern is that blind evaluations are not truly blind. For instance, reviewers may find papers online 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 know author identities and act as they would have if they had received the author list (i.e., been "non-blind"), then this would underestimate my estimates of blinding effects. On the other hand, if blind reviewers are aware of author identities and behave more favorably towards groups that they otherwise would not have, then my estimates would be over-inflated. With these concerns in mind, I repeat the main analyses with the 83% of submissions (551 out of 657) that were not available online during the time of review, and find similar conclusions as above (see Table A16). Moreover, an optional question in the reviewer portal asked blind reviewers

[20] Taking point estimates literally suggests that female applicants are better off by blinding, but female PIs are worse off, although 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. Past work suggests that discrimination may change with reputation and seniority (e.g. Bohren et al., 2019; Petersen and Saporta, 2004).

to guess the author of their submissions: guesses are wrong the majority of the time, although the evidence is suggestive because the choice to respond to the question is endogenous. Blinding effects persist even after dropping the blind reviews where the reviewer correctly guessed the author (see Appendix C.9).

3.2 Effects of Blinding on Conference Acceptance

I next test whether blinding effects on scores translate to changes in acceptances. The relationship is not mechanical, because changes in scores could be driven by infra-marginal candidates who are not on the margin of acceptance. The overall effect depends on who is at the margin, the distribution of scores (Figure A5), and the effects of blinding along the score distribution.

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.[21] In the experimental year, the conference accepted around 60% of submissions[22] (402 out of 657 papers), using both a paper's non-blind and blind scores.[23] I first test for disparities in realized acceptances by estimating, for paper p:

$$Y_p = \omega_0 + \omega X_p + \epsilon_p \tag{2}$$

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 3 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.

To test whether the use of blind review affected acceptance outcomes, I predict which papers would be accepted if the conference used only non-blind or only blind scores to determine acceptances. I then estimate Equation 2 using these predicted acceptance statuses as the outcome variable. Column 2 of Table 3 summarizes disparities in acceptances that depend only on non-blind scores, under the realized acceptance rate. I find that student applicants would be 12 percentage

[21] Appendix C.14 corroborates this relationship between reviewer scores and acceptances.

[22] Neither applicants nor reviewers in this experiment were aware of the overall acceptance rate, 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).

[23] The conference chose to place more weight on blind scores to calculate a paper's overall average score that determines acceptance decisions, so a paper's score was constructed by 2/3 of the paper's average blind score and 1/3 of its non-blind score.

Table 3: Effects of Blinding on Acceptances

		Counterfactual		
	Actual	Non-Blind	Blind	Diff
	(1)	(2)	(3)	(4)
Student	-0.08**	-0.12***	-0.01	0.11**
	(0.04)	(0.04)	(0.04)	(0.05)
Lower Rank Inst.	-0.20***	-0.18***	-0.17***	0.02
	(0.04)	(0.04)	(0.04)	(0.05)
Female	-0.01	-0.09*	-0.03	0.06
	(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)
Subfield FE	×	×	×	×
N	657	657	657	657
R^2	0.07	0.08	0.05	0.03

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 second column shows what disparities would have been if the conference only used non-blind scores to determine acceptances, the third column for if only blind scores were used, and fourth column the effects of blinding on these acceptance gaps. Dependent variable is (1) whether the paper was admitted to the conference, (2) whether a paper scored in the top 61% of non-blind scores, (3) whether a paper scored in the top 61% of blind scores, (4) difference between (3) and (2). Observations are at the paper-level. Heteroskedastic robust standard errors in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

points less likely to be admitted than non-students who are admitted 70% of the time. Applicants affiliated with lower-ranked institutions would be 20 percentage points less likely to be admitted than applicants from top 20 ranked institutions who are admitted 71% of the time, and female applicants are 9 percentage points less likely relative to the 62% for male applicants.

I find that blinding meaningfully alters conference acceptances. If only blind scores are used to determine acceptances, 30% of papers that would have been accepted under non-blind review would be crowded out: estimating a deconvolved distribution of reviewer scores reveals that around 25 percentage points of the 30 of this can be attributed to paper-reviewer specific idiosyncracies, and the rest to changes in the scores that papers receive in expectation over the population of potential reviewers (see Appendix C.11). These changes are not substituting across similar authors. In fact, using only blind scores to determine acceptances would eliminate the student acceptance rate gap (column 3 of Table 3). Simulating acceptance outcomes for various overall acceptance rates reveals that students benefit from blinding across the distribution of acceptance thresholds (Figure A9).[24]

For institution rank, blinding does not significantly change the acceptance gap at the overall

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

acceptance rate used by the conference. This is because the effects of blinding on the institution rank score gap are concentrated among applicants higher in the score distribution (Figure A9b): blinding significantly reduces the institution rank acceptance rate gap under more selective overall acceptance rates. For instance, under a 30% overall acceptance rate, non-blind review leads to a 17 percentage point institution rank acceptance rate gap, and blind review reduces this gap by 10 percentage points (59%).

Blinding directionally reduces the gender acceptance rate gap, and slightly widens the PI gender acceptance rate gap, but these changes are not statistically significant.

Putting these effect sizes in the context of past work, Blank (1991) finds that among non-blind reviewers for the American Economic Review, 23% of submissions by top 20 ranked institutions are accepted, and those by authors from worse-ranked institutions are accepted 11.5 percentage points less often. Blinding reduces this unconditional acceptance rate gap by 2 percentage points,[25] and the gap conditional on gender by 1.42 percentage point (see Appendix C.12 for further comparisons).[26]

3.3 Effects of Blinding on Quality

Given that blinding changes the allocation of scores and acceptances 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, five years after the experiment. I follow past work in interpreting these measures as proxies for underlying paper quality (e.g. Laband and Piette, 1994; Smart and Waldfogel, 1996; Li, 2017; Card et al., 2020; Carrell et al., 2024). 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 are used to determine consequential outcomes such as career promotions. 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 5.3 and introduce data on other papers aspects that reviewers may consider in the second experiment. Table 4 shows the raw means. Nearly 80% of the papers were available online. 71% of papers had at least one citation, and 55%

[25] This number is calculated using the sample sizes and acceptance rates in Table 5 of Blank (1991): the unconditional non-blind acceptance gap is given by $(28.2*149+20.9*274)/(149+274)-(15*239+7.1*426)/(239+426) = 13.5$ ppt, and the blind by $(29.5*149+13.4*274)/(149+274)-(10.1*239+6.2*426)/(239+426) = 11.5$ ppt.

[26] This number is calculated using the sample sizes and acceptance rates in Table 6 of Blank (1991): the change in acceptance rate for those affiliated with top 20 ranked institutions is given by $(1.8*149+(-7)*274)/(149+274) = -3.9$ ppt and for those affiliated with lower ranked institutions by $((-5.3)*239+(-0.9)*426)/(239+426) = -2.5$ ppt.

were published.

Table 4: Paper Citations and Publication Statuses Five Years Later

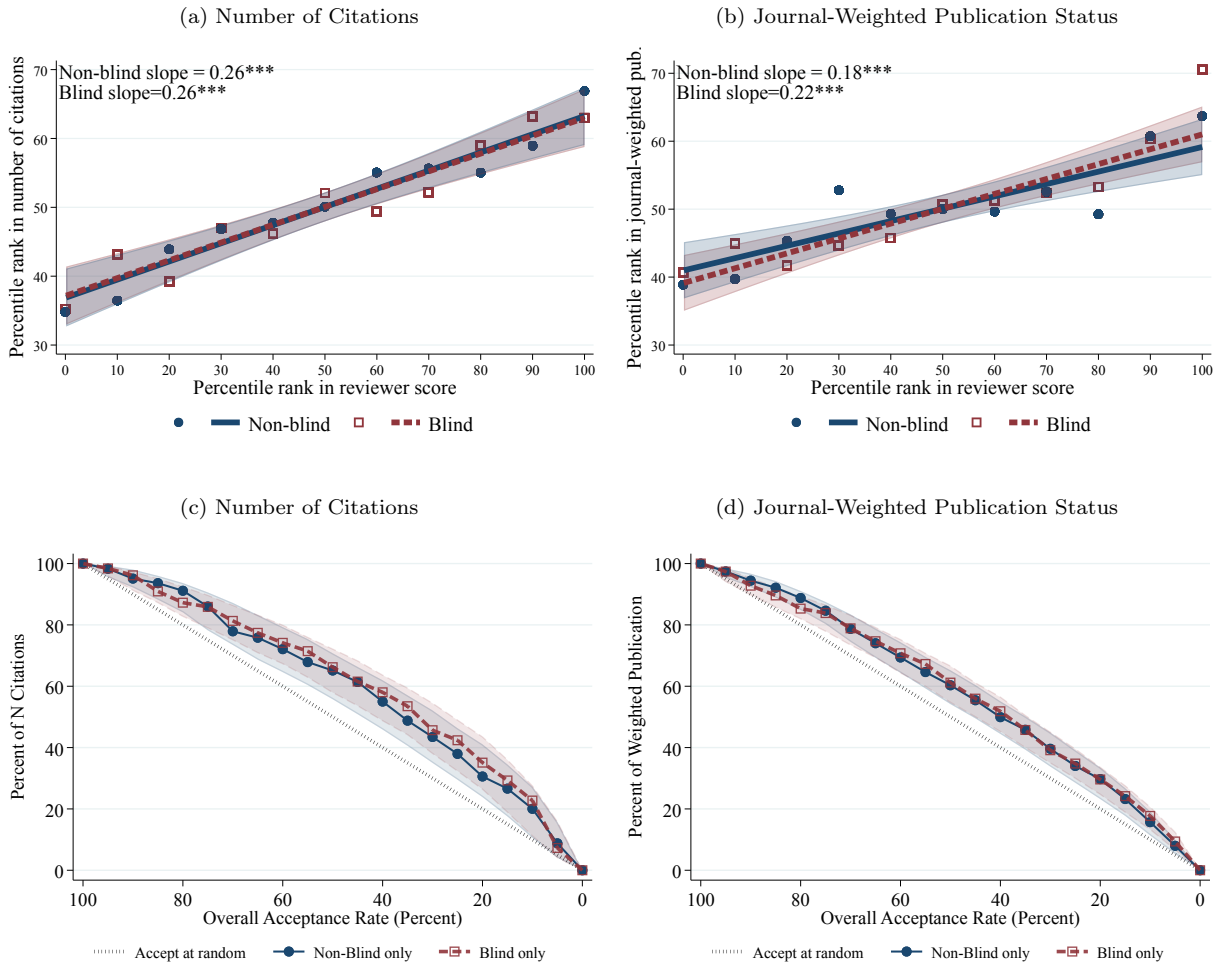
	Citations			Publication Status		
	Has	N	N | Has	Online	Pub	Weighted
All	0.71	25.57	36.05	0.79	0.55	5.37
	(0.45)	(53.79)	(60.86)	(0.41)	(0.50)	(6.93)
			By Applicant Student Status			
Student	0.73	22.99	31.36	0.81	0.58	5.21
	(0.44)	(42.12)	(46.47)	(0.39)	(0.49)	(6.47)
Not Student	0.68	28.29	41.34	0.77	0.53	5.54
	(0.47)	(63.77)	(73.55)	(0.42)	(0.50)	(7.40)
Difference	0.05	-5.31	-9.98*	0.04	0.05	-0.33
	[0.04]	[4.20]	[5.64]	[0.03]	[0.04]	[0.54]
			By Applicant Institution Rank			
Lower Ranked	0.67	19.65	29.48	0.75	0.53	4.42
	(0.47)	(40.96)	(47.22)	(0.44)	(0.50)	(5.53)
Top 20	0.76	32.66	42.75	0.84	0.58	6.10
	(0.43)	(67.07)	(73.90)	(0.37)	(0.49)	(7.28)
Difference	-0.10**	-13.01***	-13.27**	-0.09***	-0.05	-1.68***
	[0.04]	[4.70]	[6.30]	[0.03]	[0.04]	[0.55]
			By Applicant Gender			
Female	0.66	16.70	25.22	0.76	0.51	4.18
	(0.47)	(25.93)	(28.31)	(0.43)	(0.50)	(5.45)
Male	0.72	28.29	39.09	0.80	0.56	5.74
	(0.45)	(59.54)	(66.92)	(0.40)	(0.50)	(7.29)
Difference	-0.06	-11.59**	-13.87**	-0.04	-0.05	-1.55**
	[0.04]	[4.94]	[6.79]	[0.04]	[0.05]	[0.64]
			By PI Gender			
Female	0.69	22.25	32.21	0.79	0.56	5.91
	(0.46)	(38.81)	(43.17)	(0.41)	(0.50)	(6.93)
Male	0.71	26.27	36.88	0.79	0.55	5.27
	(0.45)	(56.38)	(63.82)	(0.41)	(0.50)	(6.94)
Difference	-0.02	-4.02	-4.67	0.00	0.01	0.64
	[0.05]	[5.63]	[7.64]	[0.04]	[0.05]	[0.73]

Notes. This table shows summary statistics for the citation and publication statuses for each paper five 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. "Weighted" refers to journal-weighted publication status, which is the impact factor associated with the journal that a paper was published in, and takes 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. P-values for t-test comparisons of means are represented by stars: * $p < .1$, ** $p < .05$, *** $p < .01$

I first regress a paper's percentile rank in quality measures on its percentile rank in blind and non-blind scores (Figure 2a and 2b), and find that both blind and non-blind scores are significant predictors of paper quality. Predicting citations using author observables (applicant student status, institution rank, gender, PI gender) leads to a significantly worse prediction of quality than reviewer scores (Table A20). Moreover, I fail to detect any significant difference in the predictive power of blind scores relative to non-blind scores. The point estimates of the slopes on blind and non-blind

scores are nearly identical for predicting citations (Figure 2a), and the point estimate is if anything steeper for journal-weighted publication status (Figure 2b), suggesting that blind scores capture quality at least as well as non-blind scores.

Figure 2: Effects of Blinding on Quality



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 dashed line with hollow squares represents the shares when acceptance outcomes are determined by blind scores only, and the blue solid line with solid circles for when outcomes are determined by non-blind scores only. Paper quality measures are collected 5 years after the experiment. 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. Shaded areas correspond to 95% confidence intervals, based on 1000 clustered bootstrap replications that are drawn at the paper level.

I next examine how the predictive power of reviewer scores translate to the quality of admitted submissions. To do so, I inspect the quality of submissions that would be accepted under blind and non-blind review. I additionally benchmark the predictive power of reviewer scores against accepting

papers at random, where accepting $X\%$ of papers corresponds to admitting, in expectation, $X\%$ of the cumulative number of citations associated with the submission pool. Figure 2c and 2d show the fraction of citations and publication statuses that are associated with admitted submissions, relative to the total number of citations associated with the applicant pool (Figure A12 shows by trait). The lines corresponding to blind and non-blind review are significantly more outwards relative to the at-random benchmark, which implies that reviewer scores perform significantly better at admitting highly-cited papers than random acceptance at each margin of acceptance, consistent with the result that reviewer scores are informative of paper quality.

Turning to the comparison between admissions based on blind versus non-blind scores, I find that the two lines are essentially overlapping, and if anything, slightly more outwards in Figure 2c, suggesting that blind scores perform as well as non-blind scores in admitting high-quality submissions. These conclusions persist with the inclusion of subfield fixed effects, and when considering a paper's within-reviewer ranking instead of its within-sample one (Table A21). Comparing the citation counts of papers that would be accepted under blind review and those that would be under non-blind review, generally across overall acceptance rates, I reject that blinding worsened quality by more than 5 citations (≈ 0.1 standard deviations) (Figure A11).

These results suggest that blinding does not pose a significant tradeoff between representation and quality: using only blind review admits just as high-quality papers as using only non-blind review. Using other measures of paper quality—whether a submission becomes a paper that is available online, has at least one citation, is in the top decile of submissions in the number of citations, is published—produces similar conclusions (Figure A13), as well as when considering citations 2 or 4 years after the experiment instead of 5 (Figure A14), implying that the result is not sensitive to the measure of paper quality. If the citations process is biased against traditionally low-scoring groups that benefitted from blinding, then this is likely an under-estimate of the effect of blinding on true paper quality, meaning that blinding may even improve it.

One concern may be that the patterns above are driven by the fact that conference acceptance directly affected citation and publication outcomes. I find that is likely not the case through four analyses. First, using the fact that the conference's acceptance rule creates a discontinuity in reviewer scores that determines acceptance, I conduct a regression discontinuity analysis and do not find significant evidence that acceptance meaningfully affects citation and publication outcomes 5 years later (Figure A16). Second, I repeat the main analyses while deflating citation counts for

accepted papers by the upper 95% confidence interval of my regression discontinuity estimate for the impact of acceptance and find similar conclusions (Section C.17). Third, I residualize out acceptance status, author traits, and their interactions, from citations (Figure A18b). Finally, 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 (Table A22). In each of these robustness checks, I find the conclusion that blind review predicts quality at least as well as non-blind.

4 Framework for 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, implying that accurate statistical discrimination—wherein reviewers use author names solely as informative signals of quality—alone cannot explain the blinding effects.

I interpret the results from the lens of a model of reviewer scores, and use the additional model structure to decompose baseline (non-blind) disparities into four 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, (2) inaccurate statistical discrimination (Bordalo et al., 2019; Bohren et al., 2019; Coffman et al., 2021), wherein reviewers discriminate using beliefs that deviate from realized outcomes, (3) pursuit of alternative objectives such as favoring applicants who would benefit the most from being accepted to the conference, and (4) other determinants of disparities, including animus (Becker, 1957).

I build on the framework of Canay et al. (2024) who present a decision model to interpret the literature on "outcome tests," which compare subgroup differences in post-evaluation outcomes to identify evaluator bias. In the context of academia, Laband and Piette (1994), Smart and Waldfogel (1996), Card et al. (2020), and Carrell et al. (2024) test for bias by referees and editors by examining disparities in papers' citation counts conditional on referee and editor decisions. While these works focus on gathering insights from non-blind evaluations and subsequent outcomes, I add in the notion of blind evaluation outcomes, and show how they can proxy for model components that are often unidentified in past approaches. I explain my model and identification approach in this section, and

the estimation and results in Section 5.

4.1 A Model of Non-Blind Reviewer Scores

Let submissions be characterized by two terms: x_p, a vector of author traits that are inferred by the reviewer through the author list (e.g. applicant's student status, gender), and v_p, submission content aside from author identity that is observed by the reviewer at the time of review but not necessarily observed by the researcher. Submission content can be multi-dimensional, including whether idea described in the submission is novel, how clearly it is written, and the number of figures or tables provided. Define $\mathcal{X}$ as the set of all possible x_p and $\mathcal{V}$ analogously for v_p.

Non-blind reviewers observe x_p and v_p, and use these two pieces of information to form posterior beliefs of paper quality (Q_p) and potentially alternative objectives beyond paper quality (A_p), such as how much an applicant benefits from acceptance, as well as to cater to biased preferences. Following this notation, non-blind reviewer r's score for paper p can be attributed to:

$$\tilde{S}_r^{NB}(x_p, v_p) = \delta_{0,r} + \delta_{1,r} \underbrace{\mathcal{E}_r[Q_p|x_p, v_p]}_{\text{Belief of paper quality}} + \delta_{2,r} \underbrace{\mathcal{E}_r[A_p|x_p, v_p]}_{\text{Belief of alternative objectives}} - \underbrace{\tau_r(x_p, v_p)}_{\text{Residual}} \tag{3}$$

where $\mathcal{E}_r[\cdot|x, v]$ represents reviewer r's posterior beliefs upon viewing author traits x and submission v. The δ coefficients are constants that scale beliefs to scores, and their r subscripts reflect the notion that some reviewers may place a greater weight on perceived quality relative to other components. $\tau_r(x_p, v_p)$ captures the residual variation in reviewer scores, and includes animus (Becker, 1957), as well as reviewer-specific idiosyncratic tastes.

Reviewers' posterior beliefs of paper quality, $\mathcal{E}_r[Q_p|x_p, v_p]$, can differ from the objective expectation of realized quality, $\mathbb{E}[Q_p|x_p, v_p]$. Let $\lambda_r(x_p, v_p)$ capture reviewer r's deviation in belief for paper p:[27]

$$\lambda_r(x_p, v_p) \equiv \underbrace{\mathbb{E}[Q_p|x_p, v_p]}_{\text{Expectation of realized paper quality}} - \underbrace{\mathcal{E}_r[Q_p|x_p, v_p]}_{\text{Belief of paper quality}} \tag{4}$$

Taking these together, consider the thought experiment: two submissions with exactly the same content V are scored by the same reviewer r. One of the submissions is by authors of trait x, and the

[27] One may also decompose consideration of alternative objectives into realized outcomes (e.g. accurate statistical discrimination based on talk quality) and misbeliefs (e.g. deviations between perceived and realized talk quality) by collecting data on these outcomes, although the usual challenge is that they are only observable among accepted applicants.

other by authors of trait x'. Differences in scores between the two submissions are attributable to four distinct channels: differences in realized paper quality, misbeliefs of paper quality, alternative objectives, and a residual term that includes animus.

$$\underbrace{S_r^{NB}(x,V) - S_r^{NB}(x',V)}_{\text{Gap in non-blind score}} = \underbrace{\delta_{1,r}\left(\mathbb{E}[Q|x,V] - \mathbb{E}[Q|x',V]\right)}_{\text{Accurate statistical discrimination}} - \underbrace{\delta_{1,r}\left(\lambda_r(x,V) - \lambda_r(x',V)\right)}_{\text{Inaccurate statistical discrimination}}$$
$$+ \underbrace{\delta_{2,r}\left(\mathcal{E}_r[A|x,V] - \mathcal{E}_r[A|x',V]\right)}_{\text{Alternative objectives}} - \underbrace{\left(\tau_r(x,V) - \tau_r(x',V)\right)}_{\text{Residual (including animus)}} \tag{5}$$

The last component, $\tau_r(x,V) - \tau_r(x',V)$, represents determinants of disparities that are not explained by accurate and inaccurate statistical discrimination based on paper quality, and considerations of alternative objectives. This term therefore includes animus that authors of trait x receive from reviewer r relative to authors of traits x', conditional on submission content V. Other drivers may contribute to this residual component, such as alternative objectives that are not captured by the ones I elicit (see Kleinberg et al. (2018) for a discussion of these "omitted payoffs") and specification errors (Canay et al., 2024).

The objective for the remainder of this paper will be to identify and estimate the decomposition in equation 5. Equation 5 gives a decomposition for a single value of submission content: the main results will focus on integrating over the population distribution of submission content, and averaging over reviewers. I refer to additional results in Section 5.2 that condition on specific values.

4.2 Model Identification

There are three main challenges associated with identifying the decomposition described above. First, identifying accurate statistical discrimination requires observing measures of paper quality. Second, identifying reviewer misbeliefs and consideration of alternative objectives, as well as the δ weights that reviewers place on beliefs, requires observing reviewer beliefs and their link to non-blind scores. Third, each component requires conditioning on submission content, which is typically unobserved to the researcher. Without conditioning, subgroup differences in each outcome (paper quality, reviewer scores, and beliefs) may be entirely explained by differences in submission content.

To address the first challenge, I collect data on paper quality measures for each submission as described in Section 3.3. My approach does not require directly observing underlying paper quality, Q_p, but instead unbiased (in a statistical sense) measures of it. For the second challenge, I use the

data from my second experiment that directly elicited reviewer beliefs (see Section 2.3). To tackle the third challenge, I turn to data on blind scores. Generally speaking, my approach conditions on blind scores in order to calculate subgroup differences among papers with comparable submission content. As I describe below, I allow for observed reviewer scores to contain idiosyncratic noise, by leveraging the fact that my experiments collected more than one blind score for each submission.

In order to introduce my assumptions, I first establish notation on the mapping between submission content and blind scores. In contrast to non-blind reviewers, blind reviewers do not receive author information, x_p. Instead, blind reviewers assign scores using only v_p, the submission content aside from author traits. As a result, without any loss of generality, reviewer r's blind score for paper p can be expressed as:

$$\tilde{S}^B_{p,r} = S^B(v_p) + u^B_{p,r} \tag{6}$$

where $S^B(v_p)$ captures the blind score that a submission with content v_p receives in expectation over the population of potential reviewers, so that $u^B_{p,r}$ is mean zero by construction. Moreover, random assignment of reviewers to papers implies that reviewer deviations are mean zero conditional on author traits and submission content: $\mathbb{E}[u^B_{p,r}|x_p, v_p] = 0$. Non-blind scores can be written in an analogous way, with $S^{NB}(x_p, v_p)$ capturing the non-blind score that paper p receives in expectation over the population of potential reviewers.

4.2.1 Identification Assumptions

I first assume that a submission's expected blind score across reviewers sufficiently captures its submission content for realized and perceived quality, and alternative objectives:

Assumption 1. *Conditional on expected blind scores $S^B(v_p)$ and author traits x_p, v_p is independent of realized and perceived quality, perceptions of alternative objectives, and expected non-blind scores.*

This is violated if non-blind reviewer beliefs respond to aspects of submission content that are not captured by the expected blind score. Relatedly, this is also violated if blinding simply causes reviewers to disengage: I show in Section 6 that this is likely not the case.

Later in Section 5.2, I relax this assumption by applying modern natural language processing approaches to generate additional, tractable controls from the observed submission text data: this then allows submission content v_p to additionally enter the quality and alternative objective ex-

pectation functions through a richer set of observables. In general, I find that incorporating these text-based controls adds explanatory power but does not change the decomposition conclusions.

Second, assume that reviewer-specific deviations in blind scores are uncorrelated across reviewers:[28]

Assumption 2. $\mathbb{E}[u^B_{p,r} u^B_{p,r'}] = 0 \quad \forall p, r \neq r'$.

This is in fact implied by random assignment of reviewers to papers. Assumptions 1 and 2 allow me to use observed blind scores as proxies of submission content, as I describe in detail below. Finally, I assume that conditional on reviewer observables, reviewer-specific scaling coefficients are uncorrelated with posterior beliefs and the residual determinants of disparities:

Assumption 3. *Conditional on reviewer observables, δ_r on beliefs are uncorrelated with beliefs themselves* $(\mathcal{E}_r[Q_p|x_p, v_p], \mathcal{E}_r[A_p|x_p, v_p], \tau_r(x_p, v_p))$

This holds when δ_r or posterior beliefs are homogeneous across reviewers, but homogeneity is not necessary. This assumption is violated if, for example, conditional on observables, reviewers who are more biased against female applicants are also more responsive to changes in quality beliefs. An implication of Assumption 3 is that the behavior of the average reviewer within a reviewer observable trait can be identified using the average scaling coefficient and average beliefs: for instance, the average role of misbeliefs can be rewritten as $\mathbb{E}_r[\delta_{1,r}\lambda_r(x, v)] = \mathbb{E}_r[\delta_{1,r}]\,\mathbb{E}_r[\lambda_r(x, v)]$ where the expectation is taken over reviewers. Reviewer observables can be used to weaken this assumption, by estimating reviewer weights and beliefs within a reviewer observable cell (for instance, estimating separately across reviewer institution-by-gender).

I now turn to how each of these assumptions are used. Identifying the decomposition (eq. 5) consists of three main steps: identifying (1) the average non-blind score gap conditional on submission content (the left hand side of the decomposition), (2) the conditional expectations in realized quality, misbelief, and alternative objectives, integrated over the distribution of submission content, for each trait of interest (e.g. $\int_{V \in \mathcal{V}} \mathbb{E}[Q|x, V]\, dF(V)$), (3) the δ coefficients of the non-blind scoring equation (eq. 3) which relate beliefs to scores. I detail each step below.

[28] While I have omitted notation of subfields in the assumptions for simplicity, assignment of reviewers to papers was conditional on subfields so all analyses in the decomposition will be conditional on subfields.

4.2.2 Identifying Group Differences Conditional on Submission Content

The first two steps consist of calculating subgroup differences in outcomes conditional on submission content. They follow largely the same steps, besides for changing the outcome variable of interest. Let Y_p denote some outcome for paper p, such as non-blind scores, realized paper quality, or reviewer beliefs. First, Assumption 1 implies that expectations conditional on submission content v is equivalent to conditioning on the corresponding expected blind score across reviewers $S^B(v)$: $\mathbb{E}[Y|x, v] = \mathbb{E}[Y|x, S^B(v)]$. Then, consider a linear approximation of the conditional expectation function that depends on a paper's expected blind score, author traits, and their interaction:

$$\mathbb{E}[Y_p|x_p, S^B(v_p)] = \beta_0 + \beta_1 S^B(v_p) + \beta_2 x_p + \beta_3 S^B(v_p) x_p \tag{7}$$

where β_2 and β_3 reflect the extent of subgroup differences. While I proceed with a linear approximation in the main analyses for simplicity, this approach can be extended for more flexible functional forms. I explain this in greater detail in Appendix F.1.

Simply estimating equation 7 by regressing outcomes (such as measures of paper quality) on observed blind scores leads to attenuation bias because observed blind scores contain idiosyncratic paper-reviewer noise relative to the expected blind score, $S^B(v_p)$, as captured by $u^B_{p,r}$ in equation 6. Drawing from the longstanding approach of using instrumental variables to account for errors-in-variables (Reiersøl, 1941), I rely on Assumption 2 and the fact that I collected 2 blind scores for each submission. Then, a consistent estimate of the coefficient can be recovered by instrumenting for a paper's blind score from one reviewer with another blind reviewer's (in other words, for paper p, instrument $\tilde{S}_{p,r}$ with $\tilde{S}_{p,r'}$ for $r \neq r'$).[29]

With the β coefficients in equation 7 identified, the conditional non-blind score gap, integrated over the distribution of submission content, is identified using data on blind scores and author traits:

$$\int_{V\in\mathcal{V}} (\mathbb{E}[Y|x, V] - \mathbb{E}[Y|x', V]) dF(V) = \int_{V\in\mathcal{V}} \Big(\mathbb{E}[Y|x, S^B(V)] - \mathbb{E}[Y|x', S^B(V)] \Big) dF(V) \tag{8}$$

$$= \beta_2(x - x') + \beta_3(x - x') \int_{V\in\mathcal{V}} S^B(V)\, dF(V) \tag{9}$$

$$= \beta_2(x - x') + \beta_3(x - x')\, \mathbb{E}[\tilde{S}^B_{p,r}] \tag{10}$$

[29] I test for heterogeneity in blind scores by reviewer gender and institution affiliation in Table A12 and Table A13, respectively.

where the first equality uses Assumption 1, the second uses equation 7, and the last uses the scoring equation (equation 6) and the fact that reviewer deviations $u^B_{p,r}$ are mean-zero.

Setting Y to non-blind scores identifies conditional non-blind score gaps, the left-hand side of the decomposition. Replacing Y with measures of realized quality identifies the paper quality expectation function, $\mathbb{E}[Q|x, v]$, which corresponds to accurate statistical discrimination.[30] Replacing Y with reviewer beliefs of paper quality and alternative objectives identifies their conditional belief functions. The difference between expected quality and reviewers' perceived quality identifies the extent of misbeliefs across reviewers, $\lambda(x, v)$. The main additional assumption required for identifying misbeliefs is that the mapping from blind scores and author traits to reviewer predictions remains constant across the two experimental years (see Appendix E for a comparison).

4.2.3 Identifying Scoring Coefficients

Finally, the scoring coefficients (δ) in the non-blind score equation (3) are identified by data on elicited reviewer beliefs over submission outcomes, non-blind scores, and blind scores. Directly eliciting reviewer beliefs allows me to avoid making assumptions on which reviewer beliefs are most connected to reviewer scores, and use the data to estimate equation 3 directly. Note that under Assumption 3, I do not need to recover the full distribution of reviewer-specific coefficients, but instead the average coefficient. Subtracting out conditional subgroup differences in realized quality, misbeliefs, and alternative outcomes from the conditional score gap identifies the residual component.

4.3 Identification Relative to Past Work

Identification of the decomposition relies on the rich data generated by my two experiments: the first experiment which links 2 blind scores, non-blind scores, and measures of paper quality for each submission, and the second which collects non-blind scores, blind scores, and reviewer beliefs of paper quality and alternative objectives. My approach differs from past work in three main ways. First, I use blind scores as a proxy for submission content, v_p, a component that the outcomes test literature generally takes as unobserved to the researcher. This allows for testing for discrimination among applicants with comparable submission content. Using the data on blind scores avoids im-

[30] Note here that I do not rely on the assumption that I directly observe a paper's latent quality, Q_p, but instead that I observe an unbiased measure of it ($\tilde{Q}_p$), so that the predicted values from regressing $\tilde{Q}_p$ on $(x_p, S^B(v_p))$ recovers an unbiased estimate for $\mathbb{E}[Q_p|x_p, S^B(v_p)]$.

posing additional assumptions on the distribution of unobserved submission content, and allows for unobserved submission content to enter more flexibly into evaluator decisions (Canay et al., 2024). Moreover, this avoids "infra-marginality bias", which arises from the fact that simply comparing average subgroup differences need not identify discrimination because average differences are generally uninformative of marginal ones (Ayres, 2002): I can identify marginal candidates directly from the data, because I observe reviewer scores for each paper and the conference's decision rule.[31]

Second, I collect measures 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, or on-the-job productivity among hired workers—but this outcome is by construction unobserved for defendants judges chose not to release or applicants who were rejected.[32] Data on paper quality for all submissions allows for identifying conditional expectations of underlying paper quality without relying on matching or extrapolation models.

Third, I link data on blind and non-blind scores to data on reviewer beliefs. This provides two main advantages. One is that while past work using outcomes tests have focused on distinguishing accurate statistical discrimination from all other determinants of disparities, this combines traditional notions of bias with other distinct sources, such as inaccurate beliefs by the decision-maker or consideration of alternative objectives. My approach separates each these channels. Second, data on reviewer scores and beliefs identifies the mapping between a reviewer's beliefs over various submission outcomes and the score the reviewer ultimately assigns. Past studies have tested whether disparities are consistent with accurate statistical discrimination using various outcome measures one at a time (e.g. Kleinberg et al., 2018). In contrast, my approach avoids relying on a single metric and jointly uses the realized quality and belief measures.

[31] 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) take 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. Simoiu et al. (2017) present a threshold test that allows for distributions of unobservables to differ across subgroups, which imposes a distributional assumption on unobservables.

[32] 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).

5 Disentangling Sources of Discrimination: Estimation and Results

5.1 Estimation

As in the prior sections, I take author traits, x_p, as the vector of applicant student status, institution rank, gender, PI gender, and subfields. The assumption going forward is that these dimensions of author traits are the ones that author names convey to reviewers.[33]

For paper quality, I use a paper's citation count and indicators for the paper being available online and for being published, all observed 5 years after the experiment. In Section 5.3, I discuss the implications of using these measures which themselves may contain bias in their production process and show robustness to alternative measures.

To conduct the decomposition in equation 5, I impose functional forms on both the conditional paper quality expectation function and beliefs functions. This allows me to integrate disparities in scores, paper quality, and reviewer beliefs over the marginal distribution of submission content. For simplicity, in the subsequent results section, I show estimates from estimating expectation functions that depend solely on author traits and blind scores and omit their interactions. In Appendix F.2, I show that including interactions produces similar conclusions.

When estimating the conditional expectation functions associated with each component of the decomposition, I account for noise in blind reviewer scores, as discussed in Section 4.2. I instrument for a paper's blind score from one reviewer with its other blind score, using a stacked regression that uses a given paper-(blind) reviewer observation once as the endogenous variable and once as the instrument.[34]

Next, I estimate the scaling coefficients (δ) in the non-blind score equation (3) using data from the second experiment that directly elicited reviewer beliefs. I regress reviewer r's non-blind score for paper p on r's beliefs of p's paper quality (number of citations, publication status) and alternative

[33]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 (Section C.5).

[34]This is the "Obviously Related Instrumental Variables" method that builds on the traditional errors-in-variables instrumental variables approach, and produces consistent estimates of the true relationship between the latent regressor (in this case, a paper's expected blind score) and outcome Gillen et al. (2019).

objectives (talk quality, extent conference benefits, extent author benefits):

$$S_{p,r}^{NB} = \hat{\delta_0} + \sum_k \hat{\delta_1^k}\, \mathcal{E}Q_{p,r}^k + \sum_a \hat{\delta_2^a}\, \mathcal{E}Q_{p,r}^a + \xi_r + \eta_{p,r} \tag{11}$$

where k indexes over the paper quality measures and a over the alternative objectives, $\mathcal{E}Q_p^k$ is paper p's reviewer belief for quality measure k, and $\eta_{p,r}$ is an error term. Intuitively, I estimate an index that summarizes reviewer beliefs of paper quality, and another index that summarizes beliefs over alternative objectives. ξ_r are reviewer fixed effects which account for potential reviewer-specific differences in optimism.

Finally, to aggregate over the multiple paper quality measures, I multiply the estimated coefficients to the estimated subgroup differences associated with that outcome. For instance, to quantify the extent of accurate statistical discrimination proxied by citation count, I estimate equation 7 with citations on the left hand side, and aggregate over student versus non-student differences. I then multiply this value to the δ_1 coefficient on reviewer beliefs over citations. Note that the δ_1 coefficient on perceived citations will be the same coefficient used to scale both realized citations and misbeliefs over citations in the decomposition (equation 5).

5.2 Decomposition Results

Before showing the aggregated model estimates, I first present the main ingredients that are used. The sample for the analyses subsets to papers with at least 2 blind scores (98% of the full sample). Table 5a shows the unconditional and conditional score gaps. Table 5b presents the estimated coefficients of the non-blind scoring equation (δ).[35] Table 6 presents the estimated conditional expectation functions:[36] panel A for realized paper quality measures, panel B for reviewer beliefs over those outcomes, and panel C for reviewer beliefs over alternative objectives.

Figure 3 summarizes the decomposition results, where the sum of the bars is the unconditional non-blind score gap, and net of the white bars are the gaps conditional on submission content. Starting with the student score gap, I find that on average, non-blind reviewers scored students around 0.28 points ($\approx$ 0.14 SD) lower than non-students, after controlling for subgroup differences

[35] Consistent with the notion that beliefs of paper quality affect scores, the paper with the higher predicted citation impact is awarded the higher score in 82% of all possible paper pairs within a reviewer's assigned submissions.

[36] Because the interaction terms between submission content and applicant traits are generally statistically insignificant, and the final decomposition estimates are qualitatively similar to the results without the interaction terms, I present the results without interaction terms in the main text for simplicity and show the results with interaction terms in Appendix F.2 (Table A51 shows the decomposition, and Table A49 the regression estimates).

Table 5: Differences in Non-blind Scores by Traits and Reviewer Beliefs

(a) Non-blind Score Gaps Conditional on Submission Content

	(1)	(2)
Student	-0.45***	-0.28*
	(0.12)	(0.14)
Lower Rank Inst.	-0.73***	-0.19
	(0.14)	(0.18)
Female	-0.16	-0.11
	(0.15)	(0.16)
Has Female PI	-0.20	0.13
	(0.16)	(0.19)
Submission Content (blind score)		1.05***
		(0.19)
Subfield FE	×	×
First stage F-stat		26.15
N Paper-(non-blind)reviewer	1265	1265
N Papers	645	645

(b) Non-blind Reviewer Scores and Beliefs

	(1)
Reviewer belief: N citations	0.01**
	(0.00)
Reviewer belief: paper is online	0.48
	(0.38)
Reviewer belief: paper is published	0.64***
	(0.22)
Reviewer belief: engaging talk	0.69***
	(0.10)
Reviewer belief: benefit conference	0.56***
	(0.10)
Reviewer belief: benefit author	0.15*
	(0.09)
Reviewer FE	×
N	434
N Clusters	217
R^2	0.90

Notes. This table shows (a) disparities in non-blind scores from the main experiment, conditional on submission content as proxied by blind scores, following equation 7 (see Tables A49 and A50 for estimates with interactions) and (b) the relationship between non-blind reviewer beliefs and scoring for their assigned submissions, following equation 11. (a) uses the sample from the main experiment, and (b) from the second experiment. Dependent variable for both tables is the non-blind score that a paper receives from a reviewer. Observations are at the paper-reviewer level. (a) subsets to papers with at least 2 blind scores (98% of the full sample). Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

in submission content. I find little support that this gap can be rationalized by accurate statistical discrimination based on paper quality, as proxied by citation and publication outcomes. The point estimate for the extent of accurate statistical discrimination is positive, which suggests that papers submitted by students on average have higher quality conditional on submission content. This is because conditional on submission content, student submissions are significantly more likely to be online and published than non-student ones five years later, with only a slightly negative but statistically insignificant difference in citations (panel A of Table 6; Table A23 shows for earlier years). However, the overall magnitude of the decomposition is small and not statistically significant, implying that this channel plays a minimal role in explaining the student non-blind score gap.

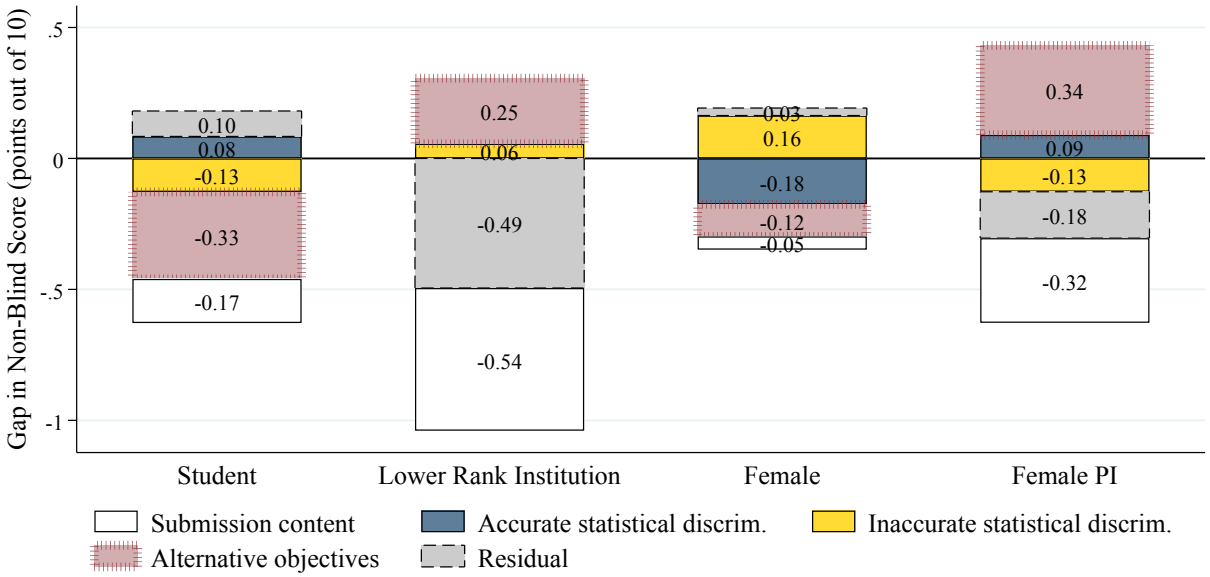
I find suggestive evidence that inaccurate beliefs of paper quality contribute to the student score gap. Point estimates suggest that reviewers are generally over-pessimistic about student submissions relative to non-students' (panel B of Table 6), relative to realized differences. For instance, reviewers predict that student submissions will have 5 fewer citations 5 years later than non-students' (panel B of Table 6), but this difference is around 0.5 in reality among the sample of papers from the main experiment. This may reflect over-reliance on stereotypes Bordalo et al. (2019) and attention discrimination (Bartoš et al., 2016), wherein reviewers pay less attention to the scientific content of student submissions relative to non-students'. While point estimates suggest that inaccurate

Table 6: Disparities in Realized Paper Quality, Perceived Paper Quality, and Alternative Objectives

	Panel A: Realized Paper Quality					
	N Citations		Is Online		Is Published	
	(1)	(2)	(3)	(4)	(5)	(6)
Student	-3.28	-0.49	0.05	0.07**	0.06	0.08**
	(4.09)	(3.99)	(0.03)	(0.03)	(0.04)	(0.04)
Lower Rank Inst.	-12.05**	-2.33	-0.09***	-0.01	-0.05	0.04
	(4.85)	(5.06)	(0.03)	(0.04)	(0.04)	(0.05)
Female	-12.25***	-11.39***	-0.04	-0.03	-0.08*	-0.08
	(3.43)	(3.77)	(0.04)	(0.04)	(0.05)	(0.05)
Has Female PI	-3.53	2.20	0.01	0.06	0.01	0.07
	(4.42)	(4.63)	(0.04)	(0.04)	(0.05)	(0.06)
Submission Content		18.43***		0.16***		0.18***
		(5.71)		(0.04)		(0.04)
Subfield FE	×	×	×	×	×	×
Outcome mean	25.79	25.79	0.79	0.79	0.55	0.55
First stage F-stat	-	26.59	-	26.59	-	26.59
N Papers	645	645	645	645	645	645
	Panel B: Reviewer Beliefs of Paper Quality					
	N Citations		Is Online		Is Published	
	(1)	(2)	(3)	(4)	(5)	(6)
Student	-4.62*	-5.09**	-0.02	-0.01	0.01	0.02
	(2.45)	(2.45)	(0.03)	(0.03)	(0.05)	(0.05)
Lower Rank Inst.	-2.39	1.63	0.03	0.06	-0.03	0.02
	(2.85)	(2.85)	(0.03)	(0.04)	(0.05)	(0.06)
Female	1.38	-0.96	0.02	0.00	0.02	-0.01
	(2.60)	(2.75)	(0.03)	(0.03)	(0.05)	(0.05)
Has Female PI	-2.00	0.56	-0.03	-0.01	-0.08	-0.06
	(2.62)	(2.67)	(0.04)	(0.04)	(0.06)	(0.06)
Submission Content		8.48***		0.06***		0.10***
		(1.94)		(0.02)		(0.03)
Subfield FE	×	×	×	×	×	×
Outcome mean	23.61	23.61	0.91	0.91	0.70	0.70
First stage F-stat	-	66.89	-	84.89	-	84.89
N Papers	383	383	408	408	408	408
	Panel C: Reviewer Beliefs of Alternative Objectives					
	Talk Quality		Benefit Conference		Benefit Author	
	(1)	(2)	(3)	(4)	(5)	(6)
Student	-0.35***	-0.31**	-0.18	-0.15	-0.23*	-0.21*
	(0.12)	(0.12)	(0.12)	(0.12)	(0.13)	(0.13)
Lower Rank Inst.	-0.13	0.13	0.07	0.24*	0.03	0.16
	(0.15)	(0.15)	(0.14)	(0.15)	(0.15)	(0.15)
Female	-0.05	-0.18	0.05	-0.03	0.16	0.10
	(0.13)	(0.13)	(0.12)	(0.12)	(0.13)	(0.13)
Has Female PI	0.18	0.35***	0.06	0.17	-0.07	0.02
	(0.14)	(0.13)	(0.14)	(0.13)	(0.16)	(0.15)
Submission Content		0.55***		0.37***		0.29***
		(0.08)		(0.09)		(0.09)
Subfield FE	×	×	×	×	×	×
Outcome mean	2.93	2.93	3.67	3.67	3.40	3.40
First stage F-stat	-	86.86	-	86.86	-	86.86
N Papers	407	407	407	407	407	407

Notes. This table presents disparities in (panel A) realized paper quality, (panel B) reviewers' predictions of paper quality outcomes and (panel C) alternative objectives, following equation 7 (see Tables A49 and A50 for estimates with interactions). Alternative objectives are: the extent to which the talk would be engaging if it were accepted (1-5 scale), extent to which the author benefits from presenting a talk at the conference (1-5 scale), and the extent to which the conference would benefit from having the author team attend (1-5 scale). Panel A uses data from the main experiment, and Panels B and C use data from the second experiment. All panels subset to papers that have more than one blind score. Even-numbered columns show subgroup differences conditional on submission content as proxied by a paper's blind scores: to account for noise in blind scores, a given blind review score for a paper is instrumented by the average score that the paper's other blind reviewers gave, as described in Section 5.1. Sample is papers with at least 2 blind scores (98% of the full sample). Standard errors are clustered at the paper-level. * $p < .1$, ** $p < .05$, *** $p < .01$

Figure 3: Decomposing Non-Blind Score Gaps



Notes. This table decomposes disparities in non-blind scores. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5) and subfield. Table A43 shows the corresponding confidence intervals based on 1000 clustered bootstrap replications that are drawn at the paper level. Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. Sample is papers with at least 2 blind scores (98% of the full sample). Table A44 shows the decomposition for each sub-component.

statistical discrimination explains a meaningful magnitude (0.13 points) of the total student score gap, these differences between realized and perceived differences are not statistically significant (p-values are $0.33, 0.06, 0.29$ for citations, online, published, respectively). Supplemental analyses that leverage text data from comments that reviewers left for authors additionally support the conclusion that blinding changes reviewer perceptions of the student gap in paper quality (Appendix D).

In fact, model estimates imply that almost the entirety of the student score gap can be explained by differences in reviewers' beliefs over alternative objectives (Figure 3), and this is largely driven by talk quality (Table A44), reflecting the finding that reviewers predict that student submissions will convert to less engaging talks (panel C of Table 6), and the belief that students benefit the conference less than non-students.

In contrast to the student score gap, the non-blind score gap between applicants from top 20 ranked and lower ranked institutions is not significantly explained by alternative objectives

(Table A43). I find little evidence that accurate statistical discrimination explains a significant magnitude of the score difference. In fact, a large portion of the unconditional non-blind score gap is explained by differences in submission content. The rest of the gap remains largely unexplained, which is consistent with the presence of institutional bias, although I fail to reject zero.

The decomposition for the score gap by institution rank illustrates the consequentiality of conditioning on submission content. This can be seen from the first 2 columns of Panel A of Table 6, where applicants from lower-ranked institutions have 12 fewer citations than those from top 20 ranked institutions, but this gap shrinks to 2 and becomes statistically insignificant after accounting for submission content using blind scores. If the decomposition were conducted without proxying for submission content, accurate statistical discrimination would explain a significant portion of the institution rank score gap (Table A52): without any controls for blind scores, around 0.19 points of the score difference by institution rank would be attributed to accurate statistical discrimination. In contrast, this point estimate lies outside of the 90% confidence interval for the decomposition using blind scores (Table A43).

Turning to gender, point estimates suggest that accurate statistical discrimination and alternative objectives can fully explain the applicant gender score gap, although these mechanisms do not significantly differ from zero. In contrast, for PI gender, I find that submission content can explain the entirety of the unconditional score gap. This is consistent with past work exploring how the presence and forms of discrimination can change with rank (e.g. Bohren et al., 2019). In sum, I find little evidence of gender bias.

Including interaction terms between author traits with submission content into the conditional expectation functions: allows for gaps in beliefs and realized quality to differ across the submission content distribution, and to test whether drivers of non-blind score gaps differ across this dimension. I find that doing so generates similar conclusions (Table A51). Decomposing score differences among submissions with median expected blind score (in other words, decomposing $S^{NB}(x, v|S^B(v) = 6) - S^{NB}(x', v|S^B(v) = 6)$) produces similar estimates to the average, likely reflecting that the mass of submission content is near the mean (Table A53).

Finally, in Appendix G, I leverage recent advances in natural language processing approaches, in particular a neural network-based sentence embedding model specifically designed to process scientific papers in fields such as computer science and biology (Singh et al., 2022), to encode the text data from each submission into a high-dimensional vector that summarizes its semantic

content. I find that accounting for these observables adds explanatory power but does not alter the decomposition conclusions, which is consistent with the notion that blind scores sufficiently capture variation in submission content for beliefs and paper outcomes.

5.3 Measuring Paper Quality

The analyses above use citation and publication outcomes as measures of paper quality, which are often used for this purpose both in past research (e.g. Smart and Waldfogel, 1996) as well as in consequential, real-world hiring and promotion decisions. In reality, true paper quality is likely multidimensional and unobserved. Below, I explore a few potential concerns one may have with these measures of paper quality.

First, it is possible that citations and publication processes themselves contain evaluator biases (e.g. Jin et al., 2019; Card et al., 2020; Koffi, 2021), although empirically there is mixed evidence on this (e.g. Smart and Waldfogel, 1996). In my sample, the subgroups that perform worse under non-blind review (those from lower ranked institutions, female applicants and PIs) typically have fewer citations, which may in part reflect biases in the citation process. This would lead me to underestimate the role inaccurate beliefs and overstate the role of accurate statistical discrimination, relative to "true" paper quality. I therefore explore how results change if I inflate the citations of traditionally lower-scoring subgroups by 20 percent, holding all else, such as the scoring equation coefficients, constant (Table A48). I find that in this case, inaccurate statistical discrimination explains a slightly larger portion (0.16 points) of the student score gap.

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 paper quality using an unweighted average over paper's citation count, reviewers may particularly reward submissions with an exceptionally high or high expectation in future paper quality This also reflects the possibility of reviewers' risk aversion. To explore this, I test whether conclusions change with the inclusion of additional functional forms of paper citations: for instance, an indicator for whether the submission placed in the top decile of citation (capturing the notion that reviewers may place greater weight on rewarding the highest-citation papers), or an indicator for whether the submission has at least one citation (capturing the notion that reviewers may seek to prevent the lowest-quality submissions from being accepted). I find that these additional functional forms do not explain a meaningful

share of the variation in non-blind scores beyond the components considered in the decomposition (Figure A25), suggesting that the decomposition results are not skewed due to omission of key functional forms of quality measures.

Finally, to address concerns that quality measures may be endogenously affected by acceptance to the conference, in addition to the evidence presented in Section 3.3, I repeat the model estimation while deflating citation counts for accepted papers by the regression discontinuity estimate for the effect of acceptance on this outcome (Figure A16). These minimally change the decomposition conclusions (see Section C.17).

6 Other Mechanisms

I consider additional explanations for why blinding altered score gaps beyond those considered in the decomposition. One possibility is that blinding reduces reviewer effort. I find little evidence of this. After the 2020 review process, the conference committee asked reviewers an optional question, "How many minutes did you spend per abstract?" Of the 103 reviewers who responded (50 non-blind, 53 blind, out of 245 total), I fail to reject that blind reviewers on average report spending the same amount of time as non-blind ones, both in terms of the average and distribution (Figure A17). Moreover, I find no significant effect of blinding on reviewers' comment lengths, further suggesting that reviewers were not disengaged by blinding (Figure A20). In a similar spirit, reviewers who were randomly assigned to blinding were not any less likely to engage with the conference 2, 3, 4, or 5 years later than those who were randomly assigned to non-blind (Table A27).

Another possibility is that blinding alters reviewers' mental models. First, I find little evidence that blinding causes reviewers to switch to heuristics like rewarding submissions with more figures or easier-to-read text: controlling for text difficulty and figure and word counts minimally changes the main results, and blinding does not significantly affect the returns to these traits (see Appendix G). Similarly, the correlation between reviewers' comment sentiments and scores do not significantly change with blind status (Table A31), implying that the mapping between reviewer beliefs of paper quality and reviewer scores was not substantially affected by blinding.

7 Conclusion

Improving the representation and quality of evaluations requires understanding not just the extent of discrimination, but the specific forms of discrimination driving differences in evaluation outcomes. 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 author traits: previously under-performing groups, particularly students and authors from lower-ranked institutions, benefit. By collecting citation and publication for each submission 5 years later, I find that despite these compositional changes, blind scores are no less predictive of subsequent paper quality than non-blind scores. These results contribute directly to debates over whether blinding policies equalize representation at the expense of quality.

Through the lens of a model of reviewer scores, the results suggest that the effects on composition and lack of change in quality are because reviewers use author identities to both update beliefs about paper quality, and consider alternative objectives such as talk quality. Blinding prevents accurate statistical discrimination, but also reduces scope for bias and considerations of alternative objectives, and these mechanisms pull the effects of blinding on quality in competing directions.

Ultimately, identifying drivers of disparities in non-blind evaluation outcomes is important for understanding optimal policy responses. Past work finds that acceptance to other conferences increase authors' future citations (de Leon and McQuillin, 2020) and co-authorships (Campos et al., 2018), suggesting the importance of studying this context. Moreover, the true value of conferences is likely cumulative, implying that it is important to structure them fairly and effectively in the aggregate.

More broadly, this paper offers a methodological contribution by showing how to combine data on blind evaluation outcomes, quality measures, and belief elicitation, to identify underlying forms of discrimination. My methodology leverages insights from two literatures: research on blinding, which has focused on documenting disparate impacts, and work on disentangling forms discrimination, which has focused on non-blind evaluations. Combining the two approaches allows for relaxing some identification assumptions present in past work. Given that blinding been considered, implemented, and tested in numerous contexts, the approach developed here—leveraging data from blind evaluations to disentangle distinct forms of discrimination—offers a framework that can be applied to many other settings to further understand the sources of disparities.

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.

Alesina, A. and E. La Ferrara (2014). A test of racial bias in capital sentencing. *American Economic Review 104*(11), 3397–3433.

Altmejd, A., A. Barrios-Fernández, M. Drlje, J. Goodman, M. Hurwitz, D. Kovac, C. Mulhern, C. Neilson, and J. Smith (2021). O brother, where start thou? sibling spillovers on college and major choice in four countries. *The Quarterly Journal of Economics 136*(3), 1831–1886.

Aneja, A., M. Luca, and O. Reshef (2025). The benefits of revealing race: Evidence from minority-owned local businesses. *American Economic Review 115*(2), 660–689.

Anwar, S. and H. Fang (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review 96*(1), 127–151.

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.

Ash, E. and S. Hansen (2023). Text algorithms in economics. *Annual Review of Economics 15*, 659–688.

Åslund, O. and O. N. Skans (2012). Do anonymous job application procedures level the playing field? *ILR Review 65*(1), 82–107.

Autor, D. H. and D. Scarborough (2008). Does job testing harm minority workers? evidence from retail establishments. *The Quarterly Journal of Economics 123*(1), 219–277.

Avivi, H. (2024). Are patent examiners gender neutral. *Unpublished manuscript*.

Ayres, I. (2002). Outcome tests of racial disparities in police practices. *Justice research and Policy 4*(1-2), 131–142.

Ayres, I. and J. Waldfogel (1994). A market test for race discrimination in bail setting. *Stanford Law Review*.

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

Barrett, G. F. and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica 71*(1), 71–104.

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.

Boring, A., K. Coffman, D. Glover, and M. J. González-Fuentes (2025). Discrimination, rejection, and willingness to apply: Effects of blind hiring processes. *Unpublished manuscript*.

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.

Brunner, E. J., S. M. Dougherty, and S. L. Ross (2023). The effects of career and technical education: Evidence from the connecticut technical high school system. *Review of Economics and Statistics 105*(4), 867–882.

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.

Campos, R., F. Leon, and B. McQuillin (2018). Lost in the storm: The academic collaborations that went missing in hurricane issac. *The Economic Journal 128*(610), 995–1018.

Canay, I. A., M. Mogstad, and J. Mountjoy (2024). On the use of outcome tests for detecting bias in decision making. *Review of Economic Studies 91*(4), 2135–2167.

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.

Carrell, S., D. Figlio, and L. Lusher (2024). Clubs and networks in economics reviewing. *Journal of Political Economy 132*(9), 2999–3024.

Cattaneo, M. D., M. Jansson, and X. Ma (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association 115*(531), 1449–1455.

Chan, A. (2022). Discrimination against doctors: A field experiment.

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

Coffman, K. B., C. L. Exley, and M. Niederle (2021). The role of beliefs in driving gender discrimination. *Management Science 67*(6), 3551–3569.

Cui, R., J. Li, and D. J. Zhang (2020). Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science 66*(3), 1071–1094.

de Leon, F. L. L. and B. McQuillin (2020). The role of conferences on the pathway to academic impact: Evidence from a natural experiment. *Journal of Human Resources 55*(1), 164–193.

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

Doleac, J. L. and B. Hansen (2020). The unintended consequences of "ban the box": Statistical discrimination and employment outcomes when criminal histories are hidden. *Journal of Labor Economics 38*(2), 321–374.

Doleac, J. L., E. Hengel, and E. Pancotti (2021). Diversity in economics seminars: who gives invited talks? In *AEA Papers and Proceedings*, Volume 111, pp. 55–59. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Edelman, B., M. Luca, and D. Svirsky (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics 9*(2), 1–22.

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.

Ferguson, M. F. and S. R. Peters (1995). What constitutes evidence of discrimination in lending? *The Journal of Finance 50*(2), 739–748.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology 32*(3), 221.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–574.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy 127*(4), 1826–1863.

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.

Grogger, J. and G. Ridgeway (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association 101*(475), 878–887.

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

Hausman, J. A., W. K. Newey, H. Ichimura, and J. L. Powell (1991). Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics 50*(3), 273–295.

Hausman, J. A., W. K. Newey, and J. L. Powell (1995). Nonlinear errors in variables estimation of some engel curves. *Journal of Econometrics 65*(1), 205–233.

Heckman, J. J. (1998). Detecting discrimination. *Journal of economic perspectives 12*(2), 101–116.

Hengel, E. (2022). Publishing while female: Are women held to higher standards? evidence from peer review. *The Economic Journal 132*(648), 2951–2991.

Hinnerich, B. T., E. Hoglin, and M. Johannesson (2011, August). Are boys discriminated in Swedish high schools? *Economics of Education Review 30*(4), 682–690.

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.

Kato, K., Y. Sasaki, and T. Ura (2021). Robust inference in deconvolution. *Quantitative Economics 12*(1), 109–142.

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.

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Kinney, R., C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski, I. Cachola, S. Candra, Y. Chandrasekhar, A. Cohan, et al. (2023). The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

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.

Kline, P., E. K. Rose, and C. R. Walters (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics 137*(4), 1963–2036.

Kline, P. and C. Walters (2021). Reasonable doubt: Experimental detection of job-level employment discrimination. *Econometrica 89*(2), 765–792.

Kline, P. M., E. K. Rose, and C. R. Walters (2024). A discrimination report card. *American Economic Review*.

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.

Kotlarski, I. (1967). On characterizing the gamma and the normal distribution. *Pacific Journal of Mathematics 20*(1), 69–76.

Krause, A., U. Rinne, and K. F. Zimmermann (2012a, December). Anonymous job applications in Europe. *IZA Journal of European Labor Studies 1*(1), 5.

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

Laband, D. N. and M. J. Piette (1994). Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy 102*(1), 194–203.

Lang, K. and A. K.-L. Spitzer (2020). Race discrimination: An economic perspective. *Journal of Economic Perspectives 34*(2), 68–89.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Public Economics 92*(10-11), 2083–2105.

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. Technical report, National Bureau of Economic Research.

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

Low, H. and L. Pistaferri (2025). Disability insurance: Error rates and gender differences. *Journal of Political Economy 133*(9).

Madden, S. and D. DeWitt (2006). Impact of double-blind reviewing on sigmod publication rates. *ACM SIGMOD Record 35*(2), 29–32.

Mountjoy, J. (2024). Marginal returns to public universities. Technical report, National Bureau of Economic Research.

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.

Pallais, A. (2014). Inefficient hiring in entry-level labor markets. *American Economic Review 104*(11), 3565–3599.

Persico, N. and P. Todd (2006). Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in wichita. *The Economic Journal 116*(515), F351–F367.

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*.

Porter, J. and P. Yu (2015). Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics 189*(1), 132–147.

Rao, B. P. (1992). *Identifiability in stochastic models*. Academic Press.

Reiersøl, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica: Journal of the Econometric Society*, 1–24.

Roberts, S. G. and T. Verhoef (2016, July). Double-blind reviewing at EvoLang 11 reveals gender bias. *Journal of Language Evolution 1*(2), 163–167. Publisher: Oxford Academic.

Rose, E. K. (2021). Does banning the box help ex-offenders get jobs? evaluating the effects of a prominent example. *Journal of Labor Economics 39*(1), 79–113.

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

Sarsons, H. (2017b). Recognition for group work: Gender differences in academia. *American Economic Review 107*(5), 141–45.

Schaub, S. and M. S. Schaub (2024). Package 'stodom'.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual review of economics 8*(1), 341–377.

Simoiu, C., S. Corbett-Davies, and S. Goel (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics 11*(3), 1193–1216.

Singh, A., M. D'Arcy, A. Cohan, D. Downey, and S. Feldman (2022). Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

Smart, S. B. and J. Waldfogel (1996). A citation-based test for discrimination at economics and finance journals. Technical report, National Bureau of Economic Research.

Steinmayr, A. (2020). Mhtreg: Stata module for multiple hypothesis testing controlling for fwer.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review 77*, 101981.

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.

Wozniak, A. (2015). Discrimination and the effects of drug testing on black employment. *Review of Economics and Statistics 97*(3), 548–566.

Appendix For Online Publication

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

Haruka Uchida

A Experiment Platform

Figure A1: Reviewer portal

Notes. This figure shows a redacted screenshot of the reviewer portal to show how paper and author information was shared with reviewers. The top row contains the title of the submission. Non-blind reviewers were presented the author names to the right of the "Authors" row. Blind reviewers had the entire row removed. The "Contact Author" was written next to the name of the applicant, so that reviewers could identify the applicant out of the coauthors.

Figure A2: Example submission

Notes. This figure shows a redacted example of a 2-page description that an applicant uploaded with their submission. Both blind and non-blind reviewers assigned to evaluate this submission were able to access this document.

B Collecting Paper Measures

B.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". Race (analyzed in Appendix C.6) was inferred using name and online profiles.

B.2 Paper-level Measures

Papers were searched during the experimental review process, and five 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 five 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.

C Additional Tables and Figures for Experimental Results

C.1 Summary Statistics

Table A1: Paper Subfields

	(1)
computational/modeling	0.63
	(0.48)
techniques	0.32
	(0.47)
neural coding/decoding	0.31
	(0.46)
cognition	0.29
	(0.45)
cortex	0.26
	(0.44)
sensory systems	0.22
	(0.41)
learning/plasticity	0.20
	(0.40)
behavior	0.20
	(0.40)
circuit	0.18
	(0.39)
motor systems	0.07
	(0.26)
disease	0.02
	(0.13)
other representations	0.06
	(0.24)
Observations	657

Notes. This table shows the share of papers that are associated with each subfield category.

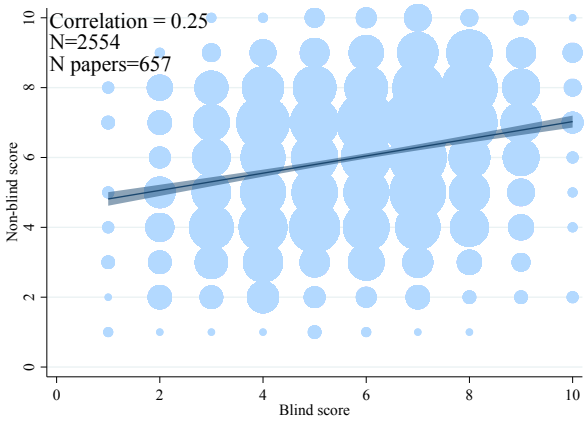
Table A2: Correlations in Author Traits

	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	0.03	1.00											
Inst Rank: 1-20	0.02	-0.01	1.00										
Inst Rank: 21+	-0.01	0.12**	-0.72***	1.00									
Inst Rank: Not University	-0.04	-0.14***	-0.35***	-0.37***	1.00								
Inst Rank: Missing	0.05	-0.02	-0.09*	-0.09*	-0.04	1.00							
Citations: N	-0.07	-0.16***	-0.02	0.04	-0.03	0.00	1.00						
Citations: Below Median	-0.01	0.11**	0.06	-0.01	-0.04	-0.07	-0.14***	1.00					
PI Female	0.04	0.05	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	-0.01	0.02	-0.01	-0.09*	1.00			
PI Citations: Below Median	-0.06	0.02	-0.11**	0.06	0.07	0.00	-0.06	0.06	0.08*	-0.27***	1.00		
Number of Authors	0.05	-0.03	0.07	-0.08*	0.04	-0.05	0.01	-0.03	0.03	0.04	-0.05	1.00	
Solo Author	-0.05	-0.18***	-0.05	0.03	0.03	-0.02	0.05	-0.08	-0.03	-0.05	0.04	-0.22***	1.00

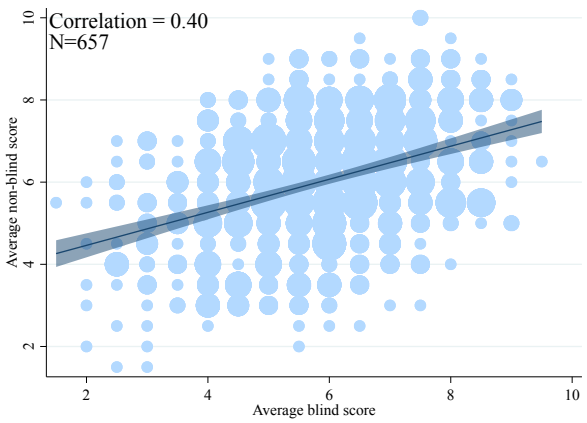
Notes. This table shows the pair-wise correlations between each trait. Observations are at the paper-level.

Figure A3: Correlation Between Blind and Non-Blind Scores

(a) At the reviewer level



(b) At the paper level

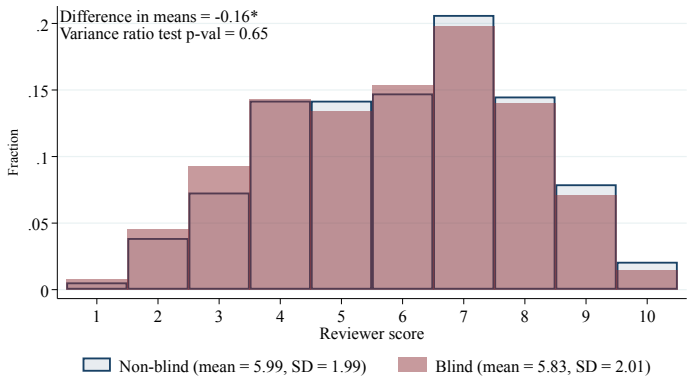


Notes. This figure shows the correlation between a paper's (a) blind score from one reviewer and non-blind score from another reviewer, and (b) average blind score and average non-blind score. In (a), observations are stacked to be at the paper-(blind reviewer)-(non-blind reviewer) level, while in (b), observations are at the paper-level.

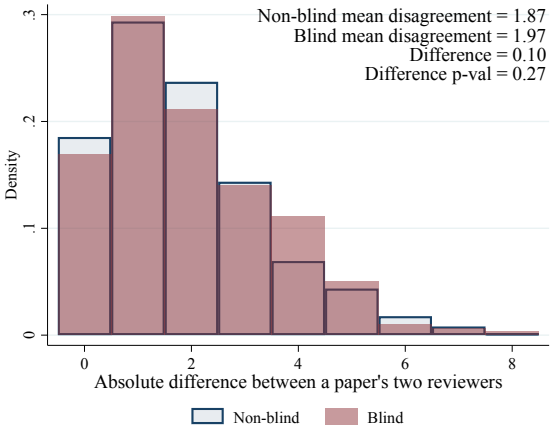
C.2 Blinding Effect on Scores

Figure A4: Effect of Blinding on the Mean and Variance of Reviewer Scores

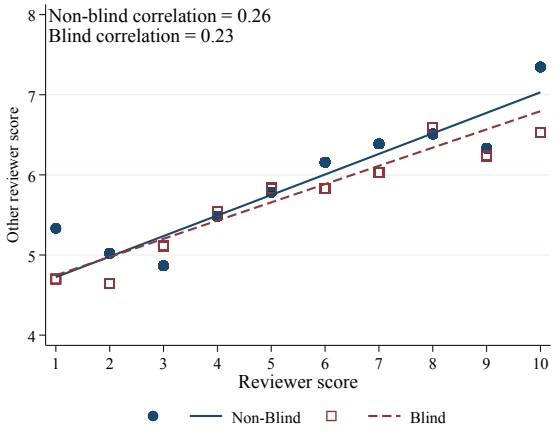
(a) Distribution of blind and non-blind scores



(b) Disagreement between reviewers for the same paper



(c) Reviewer correlation within blind status



Notes. These figures show the effects of blinding on the variance of reviewer scores, using the fact that papers were assigned to two blind reviewers and two non-blind reviewers. (a) plots the distribution of blind and non-blind scores at the paper-reviewer level, and reports the t-test comparison of means and the p-value from a variance ratio test. (b) shows the across-paper distribution of absolute difference between a paper's two reviewers of the same blind status. The clear bars with dark blue outlines show the density for the difference between a paper's two non-blind reviewers, while the red bars show for a paper's two blind reviewers. (c) shows the correlation between the two reviewers for the same paper and same blind status. Observation are at the paper level. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A3: Blind and Non-Blind Score Gaps

	Non-Blind Scores					Blind Scores				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student	-0.54***				-0.49***	-0.18				-0.17
	(0.11)				(0.11)	(0.11)				(0.12)
Lower Rank Inst.		-0.80***			-0.78***		-0.59***			-0.59***
		(0.12)			(0.12)		(0.12)			(0.12)
Female			-0.28*		-0.26			-0.05		-0.04
			(0.16)		(0.16)			(0.13)		(0.14)
Has Female PI				-0.26*	-0.29*				-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 Reviewers	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	0.13	0.13	0.18	0.12	0.13	0.11	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. * $p < .1$, ** $p < .05$, *** $p < .01$

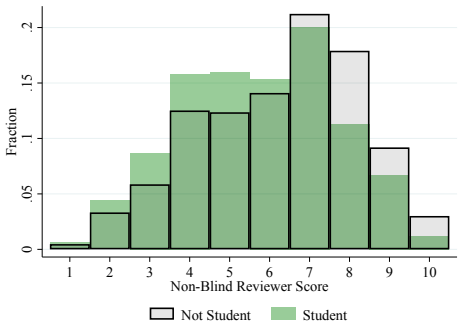
Table A4: Effects of Blinding on Reviewer Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Reviewer is Blind	-0.17*	-0.33***	-0.29**	-0.20**	-0.16*	-0.46***	
	(0.09)	(0.11)	(0.12)	(0.09)	(0.10)	(0.15)	
Student × Reviewer is Blind		0.32**				0.31**	0.31**
		(0.13)				(0.13)	(0.13)
Lower Rank Institution × Reviewer is Blind			0.25*			0.23	0.29*
			(0.15)			(0.15)	(0.15)
Female × Reviewer is Blind				0.11		0.11	0.26
				(0.17)		(0.17)	(0.16)
Has Female PI × Reviewer is 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
N Reviewers	245	245	245	245	245	245	245
N Papers	657	657	657	657	657	657	657
R^2	0.44	0.44	0.44	0.44	0.44	0.44	0.57

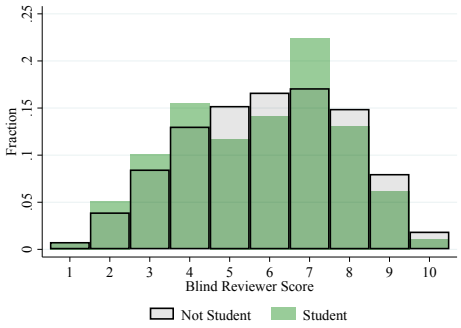
Notes. Dependent variable is the score that a reviewer gave to a paper. Observations are at the paper-reviewer level. Blind is an indicator for whether the reviewer was randomly assigned to score papers without author lists. Student is an indicator for whether the applicant is a student. "Lower ranked" 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: 0.07, 0.18, 0.19, 0.77, respectively. * $p < .1$, ** $p < .05$, *** $p < .01$

Figure A5: Distributions of Blind and non-blind Scores by Trait

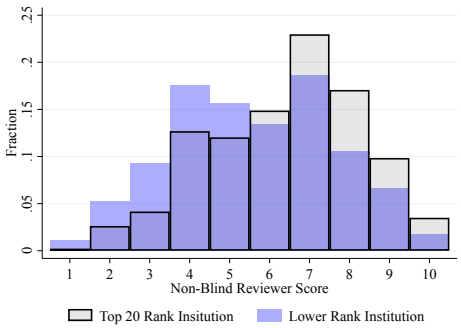
(a) Non-Blind, by Student Status



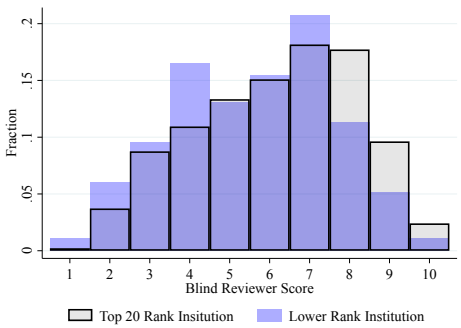
(b) Blind, by Student Status



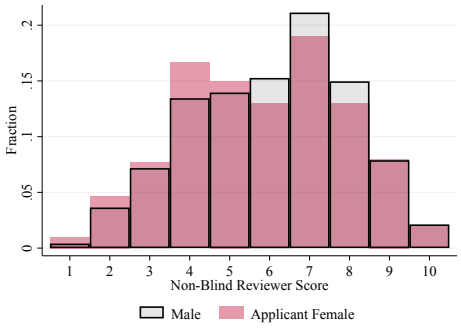
(c) Non-Blind, by Applicant Institution



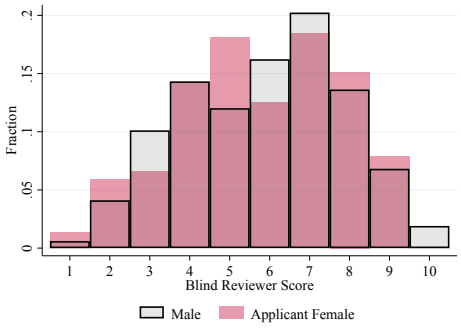
(d) Blind, by Applicant Institution



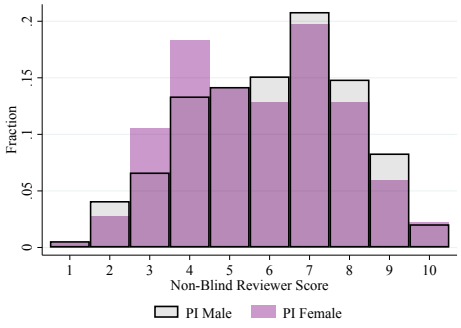
(e) Non-Blind, by Applicant Gender



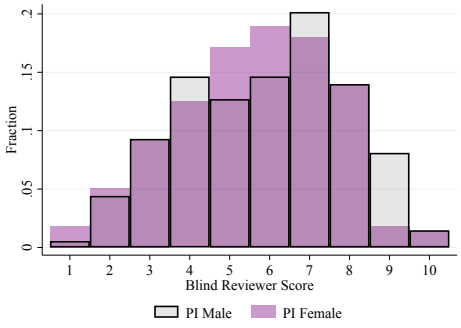
(f) Blind, by Applicant Gender



(g) Non-Blind, by PI Gender

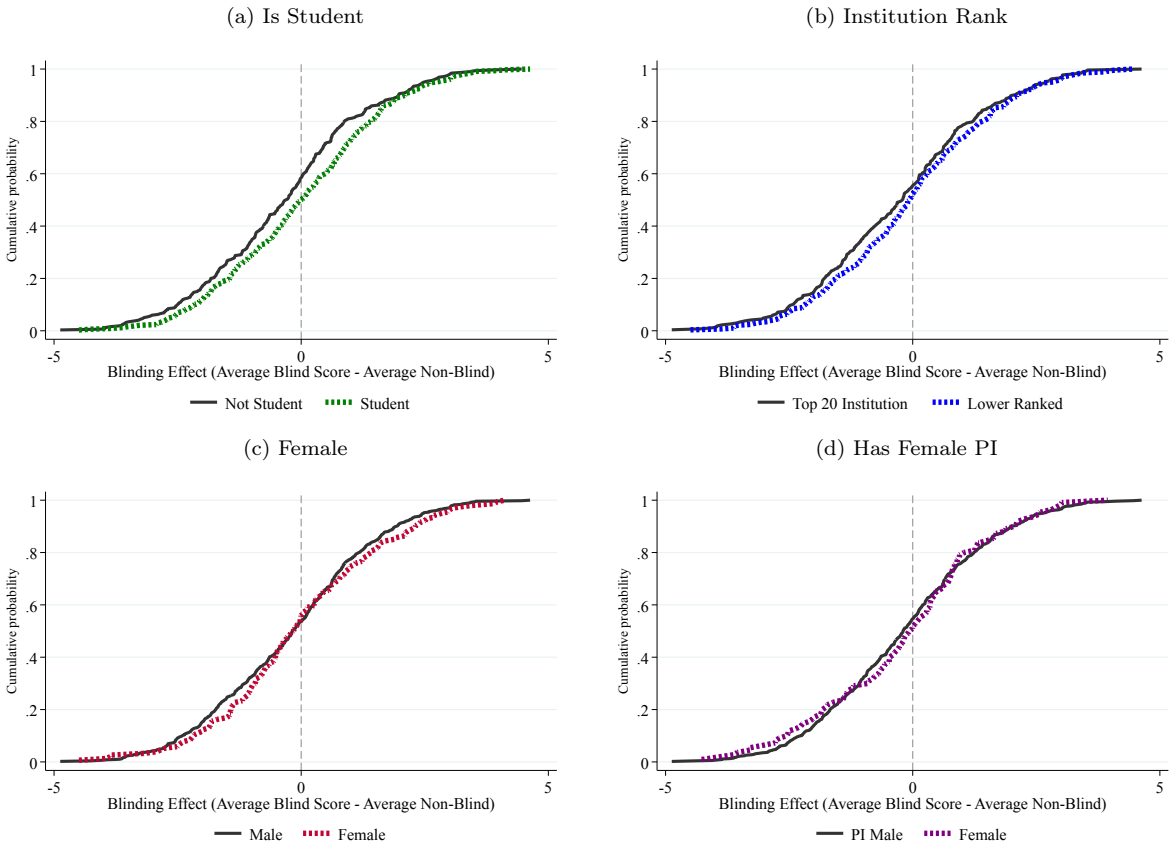


(h) Blind, by PI Gender



Notes. These figures show the distribution of blind and non-blind scores by each author trait. Observations are at the paper-reviewer level.

Figure A6: 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.

Table A5: Effects of Blinding on Reviewer Scores: Interactions

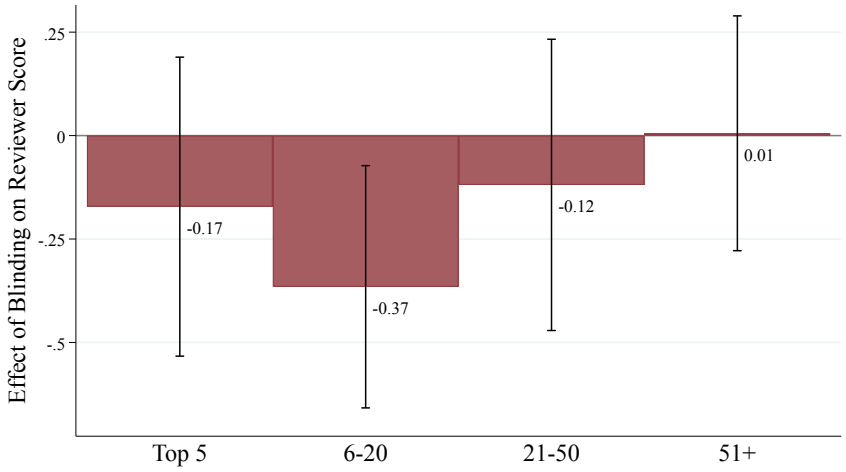
	(1)	(2)	(3)	(4)	(5)	(6)
Applicant Student × Blind	0.35**	0.31**	0.27*	0.26*	0.32**	0.26
	(0.18)	(0.13)	(0.14)	(0.14)	(0.13)	(0.20)
Lower Rank Inst. × Blind	0.33	0.27	0.28*	0.28*	0.29*	0.31
	(0.21)	(0.17)	(0.15)	(0.15)	(0.15)	(0.23)
Applicant Female × Blind	0.26	0.23	0.16	0.27*	0.20	0.07
	(0.16)	(0.23)	(0.23)	(0.16)	(0.17)	(0.29)
PI Female × Blind	-0.06	-0.06	-0.05	-0.23	-0.15	-0.34
	(0.20)	(0.20)	(0.20)	(0.28)	(0.23)	(0.30)
Student × Lower Rank Inst. × Blind	-0.08					-0.08
	(0.26)					(0.26)
Female × Lower Rank Inst. × Blind		0.07				0.06
		(0.34)				(0.34)
Student × Female × Blind			0.19			0.18
			(0.31)			(0.30)
Student × PI Female × Blind				0.32		0.33
				(0.36)		(0.36)
Female × PI Female × Blind					0.35	0.40
					(0.40)	(0.40)
Reviewer FE	×	×	×	×	×	×
Paper FE	×	×	×	×	×	×
N	2591	2591	2591	2591	2591	2591
N Clusters	245	245	245	245	245	245
R^2	0.57	0.57	0.57	0.57	0.57	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. * $p < .1$, ** $p < .05$, *** $p < .01$

C.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 A7 shows the change in scores induced by blinding, binning by finer categories of institution rank.

Figure A7: 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 show coefficients from interacting blind status with finer bins of institution rank. Bar ticks correspond to the 95% confidence intervals.

C.4 Heterogeneity By Coauthor Traits

Table A6: Blinding Effect by Applicant, Coauthor, and PI Traits

	(1)	(2)
Student × Blind	0.30**	0.31**
	(0.13)	(0.13)
Lower Rank Inst. × Blind	0.36*	0.31
	(0.21)	(0.22)
Female × Blind	0.28*	0.30*
	(0.16)	(0.16)
PI Female × Blind	-0.08	-0.09
	(0.20)	(0.20)
N Coauthors Student × Blind	-0.09	-0.09
	(0.12)	(0.12)
N Coauthors Lower Ranked Inst × Blind	-0.05	-0.03
	(0.08)	(0.08)
N Coauthors Female × Blind	0.03	0.01
	(0.12)	(0.12)
N Coauthors × Blind	0.02	0.02
	(0.07)	(0.07)
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. * $p < .1$, ** $p < .05$, *** $p < .01$

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

Table A7: Blind and Non-blind Score Gaps, with Authors' Past Citations

	Non-Blind Scores				Blind Scores			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
App. Citations: Above Median	0.24*	0.05			0.03	-0.00		
	(0.14)	(0.14)			(0.16)	(0.16)		
PI Citations: Above Median	0.10	0.03			0.22	0.13		
	(0.13)	(0.13)			(0.14)	(0.15)		
Student		-0.48***		-0.48***		-0.11		-0.07
		(0.11)		(0.12)		(0.12)		(0.12)
Lower Rank Inst.		-0.77***		-0.77***		-0.56***		-0.50***
		(0.12)		(0.13)		(0.13)		(0.13)
Female		-0.26		-0.25		-0.04		-0.07
		(0.16)		(0.16)		(0.14)		(0.14)
Has Female PI		-0.27*		-0.26*		-0.29*		-0.32**
		(0.16)		(0.15)		(0.16)		(0.16)
App. Citations: Q2			0.32	-0.05			0.55***	0.42*
			(0.20)	(0.21)			(0.21)	(0.22)
App. Citations: Q3			0.36	-0.01			0.51***	0.40*
			(0.24)	(0.23)			(0.19)	(0.20)
App. Citations: Q4			0.44**	0.07			0.10	0.03
			(0.20)	(0.21)			(0.22)	(0.23)
PI Citations: Q2			0.31*	0.09			0.11	0.01
			(0.18)	(0.18)			(0.17)	(0.17)
PI Citations: Q3			0.17	0.05			0.27	0.15
			(0.17)	(0.17)			(0.17)	(0.18)
PI Citations: Q4			0.31*	0.12			0.21	0.08
			(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	0.15	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. * $p < .1$, ** $p < .05$, *** $p < .01$

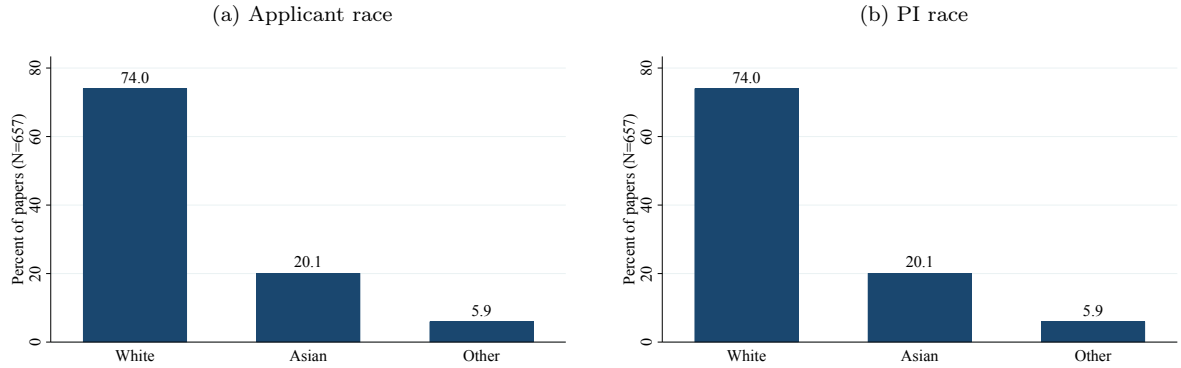
Table A8: Blinding Effect on Scores, with Authors' Past Citations

	(1)	(2)	(3)	(4)
App. Citations: Above Median × Blind	-0.14	0.04		
	(0.16)	(0.17)		
PI Citations: Above Median × Blind	0.15	0.14		
	(0.13)	(0.13)		
Student × Blind		0.32**		0.38***
		(0.13)		(0.15)
Lower Rank Inst. × Blind		0.30*		0.36**
		(0.15)		(0.16)
Female × Blind		0.25		0.24
		(0.16)		(0.16)
PI Female × Blind		-0.03		-0.06
		(0.20)		(0.20)
App. Citations: Quartile 2 × Blind			0.16	0.37*
			(0.18)	(0.20)
App. Citations: Quartile 3 × Blind			0.06	0.29
			(0.20)	(0.21)
App. Citations: Quartile 4 × Blind			-0.26	-0.00
			(0.21)	(0.23)
PI Citations: Quartile 2 × Blind			-0.06	0.01
			(0.18)	(0.19)
PI Citations: Quartile 3 × Blind			0.17	0.16
			(0.17)	(0.18)
PI Citations: Quartile 4 × Blind			0.07	0.10
			(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	0.56	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. * $p < .1$, ** $p < .05$, *** $p < .01$

C.6 Heterogeneity By Race

Figure A8: Racial composition of Applicants and PIs



Notes. These figures show the racial breakdown of applicants and principal investigators (PIs) associated with submissions. "Other" is non-White, non-Asian. Observation are at the paper level.

Table A9: Blind and Non-Blind Score Gaps, with Authors' Race

	Non-Blind Scores					Blind Scores				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Applicant race: Asian	-0.20		-0.21*	-0.17	-0.20	-0.00		0.00	0.01	-0.05
	(0.13)		(0.13)	(0.13)	(0.14)	(0.12)		(0.12)	(0.12)	(0.13)
Applicant race: Other	-0.35*		-0.36*	-0.21	-0.21	-0.31		-0.29	-0.21	-0.28
	(0.19)		(0.19)	(0.20)	(0.22)	(0.20)		(0.20)	(0.21)	(0.21)
PI race: Asian		-0.05	-0.04	0.07	0.09		-0.29**	-0.28*	-0.20	-0.27*
		(0.13)	(0.13)	(0.13)	(0.14)		(0.14)	(0.14)	(0.14)	(0.15)
PI race: Other		-0.41	-0.44*	-0.31	-0.34		-0.14	-0.14	-0.08	-0.19
		(0.26)	(0.26)	(0.23)	(0.23)		(0.26)	(0.27)	(0.26)	(0.26)
Student				-0.46***	-0.49***				-0.16	-0.17
				(0.11)	(0.11)				(0.11)	(0.11)
Lower Rank Inst.				-0.72***	-0.77***				-0.52***	-0.57***
				(0.12)	(0.12)				(0.12)	(0.13)
Female				-0.16	-0.24				-0.05	-0.03
				(0.14)	(0.16)				(0.14)	(0.14)
Has Female PI				-0.29*	-0.33**				-0.29**	-0.29*
				(0.15)	(0.15)				(0.14)	(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
R^2	0.00	0.00	0.01	0.07	0.18	0.00	0.00	0.00	0.03	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. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A10: Blinding Effect on Scores, with Authors' Race

	(1)	(2)	(3)	(4)
Applicant race: Asian × Blind	0.16		0.18	0.17
	(0.15)		(0.15)	(0.15)
Applicant race: Other × Blind	-0.01		0.01	-0.05
	(0.24)		(0.25)	(0.25)
PI race: Asian × Blind		-0.27	-0.28	-0.28
		(0.17)	(0.17)	(0.17)
PI race: Other × Blind		0.17	0.19	0.11
		(0.27)	(0.28)	(0.27)
Student × Blind				0.31
				(0.13)
Lower Rank Inst. × Blind				0.30
				(0.15)
Female × Blind				0.24
				(0.16)
PI Female × Blind				-0.01
				(0.20)
Reviewer FE	×	×	×	×
Paper FE	×	×	×	×
N	2591	2591	2591	2591
N Clusters	245	245	245	245
R^2	0.56	0.56	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. * $p < .1$, ** $p < .05$, *** $p < .01$

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

Table A11: Blinding Effects by Whether Applicant or PI Is New

	(1)	(2)	(3)
Applicant Student × Blind	0.42*	0.42**	0.46**
	(0.21)	(0.16)	(0.22)
Lower Rank Inst. × Blind	0.52**	0.28	0.51*
	(0.26)	(0.21)	(0.27)
Applicant Female × Blind	0.12	0.34*	0.14
	(0.27)	(0.19)	(0.27)
PI Female × Blind	-0.36	-0.24	-0.43
	(0.31)	(0.24)	(0.31)
New Applicant × Blind	0.59**		0.56*
	(0.29)		(0.30)
New applicant × Applicant Student × Blind	-0.25		-0.18
	(0.28)		(0.30)
New applicant × Lower Rank Inst. × Blind	-0.37		-0.36
	(0.34)		(0.36)
New applicant × Applicant Female × Blind	0.22		0.34
	(0.31)		(0.36)
New applicant × PI Female × Blind	0.48		0.34
	(0.39)		(0.43)
New PI × Blind		0.32	0.16
		(0.29)	(0.31)
New PI × Applicant Student × Blind		-0.31	-0.26
		(0.29)	(0.32)
New PI × Lower Rank Inst. × Blind		-0.13	-0.04
		(0.33)	(0.34)
New PI × Applicant Female × Blind		-0.24	-0.36
		(0.33)	(0.38)
New PI × PI Female × Blind		0.52	0.44
		(0.36)	(0.40)
Paper FE	×	×	×
Reviewer FE	×	×	×
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. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

C.8 Heterogeneity by Reviewer Traits

Table A12: Scoring Heterogeneity by Reviewer Gender

	Non-Blind				Blind			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student	-0.42***	-0.41***			-0.10	-0.10		
	(0.15)	(0.14)			(0.13)	(0.14)		
Lower Rank Inst.	-0.77***	-0.78***			-0.54***	-0.64***		
	(0.14)	(0.14)			(0.14)	(0.15)		
Female	-0.15	-0.23			0.04	0.06		
	(0.19)	(0.21)			(0.18)	(0.17)		
Has Female PI	-0.29	-0.28			-0.31	-0.29		
	(0.19)	(0.18)			(0.19)	(0.21)		
Reviewer female	-0.13		-0.51		-0.09		-0.05	
	(0.22)		(0.36)		(0.26)		(0.29)	
Student × Reviewer Female	-0.14	-0.23	-0.12	-0.18	-0.14	-0.16	0.08	0.21
	(0.23)	(0.23)	(0.30)	(0.30)	(0.23)	(0.25)	(0.27)	(0.29)
Lower Rank Inst. × Reviewer Female	0.13	0.02	0.83**	0.87**	0.02	0.12	0.09	0.14
	(0.26)	(0.27)	(0.40)	(0.37)	(0.24)	(0.26)	(0.32)	(0.33)
Female × Reviewer Female	-0.11	-0.10	-0.01	0.10	-0.26	-0.28	-0.29	-0.18
	(0.29)	(0.32)	(0.41)	(0.35)	(0.27)	(0.27)	(0.30)	(0.31)
Has Female PI × Reviewer Female	0.09	-0.02	0.40	0.55	-0.02	-0.09	-0.11	-0.09
	(0.31)	(0.32)	(0.41)	(0.36)	(0.29)	(0.30)	(0.34)	(0.36)
Reviewer FE		×		×		×		×
Paper FE			×	×			×	×
N	1289	1289	1264	1264	1302	1302	1290	1290
N Clusters	119	119	119	119	126	126	126	126
R^2	0.07	0.18	0.64	0.75	0.04	0.14	0.62	0.72

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. All regressions control for subfield. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A13: Scoring Heterogeneity by Reviewer Institution

	Non-Blind				Blind			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student	-0.49***	-0.39**			-0.27	-0.22		
	(0.17)	(0.17)			(0.17)	(0.20)		
Lower Rank Inst.	-0.63***	-0.56***			-0.49***	-0.47**		
	(0.16)	(0.17)			(0.17)	(0.20)		
Female	0.31	0.36			0.50***	0.64***		
	(0.21)	(0.22)			(0.17)	(0.17)		
Has Female PI	-0.63**	-0.63**			-0.28	-0.24		
	(0.26)	(0.24)			(0.22)	(0.25)		
Reviewer: Lower rank inst.	0.11		0.31		0.38*		0.62**	
	(0.21)		(0.30)		(0.22)		(0.31)	
Student × Reviewer Lower Rank Inst	-0.00	-0.14	0.25	0.20	0.21	0.16	-0.22	-0.48
	(0.23)	(0.23)	(0.29)	(0.30)	(0.24)	(0.26)	(0.30)	(0.30)
Lower Rank Inst. × Reviewer Lower Rank Inst	-0.09	-0.22	-0.37	-0.58	-0.13	-0.13	-0.51	-0.47
	(0.26)	(0.27)	(0.37)	(0.37)	(0.25)	(0.27)	(0.38)	(0.37)
Female × Reviewer Lower Rank Inst	-0.69**	-0.83***	-0.76*	-0.90**	-1.05***	-1.11***	-0.64*	-0.51
	(0.28)	(0.31)	(0.43)	(0.39)	(0.30)	(0.30)	(0.37)	(0.37)
Has Female PI × Reviewer Lower Rank Inst	0.58*	0.59*	0.63*	0.69	-0.03	-0.16	-0.09	0.12
	(0.33)	(0.33)	(0.34)	(0.43)	(0.34)	(0.34)	(0.44)	(0.50)
Reviewer FE		×		×		×		×
Paper FE			×	×			×	×
N	1289	1289	1264	1264	1302	1302	1290	1290
N Clusters	119	119	119	119	126	126	126	126
R^2	0.10	0.21	0.64	0.76	0.05	0.16	0.63	0.73

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. All regressions control for subfield. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A14: Scoring Heterogeneity by Reviewer Experience

	Non-Blind				Blind			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student	-0.67***	-0.69***			-0.34	-0.39*		
	(0.19)	(0.18)			(0.23)	(0.24)		
Lower Rank Inst.	-0.68***	-0.75***			-0.75***	-0.81***		
	(0.24)	(0.26)			(0.26)	(0.29)		
Female	-0.42	-0.53*			-0.19	-0.23		
	(0.28)	(0.30)			(0.34)	(0.32)		
Has Female PI	-0.22	-0.23			-0.02	-0.30		
	(0.23)	(0.22)			(0.27)	(0.30)		
Reviewer: Experienced	-0.19		-0.46		0.01		0.59	
	(0.24)		(0.30)		(0.27)		(0.37)	
Student × Reviewer Experienced	0.32	0.31	0.44	0.21	0.23	0.30	-0.53	-0.82**
	(0.23)	(0.23)	(0.34)	(0.37)	(0.26)	(0.27)	(0.37)	(0.39)
Lower Rank Inst. × Reviewer Experienced	-0.08	-0.05	0.09	-0.08	0.30	0.32	0.05	-0.19
	(0.27)	(0.29)	(0.38)	(0.39)	(0.29)	(0.32)	(0.35)	(0.34)
Female × Reviewer Experienced	0.38	0.44	0.26	0.40	0.20	0.28	0.40	0.25
	(0.31)	(0.35)	(0.40)	(0.40)	(0.37)	(0.34)	(0.38)	(0.33)
Has Female PI × Reviewer Experienced	-0.07	-0.10	-0.07	-0.79*	-0.41	-0.04	-0.61	-0.41
	(0.31)	(0.30)	(0.46)	(0.47)	(0.32)	(0.35)	(0.40)	(0.41)
Reviewer FE		×		×		×		×
Paper FE			×	×			×	×
N	1289	1289	1264	1264	1302	1302	1290	1290
N Clusters	119	119	119	119	126	126	126	126
R^2	0.07	0.18	0.63	0.75	0.04	0.14	0.62	0.72

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. All regressions control for subfield. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A15: Blinding Effect Heterogeneity by Reviewer Traits

	(1)	(2)	(3)	(4)
Applicant Student × Blind	0.20	0.37	0.60**	0.53
	(0.16)	(0.24)	(0.28)	(0.38)
Lower Rank Inst. × Blind	0.39**	0.32	0.09	0.32
	(0.19)	(0.24)	(0.32)	(0.43)
Applicant Female × Blind	0.36*	0.61**	0.06	0.69
	(0.20)	(0.28)	(0.32)	(0.49)
PI Female × Blind	0.24	0.27	-0.47	0.26
	(0.28)	(0.39)	(0.40)	(0.66)
Student × Reviewer Female	-0.18			-0.23
	(0.23)			(0.23)
Student × Reviewer Female × Blind	0.30			0.34
	(0.33)			(0.33)
Lower Ranked Inst × Reviewer Female	0.41			0.41
	(0.27)			(0.27)
Lower Ranked Inst × Reviewer Female × Blind	-0.25			-0.30
	(0.34)			(0.35)
Applicant Female × Reviewer Female	0.21			0.18
	(0.29)			(0.29)
Female × Reviewer Female × Blind	-0.26			-0.55
	(0.36)			(0.37)
PI Female × Reviewer Female	0.39			0.38
	(0.32)			(0.32)
PI Female × Reviewer Female × Blind	-0.75			-0.78
	(0.46)			(0.48)
Student × Reviewer Lower Inst		0.09		0.08
		(0.24)		(0.24)
Student × Reviewer Lower Inst × Blind		-0.06		-0.01
		(0.33)		(0.32)
Lower Ranked Inst × Reviewer Lower Inst		-0.35		-0.35
		(0.27)		(0.27)
Lower Ranked Inst × Reviewer Lower Inst × Blind		-0.06		-0.06
		(0.37)		(0.38)
Female × Reviewer Lower Inst		-0.37		-0.28
		(0.29)		(0.29)
Female × Reviewer Lower Inst × Blind		-0.64*		-0.84**
		(0.38)		(0.38)
PI Female × Reviewer Lower Inst		0.27		0.30
		(0.33)		(0.32)
PI Female × Reviewer Lower Inst × Blind		-0.46		-0.60
		(0.51)		(0.52)
Student × Reviewer Experienced			-0.18	-0.20
			(0.24)	(0.24)
Lower Ranked Inst × Reviewer Experienced			-0.38	-0.32
			(0.28)	(0.28)
Female × Reviewer Experienced			-0.09	-0.20
			(0.31)	(0.33)
PI Female × Reviewer Experienced			-0.71**	-0.64*
			(0.31)	(0.33)
Student × Reviewer Experienced × Blind			-0.39	-0.40
			(0.33)	(0.33)
Lower Ranked Inst × Reviewer Experienced × Blind			0.29	0.16
			(0.37)	(0.38)
Female × Reviewer Experienced × Blind			0.30	0.30
			(0.40)	(0.41)
PI Female × Reviewer Experienced × Blind			0.57	0.47
			(0.49)	(0.52)
Reviewer FE	×	×	×	×
Paper FE	×	×	×	×
N	2591	2591	2591	2591
N Clusters	245	245	245	245
R^2	0.57	0.58	0.57	0.58

Dependent variable is the score that a reviewer gave to a paper. Observations are at the paper-reviewer level. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

C.9 Robustness and Reviewer Guesses of Author Identity

In this section, I run robustness checks to address two potential concerns with the main analysis as discussed in Section 3.1: (1) that blind reviewers were not truly blind since they could have learned author identities elsewhere, and (2) not every reviewer submitted their assigned reviews. Table A16 summarizes the results from each (see Table A17 for selection into these subsamples). The first row column the main estimates. The second column subsets to 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. 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 because reviewers may have other avenues of learning about the paper and its authors, it provides suggestive evidence by removing the papers most likely to be identifiable by reviewers.

Table A16: Robustness for Effects of Blinding on Reviewer Scores

	Main	NotOnline	NotGuessed	NotMiss	IPW
	(1)	(2)	(3)	(4)	(5)
Student × Blind	0.31**	0.29**	0.37***	0.32**	0.31**
	(0.13)	(0.14)	(0.14)	(0.14)	(0.14)
Lower Rank Inst. × Blind	0.29*	0.20	0.26	0.28*	0.28*
	(0.15)	(0.16)	(0.17)	(0.16)	(0.16)
Female × Blind	0.26	0.32*	0.27	0.24	0.23
	(0.16)	(0.17)	(0.17)	(0.16)	(0.17)
Has Female PI × Blind	-0.06	-0.05	-0.28	-0.11	-0.11
	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)
Paper FE	×	×	×	×	×
Reviewer FE	×	×	×	×	×
N	2591	2170	2391	2480	2480
N Clusters	245	245	245	245	245
N Papers	657	551	657	620	620
R^2	0.57	0.59	0.59	0.57	0.57

Notes. This table examines the robustness of blinding effects on reviewer score gaps, using equation 1. The first column presents the main effects from the main text. The subsequent columns: (2) sub-set to papers that were not available online at the time of the review process (3) sub-set to paper-reviewers where reviewers did not guess the author identity correctly (4) sub-set to papers that are not missing a review (5) use inverse probability weighting to acknowledge the fact that some papers were 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. * $p < .1$, ** $p < .05$, *** $p < .01$

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 A16). Second, I implement inverse probability weighting, using a paper's applicant and PI traits to predict the likelihood that it is missing

reviews (Column 5 of Table A16; see Table A17 for prediction regression). 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.

Table A17: Selection into Robustness Subsamples

	Available Online						Missing Review					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Student	-0.05*				-0.05*	-0.05*	-0.04**				-0.04**	-0.03*
	(0.03)				(0.03)	(0.03)	(0.02)				(0.02)	(0.02)
Lower Rank Inst.		-0.03			-0.02	-0.03		-0.03			-0.03	-0.03
		(0.03)			(0.03)	(0.03)		(0.02)			(0.02)	(0.02)
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)
Female PI				-0.08***	-0.08**	-0.08**				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	0.00	0.01	0.02	0.04	0.01	0.01	0.00	0.00	0.02	0.04

Notes. This table shows which papers were removed in the robustness checks in Table A16. 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. * $p < .1$, ** $p < .05$, *** $p < .01$

Since the implementation of blind review in 2020 (the year of the main experiment), the conference included an optional text-box question in the review portal that asked "Please provide your best guess as to the identity of the lab (PI name) of this submission." Around 40% of reviews (at the paper-reviewer level) are associated with a guess (Table A18), and of those, around 12-15% explicitly state that the reviewer is not sure. Around 36-41% of guesses are correct: this is slightly less than Blank (1991) who finds that 46% of blind referees correctly identify the author.

There is selection into guessing (Table A19). Reviewers who are students are significantly less likely to give a guess, but conditional on guessing, are more likely to guess correctly. Female reviewers are significantly less likely to give a correct guess, both conditional on guessing and unconditionally.

Table A18: Reviewers' Author Guesses

Variable	(1) Main experiment	(2) Followup experiment
Has Author Guess	0.37	0.39
	(0.48)	(0.49)
Guess: correct	0.15	0.14
	(0.36)	(0.35)
Guess: says not sure (cond. on guessing)	0.15	0.12
	(0.36)	(0.32)
Guess: correct (cond. on guessing)	0.41	0.36
	(0.49)	(0.48)
Observations	1,302	3,009

Notes. This table provides summary statistics for the guesses that reviewers gave for the authors of their submissions. Columns correspond to experimental year. "Guess: says not sure" is an indicator for if the reviewer explicitly responds that they are not sure of the author, for example "not sure", "I have no idea." Variables with "cond. on guessing" subset to observations with a guess. Observation is at the paper-reviewer level, subsetting only to blind reviews. Standard deviations in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A19: Selection in Reviewers' Guesses

	Gave guess			Guess is correct | guess			Guess is correct		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Applicant: Student	-0.02		-0.01	0.01		0.01	-0.01		-0.01
	(0.03)		(0.03)	(0.05)		(0.05)	(0.02)		(0.02)
Applicant: Lower Rank Institution	-0.04		-0.03	-0.06		-0.05	-0.04*		-0.03
	(0.03)		(0.03)	(0.05)		(0.05)	(0.02)		(0.02)
Applicant: Female	-0.01		-0.02	0.07		0.06	0.02		0.02
	(0.03)		(0.03)	(0.05)		(0.06)	(0.02)		(0.03)
Female PI	-0.00		-0.00	0.03		0.01	0.01		0.01
	(0.04)		(0.04)	(0.06)		(0.06)	(0.03)		(0.03)
Reviewer: Student		-0.26***	-0.25***		0.42**	0.49***		-0.06	-0.04
		(0.05)	(0.06)		(0.18)	(0.19)		(0.05)	(0.05)
Reviewer: Lower Rank Inst.		0.03	0.02		-0.07	-0.06		-0.01	-0.01
		(0.03)	(0.03)		(0.05)	(0.05)		(0.02)	(0.02)
Reviewer: Female		-0.01	-0.00		-0.14***	-0.14***		-0.05**	-0.05**
		(0.03)	(0.03)		(0.05)	(0.05)		(0.02)	(0.02)
Subfield FE	×	×	×	×	×	×	×	×	×
N	1302	1302	1302	487	487	487	1302	1302	1302
R^2	0.01	0.01	0.03	0.01	0.03	0.08	0.01	0.01	0.02

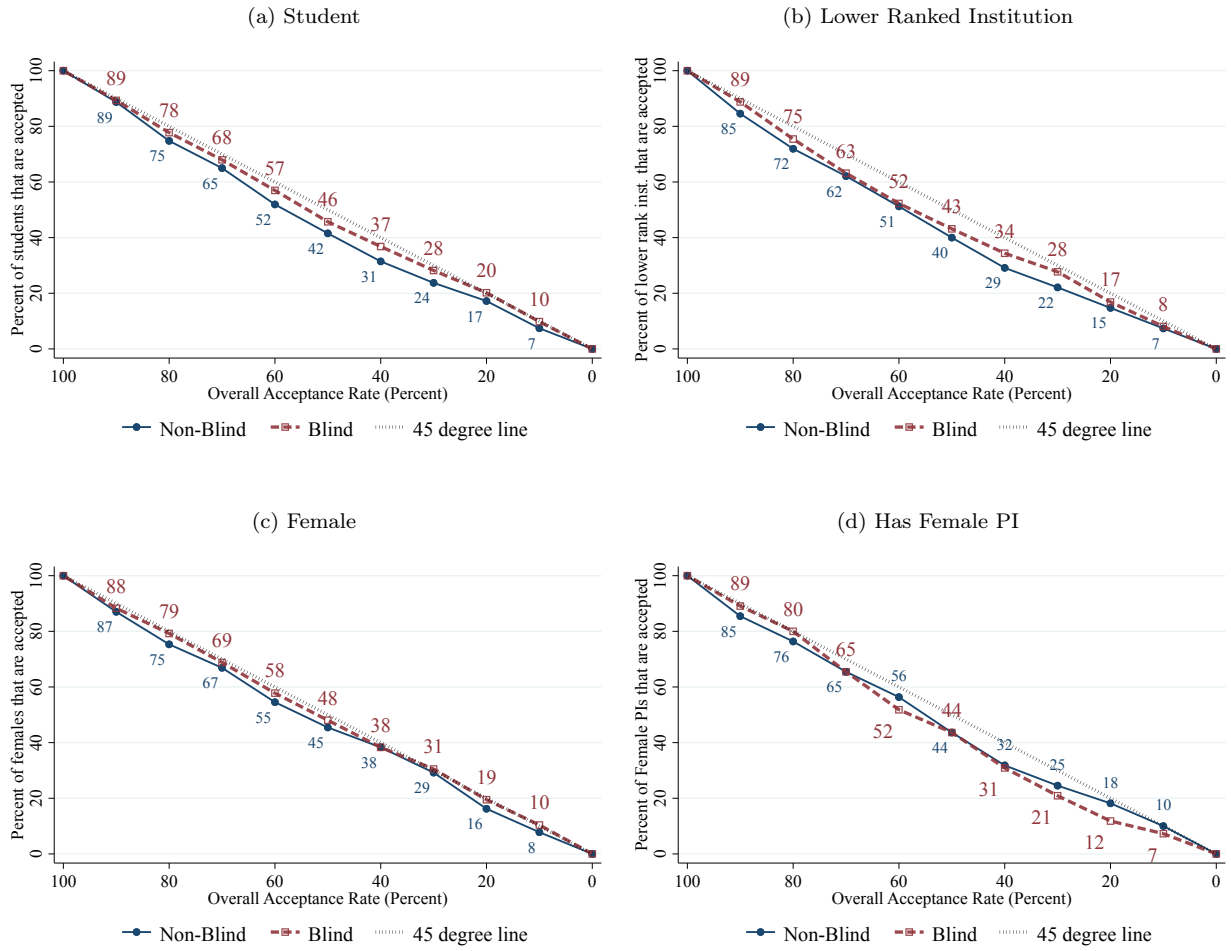
Notes. This table shows which papers were removed in the robustness checks in Table A16. The dependent variables are: (columns 1-3) an indicator for whether the reviewer gave a guess for the author of the given paper, (columns 4-6) whether the reviewer's guess for the paper was correct among those that gave guesses,(columns 5-6) whether the reviewer gave a guess and it was correct. Observations are at the paper-reviewer level. First four rows correspond to variables about the submission, while the last three correspond to characteristics of the reviewer. Heteroskedastic robust standard errors in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

C.10 Effects of Blinding on Acceptances

Figure A9 shows differences in acceptance rate gaps using predicted acceptance outcomes for various overall acceptance rates. Ultimately, comparing across acceptance thresholds 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,[37] 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, although I cannot rule it out. Neither applicants nor reviewers in this experiment were aware of the overall acceptance rate, 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 counterfactuals 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.

[37] 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).

Figure A9: Effects of Blinding on Composition, by Overall Acceptance Rate



Notes. These figures present 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 predicted 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 applicants who are affiliated with an institution that is not top 20 ranked, 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 predicted 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 predicted using a paper's reviewer-residualized average blind and average non-blind score.

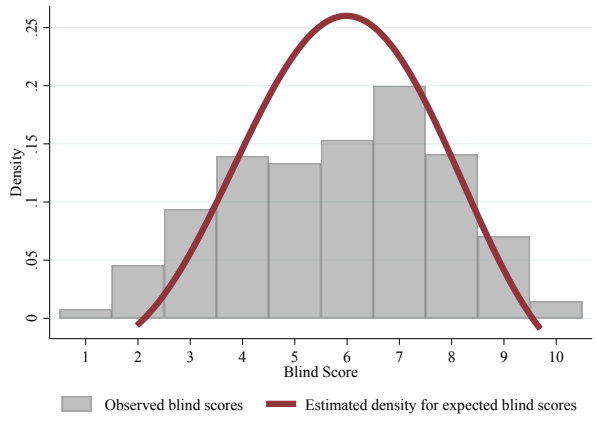
C.11 Deconvolving the Distribution of Reviewer Scores

Section 3.2 presents differences in a submission's acceptance outcome across blind and non-blind scores in the main experiment. These differences reflect both the effect of blinding on scores as well as reviewer-specific idiosyncracies. As captured by equation 6, observed blind reviewer scores are the sum of the blind score that the paper receives from the population of potential reviewers, and a paper-reviewer specific component, $u^B_{p,r}$. This section asks to what extent the differences in acceptances outcomes are explained by systematic changes in score allocation due to blinding. Because reviewers are randomly assigned to papers conditional on subfield, and I randomly assigned 2 blind reviewers for each paper, the distribution of expected blind scores can be estimated following traditional deconvolution approaches (Kotlarski, 1967; Rao, 1992).

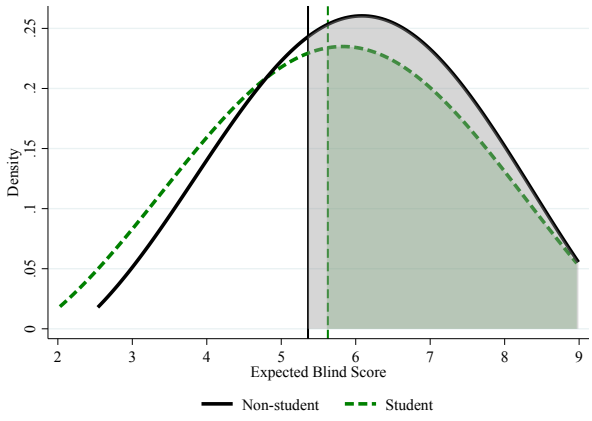
Figure A10a shows the distribution of observed blind scores, along with the deconvolved density (I first residualize out subfield from blind scores in order to account for the fact that reviewers were randomly assigned conditional on subfield). The distribution of observed blind scores as illustrated by the histogram is more dispersed than the deconvolved density, which is expected given that observed blind scores should contain more noise than a paper's expected blind score due to the paper-reviewer specific component. Figure A10b shows the deconvolved distributions estimated separately by student status.

Figure A10: Deconvolution Estimates of Reviewer Scores

(a) Blind Scores for the Full Sample



(b) Blind Scores By Student Status



Notes. This figure shows (a) the empirical distribution of blind scores at the paper-reviewer level, and the deconvolved density of expected blind scores that is estimated using the Kotlarski STATA package of Kato et al. (2021) and the (b) deconvolved distribution by student status. Vertical lines in (b) correspond to the blind score threshold for (solid line) non-students and (dashed line) students that is implied by a 60% overall acceptance rate with the same non-blind score threshold for both groups. Shaded areas therefore correspond to the applicants of each group that are accepted under a 60% overall acceptance rate.

The main text focuses on the impacts of blinding using observed blind and non-blind scores (Section 3.2). Another question is whether blinding changes acceptance outcomes when considering the score that papers receive in expectation over the population of potential reviewers.

If acceptances were based on a paper's expected blind score, then the cutoff threshold score corresponding to a 60% overall acceptance rate is around 5.5 (in other words, $F^B(5.5) \approx 0.60$, where $F^B(\cdot)$ is the CDF of expected blind scores).

Using the deconvolved distribution of expected blind scores and the estimated non-blind scoring equation (Table 5a), the implied distributions of expected non-blind scores can be computed for each applicant trait. For simplicity, focus on student status. The coefficient on the student indicator in Table 5a reflects the difference in non-blind score for any two submissions with the same expected blind score.

If acceptances were based on a paper's expected non-blind score, then the non-blind cutoff score is determined by the overall acceptance rate, the fraction of applicants who are students (0.51; Table 1), and the densities of expected blind scores by student status (Figure A10b). In other words, for a given overall acceptance rate R the expected non-blind acceptance threshold S^{NB*} and its corresponding expected blind thresholds S^{B*}_{non} for non-students and $S^{B*}_{student}$ for students is defined by:

$$R = p_{student} \times (1 - F^{B}_{student}(S^{B*}_{student})) + p_{non} \times (1 - F^{B}_{non}(S^{B*}_{non})) \tag{A1}$$

$$S^{NB*} = S^{NB}(student, S^{B*}_{student}) \tag{A2}$$

$$S^{NB*} = S^{NB}(non, S^{B*}_{non}) \tag{A3}$$

where $p_{student}$ is the proportion of students, $F^{B}_{x}(\cdot)$ is the cumulative density function of the deconvolved density of blind scores for group x, S^{B*}_{x} is the blind score threshold for group x corresponding to the expected non-blind score threshold S^{NB*}, and $S^{NB}(\cdot,\cdot)$ is the non-blind scoring equation. While this focuses on the student dimension for simplicity, this can be extended to more groups.

I find that the non-blind cutoff score for a 60% overall acceptance rate corresponds to an expected blind score of 5.6 for students and 5.3 for non-students. Note that both students and non-students are subject to the same non-blind cutoff score, and these translate to different blind scores given the non-blind scoring equation. More specifically, under the estimates of Table 5a, the cutoff expected blind score for non-students equals the cutoff expected blind score for students minus the student coefficient in the scoring equation divided by the coefficient on the expected blind score: $5.3 \approx 5.6 - 0.28/1.05$. Vertical lines in Figure A10b illustrate the thresholds, and shaded areas the density of applicants who are accepted. The threshold for students is higher, reflecting the fact that holding submission content (as proxied by blind score) constant, students receive a lower non-blind score than non-students.

Modeling acceptances in this way implies that accepting 60% of submissions using expected blind scores leads to around 56% of student papers and 63% of non-student papers accepted. In contrast, using expected non-blind scores leads to 54% of student papers and 66% of non-student papers accepted. As in prior results, fewer student papers are accepted under non-blind review than blind review. Blind review increases student acceptances and reduces non-student acceptances. More specifically, of the 60% of papers that are accepted using expected non-blind scores, 3 percentage points (5%) of the (non-student) papers' decisions are flipped by blinding. Of the 40% of papers that are not accepted using expected non-blind scores, 2.4 percentage points (5%) of (student) papers' decisions are flipped by blinding.

C.12 Comparing Effect Sizes

In Section 3.2, I find that under non-blind review, student applicants are 31% less likely to be accepted than non-students, applicants from lower-ranked institutions are 38% less likely to be accepted than those from top 20 ranked institutions, and female applicants are 9% less likely than male applicants. Relating these effect sizes to the literature on correspondence style studies, Bertrand and Mullainathan (2004) document that applicants with White names receive around 50% (3.2 percentage points from a baseline for Black applicants of 6.45) more employer callbacks than Black names. Agan and Starr (2018) document a racial callback gap prior to Ban-the-Box which removed criminal history information from job applications of around 7% (0.8 percentage points from a baseline for Black applicants of 13.1%).

In terms of policy effects, under the realized overall acceptance rate, I find that blinding nearly closes the student acceptance rate gap, while the gap by institution rank persists at around 34%. These effects are larger than past estimates on the effects of increasing gender diversity in evaluation committees on gender gaps (Bagues et al., 2017). Agan and Starr (2018) find that Ban-the-Box increases the racial callback gap to around 42% (0.8 to 4.4 percentage points).

C.13 Citations and Publication Status Five Years Later

Table A20: Reviewer Scores and Paper Quality

	Panel A: Percentile in Citations						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Trait-Predicted Percentile	0.13***			0.07*	0.09**		
	(0.04)			(0.04)	(0.04)		
Non-Blind Percentile		0.26***		0.25***		0.19***	0.19***
		(0.04)		(0.04)		(0.04)	(0.04)
Blind Percentile			0.26***		0.24***	0.18***	0.18***
			(0.04)		(0.04)	(0.04)	(0.04)
p-val H0: Reviewer = Predicted				0.00	0.01		
p-val H0: Blind = Non-Blind						0.89	0.95
Subfield FE							×
N	657	657	657	657	657	657	657
R^2	0.02	0.07	0.07	0.08	0.08	0.10	0.12
	Panel B: Percentile in Weighted Publication						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Trait-Predicted Percentile	0.10***			0.05	0.06*		
	(0.04)			(0.04)	(0.04)		
Non-Blind Percentile		0.18***		0.17***		0.11***	0.11***
		(0.04)		(0.04)		(0.04)	(0.04)
Blind Percentile			0.22***		0.21***	0.17***	0.17***
			(0.04)		(0.04)	(0.04)	(0.04)
p-val H0: Reviewer = Predicted				0.05	0.01		
p-val H0: Blind = Non-Blind						0.35	0.37
Subfield FE							×
N	657	657	657	657	657	657	657
R^2	0.01	0.04	0.05	0.04	0.06	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. Columns 2 and 3 correspond to Figures 2a and 2b. 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. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A21: 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	0.22***	0.22***	0.22***	0.16***	0.18***	0.16***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Within-Reviewer Percentile × Blind			0.00			0.02
			(0.04)			(0.04)
N	1289	1302	2591	1289	1302	2591
R^2	0.05	0.05	0.05	0.03	0.04	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. * $p < .1$, ** $p < .05$, *** $p < .01$

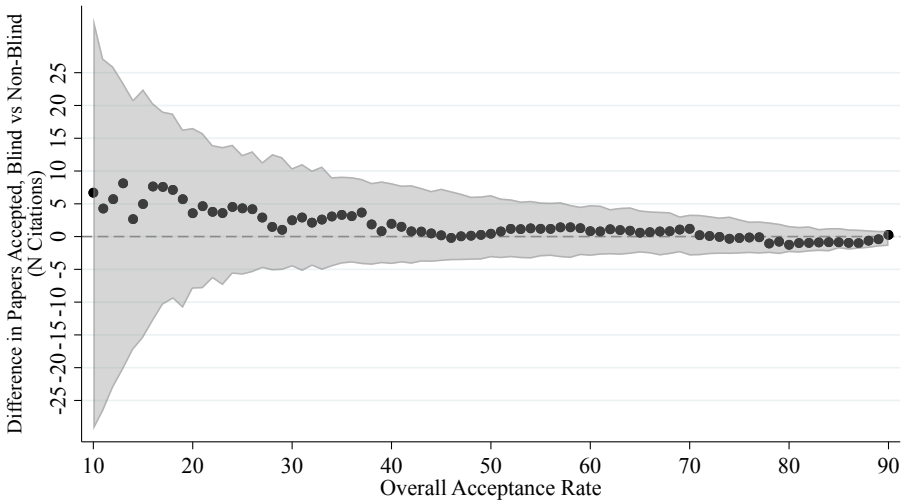
Table A22: Score and Paper Quality Rankings (Robustness Subsample)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Trait-Predicted Percentile	0.18***			0.09*	0.11**		
	(0.05)			(0.05)	(0.05)		
Non-Blind Percentile		0.33***		0.30***		0.22***	0.22***
		(0.04)		(0.04)		(0.06)	(0.06)
Blind Percentile			0.31***		0.29***	0.15**	0.14**
			(0.04)		(0.05)	(0.06)	(0.06)
p-val H0: Reviewer = Predicted				0.00	0.01		
p-val H0: Blind = Non-Blind						0.57	0.52
Subfield FE							×
N	417	417	417	417	417	417	417
R^2	0.03	0.12	0.11	0.13	0.12	0.14	0.17

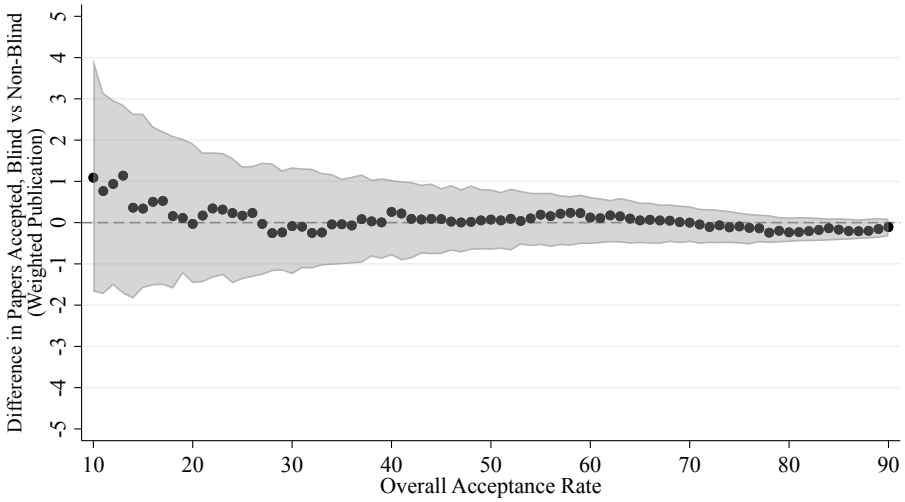
Notes. This table shows the predictive power of author traits, blind score percentile ranks, and non-blind score percentile ranks for paper quality, measured by citations. Trait-predicted percentiles are created by the predicted values from regressing paper citations on author traits (applicant student 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. * $p < .1$, ** $p < .05$, *** $p < .01$

Figure A11: Effects of Blinding on Quality

(a) Number of Citations



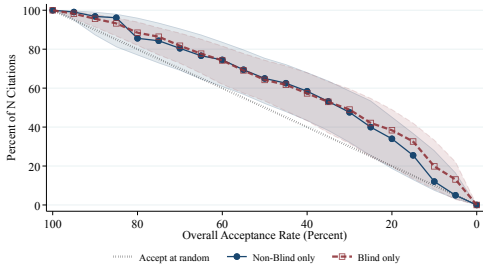
(b) Journal-Weighted Publication Status



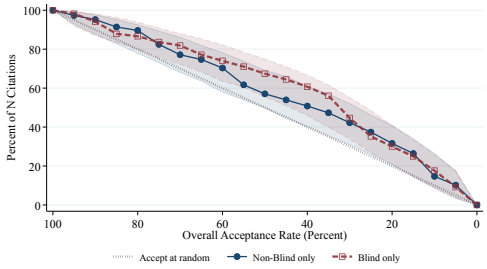
Notes. These figures present the impacts of blinding on the average quality of papers that the conference selects, where paper quality is measured by (a) number of citations (b) journal-weighted publication statuses. Each dot in a figure reflects the mean difference in paper quality across papers accepted under blind scores, and those accepted under non-blind scores. Observations are at the paper level. All acceptance outcomes are predicted 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 predicted using a paper's reviewer-residualized average blind and average non-blind score. Shaded areas correspond to 95% confidence intervals, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Figure A12: Effects of Blinding on Paper Quality by trait

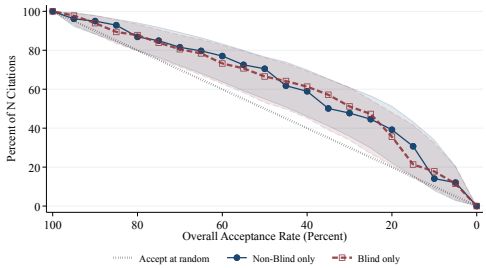
(a) Non-student



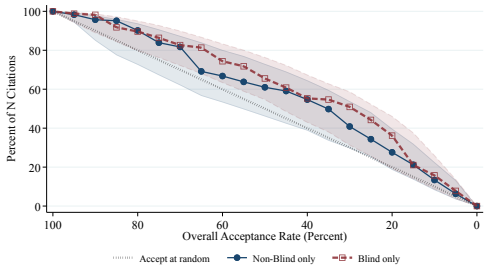
(b) Student



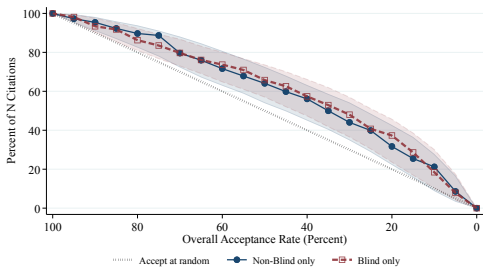
(c) Top 20 Ranked Institution



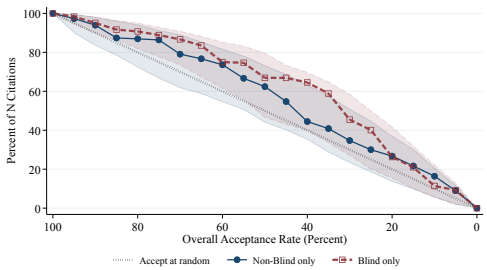
(d) Lower Ranked Institution



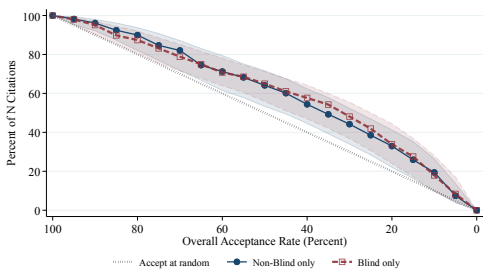
(e) Male Applicant



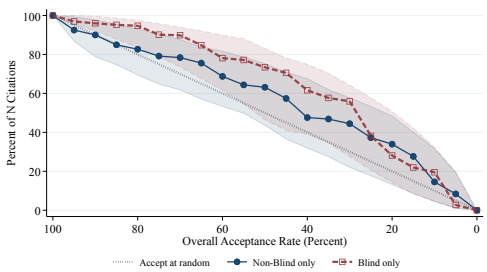
(f) Female Applicant



(g) Male PI

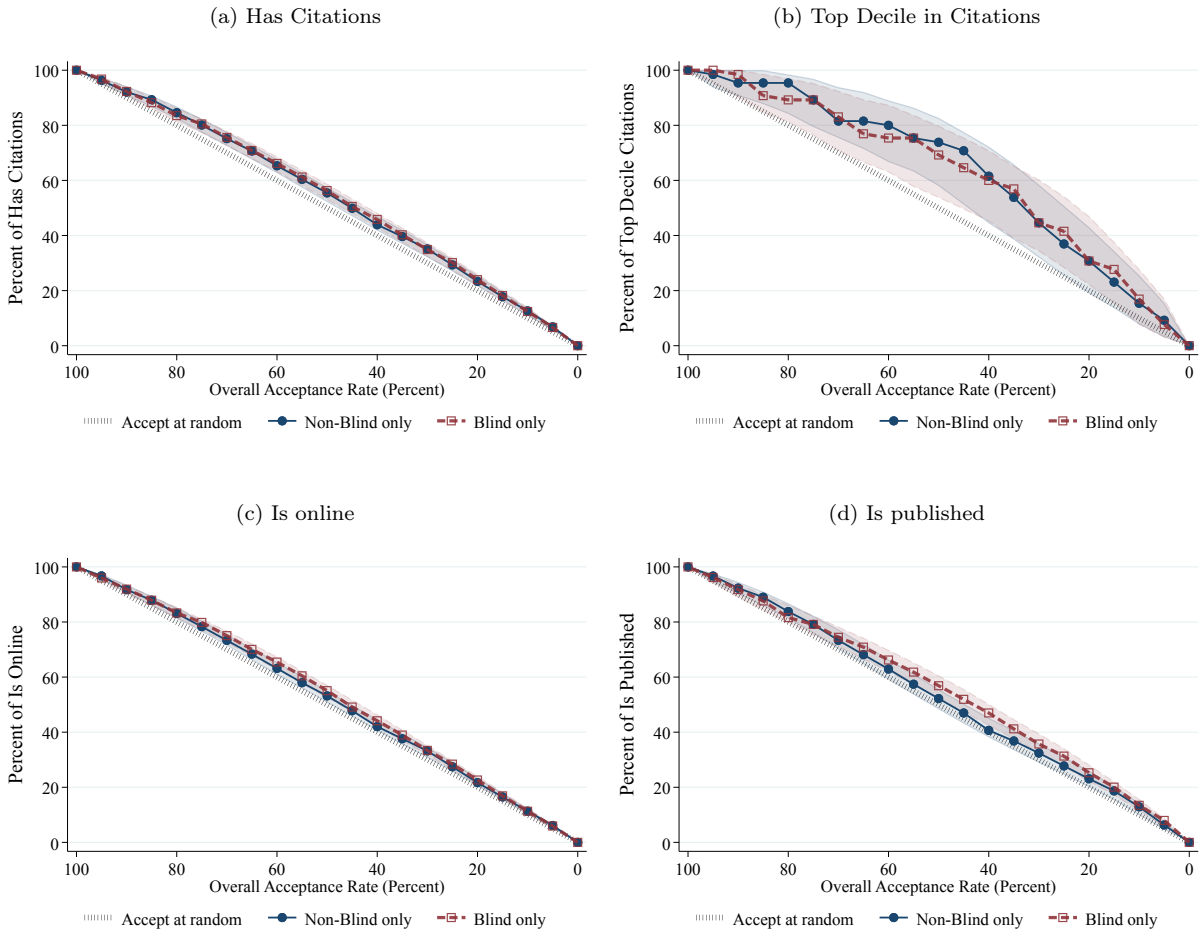


(h) Female PI



Notes. These figures illustrate the share of citations, for various overall acceptance rates, as in Figure 2, but subsetting by applicant traits. Paper ranks are calculated within the subset. For instance, (a) shows, among non-students, the share of 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). The red dashed line with hollow squares represents the shares when acceptance outcomes are determined by blind scores only, and the blue solid line with solid circles for when outcomes are determined by non-blind scores only. Paper quality measures are collected 5 years after the experiment. 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. Shaded areas correspond to 95% confidence intervals, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Figure A13: Effects of Blinding on Paper Quality

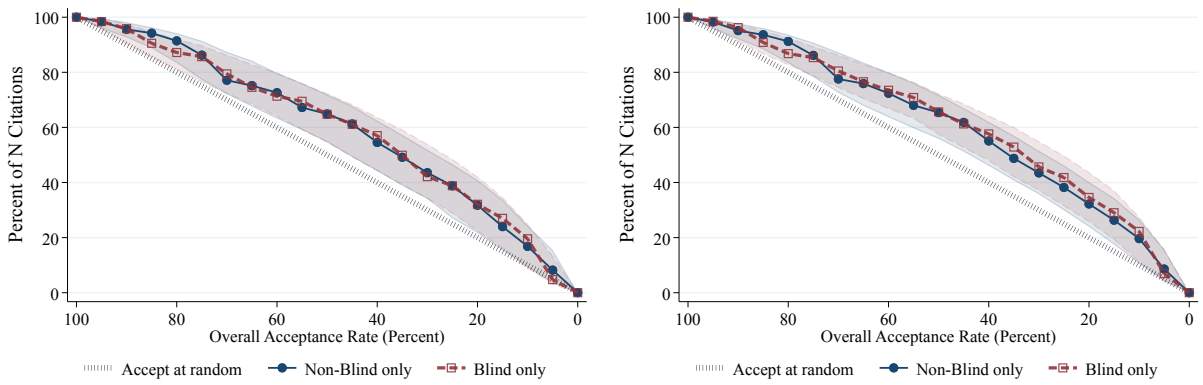


Notes. These figures illustrate the share of (a) non-zero citations, (b) citations in the top decile of the sample, (c) online statuses (paper is available online) or (d) publication statuses that are attributable to accepted papers, for various overall acceptance rates, as in Figure 2. For instance, (a) shows the share of papers with non-zero 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). The red dashed line with hollow squares represents the shares when acceptance outcomes are determined by blind scores only, and the blue solid line with solid circles for when outcomes are determined by non-blind scores only. Paper quality measures are collected 5 years after the experiment. 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. Shaded areas correspond to 95% confidence intervals, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Figure A14: Effects of Blinding on Paper Quality

(a) Number of citations 2 years later

(b) Number of citations 4 years later



Notes. These figures illustrate the share of citations (a) 2 years later or (b) 4 years later, for various overall acceptance rates, as in Figure 2. For instance, (a) shows the share of 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). The red dashed line with hollow squares represents the shares when acceptance outcomes are determined by blind scores only, and the blue solid line with solid circles 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. Shaded areas correspond to 95% confidence intervals, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A23: Disparities in Paper Outcomes 2, 4, and 5 Years Later

	Panel A: Unconditional differences											
	Has Citations			N Citations			Is Online			Is Published		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	2 year	4 year	5 year	2 year	4 year	5 year	2 year	4 year	5 year	2 year	4 year	5 year
Student	0.015	0.052	0.053	-0.900	-2.472	-3.247	0.029	0.043	0.042	0.019	0.026	0.057
	(0.039)	(0.038)	(0.037)	(1.014)	(3.014)	(4.018)	(0.040)	(0.033)	(0.033)	(0.039)	(0.040)	(0.040)
Lower Rank Inst.	-0.114***	-0.100**	-0.097**	-3.084***	-8.826**	-11.688**	-0.060	-0.100***	-0.090***	-0.007	-0.043	-0.044
	(0.043)	(0.040)	(0.039)	(1.160)	(3.564)	(4.790)	(0.042)	(0.035)	(0.034)	(0.042)	(0.043)	(0.043)
Female	-0.069	-0.029	-0.051	-3.261***	-9.396***	-11.868***	-0.059	-0.040	-0.024	-0.051	-0.053	-0.067
	(0.046)	(0.045)	(0.044)	(0.936)	(2.556)	(3.374)	(0.047)	(0.042)	(0.041)	(0.045)	(0.047)	(0.048)
Has Female PI	-0.096*	-0.042	-0.025	-1.078	-3.563	-3.957	-0.028	-0.018	0.003	-0.054	-0.004	0.015
	(0.052)	(0.050)	(0.048)	(1.270)	(3.308)	(4.293)	(0.052)	(0.044)	(0.043)	(0.051)	(0.053)	(0.053)
Subfield FE	×	×	×	×	×	×	×	×	×	×	×	×
N	657	657	657	657	657	657	657	657	657	657	657	657
R^2	0.05	0.03	0.03	0.05	0.06	0.05	0.02	0.03	0.03	0.03	0.03	0.02
	Panel B: Differences conditional on Blind score											
	Has Citations			N Citations			Is Online			Is Published		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	2 year	4 year	5 year	2 year	4 year	5 year	2 year	4 year	5 year	2 year	4 year	5 year
Student	0.042	0.082**	0.088**	-0.409	-0.517	-0.492	0.058	0.073**	0.072**	0.040	0.049	0.085**
	(0.042)	(0.040)	(0.039)	(1.003)	(2.977)	(3.993)	(0.041)	(0.035)	(0.035)	(0.040)	(0.042)	(0.042)
Lower Rank Inst.	-0.026	-0.004	0.007	-1.494	-2.089	-2.332	0.009	-0.018	-0.009	0.058	0.035	0.043
	(0.050)	(0.048)	(0.047)	(1.267)	(3.796)	(5.064)	(0.048)	(0.043)	(0.042)	(0.048)	(0.049)	(0.050)
Female	-0.066	-0.036	-0.056	-3.262***	-9.143***	-11.391***	-0.056	-0.048	-0.031	-0.051	-0.064	-0.076
	(0.047)	(0.047)	(0.047)	(0.972)	(2.839)	(3.775)	(0.048)	(0.044)	(0.043)	(0.045)	(0.050)	(0.051)
Has Female PI	-0.035	0.014	0.037	0.129	0.909	2.205	0.029	0.035	0.056	-0.007	0.038	0.065
	(0.054)	(0.051)	(0.050)	(1.357)	(3.558)	(4.629)	(0.054)	(0.042)	(0.042)	(0.054)	(0.056)	(0.056)
Submission content	0.174***	0.189***	0.199***	3.265***	13.326***	18.433***	0.135***	0.159***	0.156***	0.131***	0.155***	0.175***
	(0.044)	(0.044)	(0.045)	(1.148)	(4.151)	(5.710)	(0.041)	(0.040)	(0.040)	(0.040)	(0.044)	(0.045)
Subfield FE	×	×	×	×	×	×	×	×	×	×	×	×
N Papers	645	645	645	645	645	645	645	645	645	645	645	645

Notes. This table presents disparities in realized paper quality 2, 4, and 5 years after the experiment (a) unconditional on submission content and (b) conditional on submission content as in panel A of Table 6. For (b), submission content is proxied by a paper's blind scores: to account for noise in blind scores, a given blind review score for a paper is instrumented by the average score that the paper's other blind reviewers gave, as described in Section 5.1. Standard errors are clustered at the paper-level. * $p < .1$, ** $p < .05$, *** $p < .01$

C.14 Effects of Acceptances on Paper and Author Outcomes

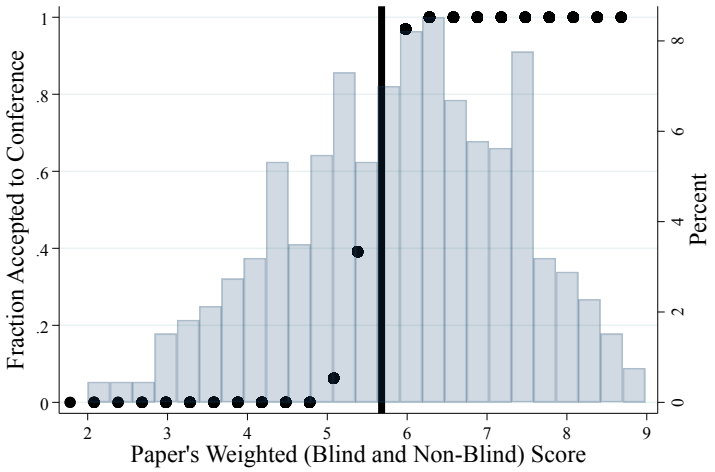
To identify the acceptance score cutoff, I follow past work (e.g. Altmejd et al., 2021; Brunner et al., 2023; Mountjoy, 2024) and estimate local linear regression discontinuities centered at each potential score cutoff, defining the cutoff as the score value that produces the largest discontinuity in acceptance. Porter and Yu (2015) discuss the statistical properties of this procedure.

Using the cutoff, I conduct diagnostic tests for the regression discontinuity. First, the running variable (a paper's average reviewer score) and the acceptance score cutoffs generate a significant discontinuity in acceptance outcomes (Figure A15). Second, the densities of the running variables do not exhibit significant discontinuities in density at the cutoff: I fail to reject a test of manipulation using local polynomial density estimation (Cattaneo et al., 2020) against the null hypothesis of equality at the cutoff ($p = 0.42$). Finally, Table A24 tests for covariate balance at the cutoff using pre-determined characteristics as outcomes. I additionally construct each submission's covariate-predicted citation count by regressing a submission's citation count 5 years later on author traits and subfield fixed effects: I include this as an outcome in the balance test and find that this does not display a significant discontinuty at the cutoff, further supporting the conclusion that pre-determined covariates were balanced at the cutoff.

Figure A16 illustrates the RD estimate for the number of citations and Table A25 shows the estimates for the rest of the outcomes. In general, I find that the estimates are noisy: point estimates suggest that acceptance increases the likelihood that the applicant remains in academia 5 years later by 6 percentage points from a baseline of 72 percent, but this is not statistically significant and coupled with negative (and statistically insignificant) estimates for the effects on whether the applicant is a professor or applies again to the conference 5 years later. I find some evidence that acceptance significantly increases the likelihood that the PI applies to the conference 5 years later (Table A25).

Point estimates suggest that students benefit from acceptance on the extensive margin of their submission having at least one citation 5 years later, while non-students benefit on the intensive margin, but neither of these estimates are not statistically significant (Table A26).

Figure A15: Realized Acceptance Status



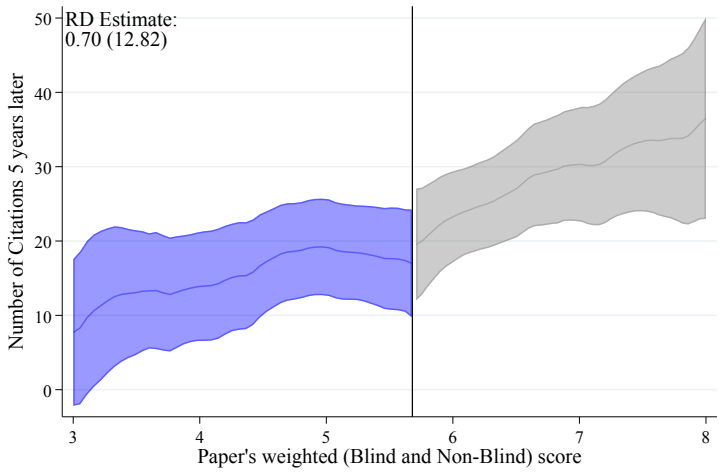
Notes. This figure shows the fraction of submissions that were accepted to the conference in 2020, given the paper's weighted-average reviewer score (constructed by summing 2/3 of the paper's average blind score and 1/3 of the paper's average non-blind score). The vertical line represents the cutoff score.

Table A24: RD Covariate Balance

	Student	Lower Rank Inst	Female	PI Female	Predicted Cites
RD Estimate	0.10	-0.28	0.22	0.14	-0.95
	(0.36)	(0.28)	(0.29)	(0.28)	(4.71)
Robust 95% CI	[-0.70,0.93]	[-1.08,0.23]	[-0.32,0.99]	[-0.39,0.89]	[-13.76,8.49]
Robust p-value	0.78	0.21	0.31	0.44	0.64
N	657	657	657	657	657

Notes. This table shows the effects of conference acceptance on author and paper outcomes for the sample of submissions from the main experiment, conditional on subfield, estimated using a regression discontinuity estimated following the bias-corrected approach of Calonico et al. (2017). Column names correspond to the dependent variable. Observations are at the paper level.

Figure A16: Effects of Conference Acceptance on Citations 5 Years Later



Notes. This figure shows a paper's citations 5 years later, 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 cutoff for acceptance. Those to the right of the cutoff were accepted to the conference and those to the left were not.

Table A25: Impacts of conference acceptance on paper and author outcomes 5 years later

	Panel A: Paper Outcomes			
	Has Cites	N Cites	Is Online	Published
RD Estimate	0.02	0.70	-0.01	-0.10
	(0.31)	(12.82)	(0.29)	(0.27)
Robust 95% CI	[-0.69,0.69]	[-33.12,28.50]	[-0.68,0.63]	[-0.77,0.51]
Robust p-value	1.00	0.88	0.94	0.70
N	657	657	657	657
	Panel B: Author Outcomes			
	In Academia	Is Professor	Applied	PI applied
RD Estimate	0.06	-0.09	-0.15	0.57
	(0.28)	(0.27)	(0.24)	(0.35)
Robust 95% CI	[-0.59,0.68]	[-0.78,0.54]	[-0.85,0.30]	[-0.02,1.55]
Robust p-value	0.89	0.72	0.35	0.06
N	657	657	657	657

Notes. This table shows the effects of conference acceptance on author and paper outcomes for the sample of submissions from the main experiment, conditional on subfield, estimated using a regression discontinuity estimated following the bias-corrected approach of Calonico et al. (2017). Observations are at the paper level.

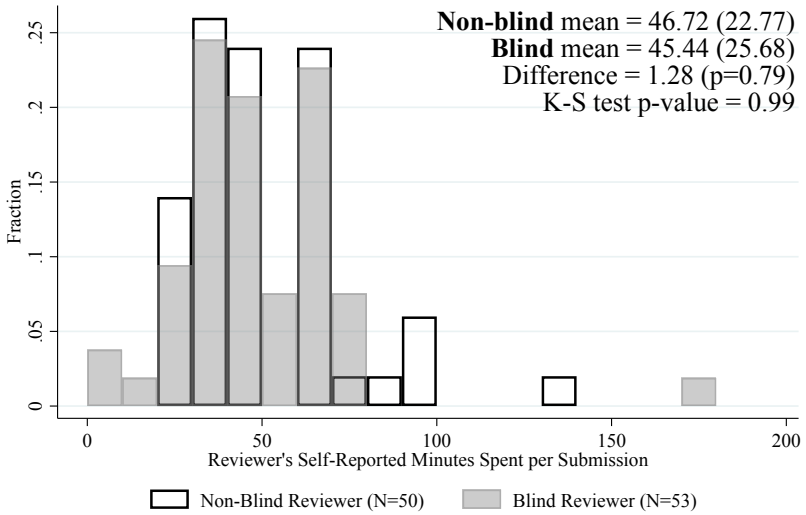
Table A26: Impacts of conference acceptance on citations by trait

	Panel A: Has at Least One Citation							
	Student Status		Institution Rank		Gender		PI Gender	
	Student	Non-Student	Top 20	20+	Female	Male	Female PI	Male PI
RD Estimate	0.15	0.07	-0.71	0.48	4.50	-0.14	0.23	0.05
	(0.76)	(0.24)	(0.83)	(0.37)	(7.34)	(0.33)	(0.62)	(0.30)
Robust 95% CI	[-1.57,1.88]	[-0.48,0.63]	[-3.10,0.93]	[-0.28,1.37]	[-11.08,23.62]	[-0.91,0.56]	[-0.91,2.35]	[-0.71,0.73]
Robust p-value	0.86	0.79	0.29	0.20	0.48	0.64	0.39	0.97
N	337	320	285	267	154	503	110	546
	Panel B: Number of Citations							
	Student	Non-Student	Top 20	20+	Female	Male	Female PI	Male PI
RD Estimate	-3.92	0.13	-38.02	19.30	29.50	5.83	5.60	5.06
	(40.86)	(13.84)	(35.58)	(17.65)	(49.91)	(13.64)	(40.06)	(12.98)
Robust 95% CI	[-112.00,85.72]	[-35.51,30.70]	[-130.45,35.61]	[-25.25,59.80]	[-81.76,150.60]	[-27.84,37.04]	[-91.05,114.90]	[-28.24,34.74]
Robust p-value	0.79	0.89	0.26	0.43	0.56	0.78	0.82	0.84
N	337	320	285	267	154	503	110	546

Notes. This table shows the effects of conference acceptance on author and paper outcomes for the sample of submissions from the main experiment, conditional on subfield, estimated using a regression discontinuity estimated following the bias-corrected approach of Calonico et al. (2017). Dependent variable is (a) an indicator for whether the submission has at least one citation 5 years later, and (b) the number of citations associated with the submission 5 years later. Observations are at the paper level.

C.15 Time Spent by Reviewers

Figure A17: Reviewer time spent



Notes. This figure shows the distribution of reviewer responses to the optional question that was asked by the committee, "How many minutes did you spend per abstract?" 103 out of the 245 reviewers responded: 50 non-blind, 53 blind. The p-value associated with "Difference" corresponds to the p-value from a t-test comparison of means. The "K-S test p-value" corresponds to a p-value from a two-sample Kolmogorov-Smirnov equality-of-distributions test.

C.16 Reviewer Engagement in Future Years

Table A27: Effects of Blinding on Reviewer Engagement with the Conference

	2 years later	3 years later	4 years later	5 years later
	(1)	(2)	(3)	(4)
Blind	0.03	-0.04	0.05	0.02
	(0.06)	(0.06)	(0.06)	(0.06)
Outcome mean	0.63	0.59	0.52	0.49
N	245	245	245	245
R^2	0.00	0.00	0.00	0.00

Notes. This table shows the effects of blinding on the reviewer's likelihood of participating in the conference (as a reviewer, applicant, or PI) in years following the experiment. Dependent variable is an indicator for whether the reviewer participated in the conference in the given year. Observation is at the reviewer level. Heteroskedastic robust standard errors in parentheses.

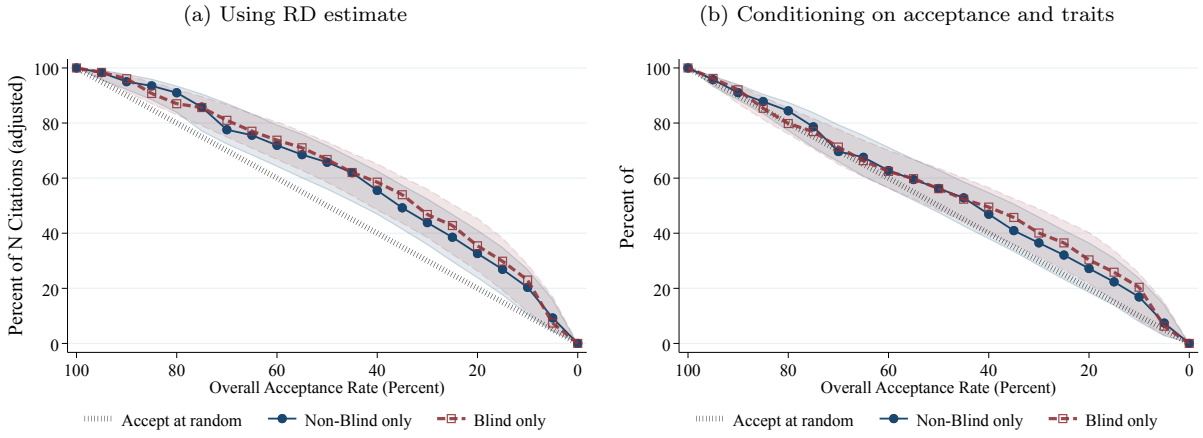
C.17 Robustness to Effects of Conference Acceptance

In this section, I explore whether conclusions change when accounting for potential impacts of conference acceptance on citation outcomes. Intuitively, the concern is that if acceptances to the conference increase future citations, then true paper quality may be mechanically overestimated for the set of papers accepted relative to the set of papers that were not. I therefore decrease the citation count of papers accepted to the conference by the point estimate for the effect on citations that estimated in Table A25 (I maintain a floor of zero citations), and re-produce the main analyses.

Figure A18a shows the analog of Figure 2, and implies that the finding that blinding does not worsen quality is not sensitive to adjusting for conference acceptance in this way. In Figure A18b, I do a related exercise of residualizing out an indicator for conference acceptance, author traits, and their interactions from citations.

Table A28 shows subgroup differences in quality using this adjusted outcome (analog to Panel A of Table 6 in the main text). As expected given that students, those from lower ranked institutions, and female applicants and PIs are less likely to be accepted (Table 3), the point estimates in column 1 are slightly less negative when adjusting for quality in this way. However, after conditioning on submission content and incorporating this into the model estimates (Table A29), I find that the conclusions remain the same as in the main text (Table A43).

Figure A18: Effects of Blinding on Quality (adjusting for conference acceptance)



Notes. These figures illustrate the share of citations that are attributable to accepted papers, as in Figure 2, but adjusting for potential effects of conference acceptance on citations by (a) decreasing the citation count of papers accepted to the conference by the estimated the effect on citations presented in Table A25, and (b) residualizing out an indicator for conference acceptance, author traits, and their interactions from citations. 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. Shaded areas correspond to 95% confidence intervals, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A28: Disparities in Realized Paper Quality (adjusting for conference acceptance)

	N Citations	
	(1)	(2)
Student	-3.23	-0.51
	(4.09)	(3.98)
Lower Rank Inst.	-11.91**	-2.44
	(4.85)	(5.04)
Female	-12.25***	-11.41***
	(3.42)	(3.75)
Has Female PI	-3.52	2.07
	(4.41)	(4.61)
Submission Content		17.97***
		(5.67)
Subfield FE	×	×
Outcome mean	25.35	25.35
First stage F-stat	-	26.59
N Papers	645	645

Notes. This table presents disparities in paper quality outcomes that are realized 5 years after the experiment as in Panel A of Table 6, but adjusting for potential effects of conference acceptance on citations by decreasing the citation count of papers accepted to the conference by the upper bound of the effect on citations estimated in Table A25. Observations for odd-numbered columns are at the paper-level, with heteroskedastic robust standard errors in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A29: Decomposing Non-Blind Score Gaps (adjusting for conference acceptance)

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	-0.28	-0.19	-0.11	0.13
	[-0.51,0.002]	[-0.44,0.23]	[-0.39,0.23]	[-0.18,0.57]
	Panel B: Decomposition components			
Submission content	-0.17	-0.54	-0.05	-0.32
	[-0.44,0.05]	[-1.02,-0.30]	[-0.37,0.23]	[-0.73,-0.05]
Accurate statistical discrimination	0.08	0.000	-0.18	0.09
	[-0.05,0.26]	[-0.14,0.21]	[-0.41,0.02]	[-0.04,0.31]
N Citations	-0.005	-0.02	-0.11	0.02
	[-0.09,0.08]	[-0.13,0.10]	[-0.28,0.02]	[-0.06,0.15]
Online	0.03	-0.004	-0.01	0.03
	[-0.03,0.15]	[-0.05,0.06]	[-0.09,0.03]	[-0.03,0.13]
Published	0.05	0.03	-0.05	0.04
	[0.002,0.15]	[-0.02,0.12]	[-0.14,0.01]	[-0.01,0.14]
Inaccurate statistical discrimination	-0.13	0.06	0.16	-0.13
	[-0.34,0.04]	[-0.19,0.27]	[-0.04,0.40]	[-0.41,0.05]
N Citations	-0.04	0.04	0.10	-0.01
	[-0.17,0.04]	[-0.09,0.18]	[-0.02,0.26]	[-0.15,0.08]
Online	-0.04	0.03	0.02	-0.03
	[-0.18,0.04]	[-0.04,0.14]	[-0.05,0.11]	[-0.16,0.04]
Published	-0.04	-0.02	0.05	-0.08
	[-0.14,0.03]	[-0.14,0.07]	[-0.04,0.16]	[-0.24,0.01]
Alternative objectives	-0.33	0.25	-0.12	0.34
	[-0.59,-0.07]	[-0.07,0.58]	[-0.41,0.14]	[0.08,0.67]
Talk quality	-0.22	0.09	-0.12	0.24
	[-0.39,-0.06]	[-0.10,0.28]	[-0.30,0.02]	[0.09,0.46]
Author benefit	-0.03	0.02	0.01	0.003
	[-0.10,0.02]	[-0.03,0.09]	[-0.03,0.06]	[-0.05,0.06]
Conference benefit	-0.08	0.14	-0.02	0.09
	[-0.21,0.03]	[-0.002,0.31]	[-0.14,0.11]	[-0.02,0.26]
Unexplained	0.10	-0.49	0.03	-0.18
	[-0.30,0.54]	[-0.95,0.07]	[-0.39,0.55]	[-0.72,0.36]

Notes. This table decomposes disparities in non-blind scores as in Table A43, but adjusting for potential effects of conference acceptance on citations by decreasing the citation count of papers accepted to the conference by the upper bound of the effect on citations estimated in Table A25. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5) and subfield. Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

D Reviewer Comments for Authors

Each year during the review process, reviewers are asked to leave comments for the authors. The comments are generally short and allude to the evaluation criteria that the reviewers are asked to follow.

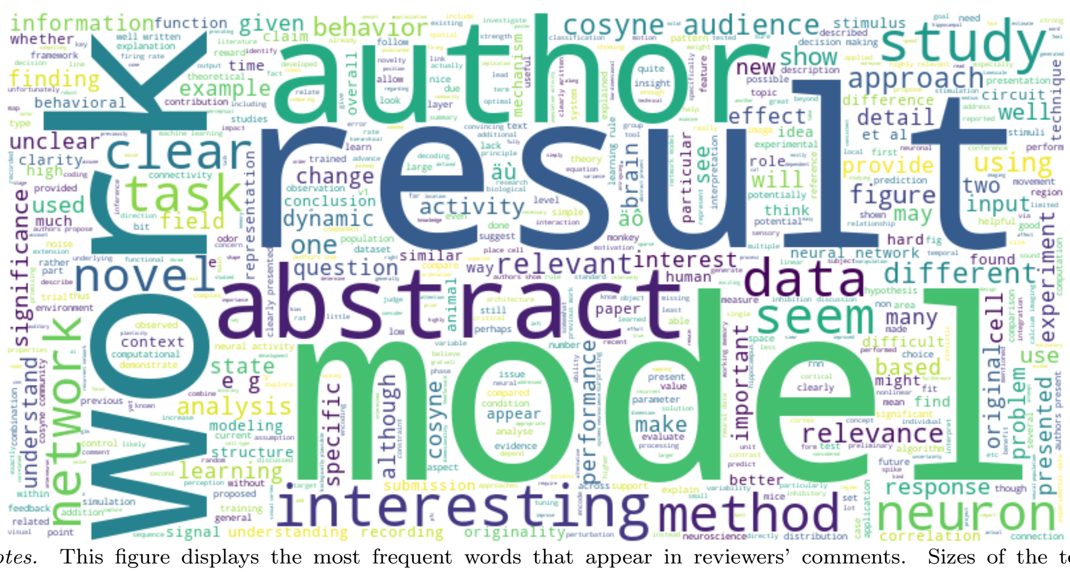
In the experimental year, all reviewer-submission pairs have a comment and the average comment contains a little over 100 words.[38] While the general structure and length of comments varies, reviewers often allude to four criteria that they are asked to consider while evaluating (the criteria are significance, originality, clarity, and relevance to audience, as described in Section 2). I use these comments to extract and better understand reviewers' beliefs of paper quality.

To enumerate the comments from the 2020 review process, each comment was assigned to at least two research assistants, who were shown only the comment and then asked to rate it according to each of these four statements:

1. Reviewer believes that study advances state of the field.
2. Reviewer believes that the concepts, approach and/or techniques are novel/new.
3. Reviewer believes that the addressed questions and the obtained results are clearly presented.
4. Reviewer believes that the submission is relevant to the conference audience.

For each, the available answers were: Strongly disagree, somewhat disagree, neither disagree nor agree, somewhat agree, strongly agree, or unclear. The last option was instructed to be chosen if the criteria was not mentioned.

Figure A19: Commonly Used Words in Comments



Notes. This figure displays the most frequent words that appear in reviewers' comments. Sizes of the text correspond to frequency counts.

[38] An example is: "No hypothesis is stated explicitly. The implicit hypothesis that some models can perform transfer learning while others cannot is not directly tested. No comparisons between models are reported. For example, the successful performance in four tasks is not compared with models that fail at this series of tasks. For future submissions, I encourage the authors to revise their work by proposing at least one hypothesis, testing it... analyzing and describing the results of their investigation."

Table A30: Reviewer Comment Sentiments

	Non-blind reviewers				Blind reviewers			
	Advance	Novel	Clear	Relevant	Advance	Novel	Clear	Relevant
All	3.82	3.90	3.32	3.93	3.75	3.82	3.29	3.93
	(0.99)	(0.96)	(1.18)	(1.02)	(1.00)	(1.00)	(1.22)	(1.01)
				By Applicant Student Status				
Student	3.74	3.84	3.23	3.84	3.72	3.82	3.30	3.92
	(1.02)	(0.99)	(1.19)	(1.05)	(1.00)	(1.00)	(1.23)	(1.02)
Not Student	3.91	3.95	3.41	4.02	3.77	3.81	3.28	3.94
	(0.96)	(0.92)	(1.17)	(0.98)	(1.00)	(1.00)	(1.21)	(1.00)
Difference	-0.17***	-0.10**	-0.18***	-0.18***	-0.05	0.01	0.02	-0.02
	[0.05]	[0.04]	[0.06]	[0.05]	[0.05]	[0.04]	[0.06]	[0.04]
				By Applicant Institution Rank				
Lower Ranked	3.75	3.83	3.21	3.85	3.67	3.74	3.17	3.87
	(0.99)	(0.98)	(1.19)	(1.02)	(1.01)	(1.03)	(1.25)	(1.04)
Top 20	3.87	3.95	3.39	3.98	3.81	3.87	3.37	3.98
	(0.99)	(0.94)	(1.17)	(1.01)	(0.98)	(0.98)	(1.19)	(0.98)
Difference	-0.16***	-0.13***	-0.23***	-0.13***	-0.16***	-0.17***	-0.27***	-0.13***
	[0.04]	[0.05]	[0.07]	[0.05]	[0.05]	[0.05]	[0.06]	[0.05]
				By Applicant Gender				
Female	3.80	3.86	3.41	3.89	3.73	3.80	3.27	3.88
	(1.00)	(0.96)	(1.19)	(1.07)	(0.96)	(0.99)	(1.19)	(1.02)
Male	3.83	3.91	3.29	3.94	3.75	3.82	3.29	3.95
	(0.99)	(0.96)	(1.18)	(1.00)	(1.01)	(1.01)	(1.23)	(1.00)
Difference	-0.03	-0.05	0.13*	-0.05	-0.02	-0.02	-0.03	-0.07
	[0.06]	[0.05]	[0.07]	[0.06]	[0.05]	[0.05]	[0.07]	[0.05]
				By PI Gender				
Female	3.76	3.90	3.35	3.88	3.66	3.83	3.23	3.87
	(1.00)	(0.96)	(1.18)	(1.00)	(1.03)	(1.02)	(1.22)	(1.06)
Male	3.83	3.89	3.31	3.94	3.77	3.81	3.30	3.95
	(0.99)	(0.96)	(1.19)	(1.02)	(0.99)	(1.00)	(1.22)	(1.00)
Difference	-0.08	0.01	0.04	-0.05	-0.11*	0.01	-0.07	-0.08
	[0.06]	[0.06]	[0.07]	[0.05]	[0.06]	[0.06]	[0.09]	[0.07]

Notes. This table shows summary statistics for reviewer comment sentiments. Observations are at the paper-reviewer-rater level, with observations weighted by the inverse number of total raters assigned to a given paper-reviewer, so that each paper-reviewer has equal weight. Each outcomet akes on a value of one through five, with one corresponding to "strongly disagree" and five as "strongly agree" by the rater. First four columns correspond to non-blind reviewers, and the subsequent four columns to blind reviewers. 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), with standard errors clustered at the reviewer level. P-values for t-test comparisons of means are represented by stars: * $p < .1$, ** $p < .05$, *** $p < .01$

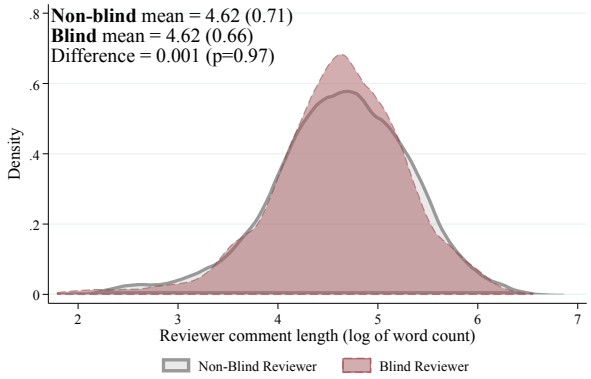
Table A31: Correlating Reviewer Comments with Scores

	(1)	(2)	(3)
Reviewer comment: Advances field	0.86***	0.67***	0.71***
	(0.06)	(0.06)	(0.08)
Reviewer comment: Novel	0.24***	0.27***	0.26***
	(0.05)	(0.05)	(0.08)
Reviewer comment: Clearly written	0.41***	0.43***	0.39***
	(0.03)	(0.04)	(0.06)
Reviewer comment: Relevant to audience	0.65***	0.54***	0.50***
	(0.05)	(0.05)	(0.07)
Reviewer comment: Advances field (not mentioned)	-0.13	-0.11	-0.08
	(0.20)	(0.20)	(0.28)
Reviewer comment: Novel (not mentioned)	0.26	0.04	0.14
	(0.17)	(0.17)	(0.24)
Reviewer comment: Clearly written (not mentioned)	0.80***	0.03	0.06
	(0.14)	(0.15)	(0.19)
Reviewer comment: Relevant to audience (not mentioned)	0.94***	0.25*	0.02
	(0.15)	(0.15)	(0.20)
Reviewer comment: Advances field × Blind			-0.08
			(0.11)
Reviewer comment: Novel × Blind			0.03
			(0.11)
Reviewer comment: Clearly written × Blind			0.07
			(0.07)
Reviewer comment: Relevant to audience × Blind			0.10
			(0.11)
Reviewer comment: Advances field (not mentioned) × Blind			-0.09
			(0.40)
Reviewer comment: Novel (not mentioned) × Blind			-0.18
			(0.32)
Reviewer comment: Clearly written (not mentioned) × Blind			-0.08
			(0.31)
Reviewer comment: Relevant to audience (not mentioned) × Blind			0.47
			(0.33)
Reviewer FE		×	×
Paper FE		×	×
N	2590	2590	2590
N Clusters	245	245	245
N Papers	657	657	657
R^2	0.47	0.74	0.74

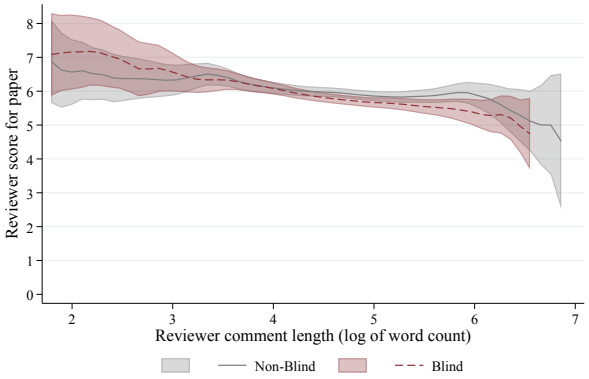
Notes. This table shows the connection between a reviewer's score for a paper and the reviewer's sentiments towards the four evaluation criteria instructed by the conference, as inferred by the reviewer's comment for the paper. Observations are at the paper-reviewer level, with the reviewer comment sentiments corresponding to averages taken over reviewers. Raters evaluated the extent to which the reviewer seemed to agree with the criteria (labeled by column) using only the comment that the reviewer left for the paper. Raters evaluated on a 5-point scale, from 1 (strongly disagree) through 5 (strongly agree), or "unclear" if the criteria was not mentioned. Those unmentioned are imputed the mean and Table A34 puts the indicator for unclear as the outcome variable. Observations are weighed by the inverse number of total raters assigned to a given paper-reviewer, so that each paper-reviewer has equal weight. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

Figure A20: Length of Reviewer Comments

(a) Distribution of word count



(b) Word count and reviewer score



Notes. These figures show the (a) distribution of reviewer comments' word counts and (b) the correlation between reviewer comments' word counts and the score that the reviewer gives the paper, separately by treatment status.

First, I find that reviewer beliefs about submissions along these dimensions are strongly linked to scoring behavior. Regressing reviewer scores on reviewer beliefs shows that submissions described as advancing the field, novel, clearly written, or relevant to the audience, also receive significantly higher scores (Table A31). This persists even after including reviewer fixed effects to account for reviewer-level differences in optimism.

Table A32: Effect of blinding on reviewer comments

	Advances		Novel		Clear		Relevant	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student	-0.16***		-0.08*		-0.18***		-0.16***	
	(0.05)		(0.04)		(0.06)		(0.05)	
Lower Rank Inst.	-0.14***		-0.12**		-0.21***		-0.12**	
	(0.05)		(0.05)		(0.06)		(0.05)	
Female	-0.01		-0.03		0.14*		-0.04	
	(0.06)		(0.05)		(0.07)		(0.06)	
Has Female PI	-0.07		0.00		0.03		-0.05	
	(0.06)		(0.06)		(0.08)		(0.06)	
Blind	-0.13**		-0.09*		-0.04		-0.04	
	(0.05)		(0.05)		(0.07)		(0.06)	
Student × Blind	0.13**	0.13**	0.09	0.10*	0.21***	0.15**	0.12**	0.13**
	(0.06)	(0.06)	(0.06)	(0.06)	(0.08)	(0.07)	(0.06)	(0.06)
Lower Rank Inst. × Blind	-0.03	0.04	-0.06	-0.04	-0.08	-0.02	-0.01	0.03
	(0.07)	(0.06)	(0.06)	(0.06)	(0.09)	(0.08)	(0.07)	(0.06)
Female × Blind	-0.00	0.00	0.01	0.02	-0.18*	-0.13	0.00	0.01
	(0.07)	(0.07)	(0.07)	(0.07)	(0.10)	(0.09)	(0.07)	(0.07)
PI Female × Blind	-0.04	0.01	-0.00	0.06	-0.11	-0.08	-0.04	-0.05
	(0.08)	(0.08)	(0.08)	(0.08)	(0.10)	(0.10)	(0.08)	(0.08)
Rater FE	×	×	×	×	×	×	×	×
Paper FE		×		×		×		×
Reviewer FE		×		×		×		×
N	6382	6382	6376	6376	6379	6379	6378	6378
N Clusters	245	245	245	245	245	245	245	245
N Papers	657	657	657	657	657	657	657	657
R^2	0.03	0.35	0.07	0.34	0.03	0.40	0.16	0.45

Notes. This table shows the effects of blinding on reviewer comments. Observations are at the paper-reviewer-rater level, where a rater evaluated the extent to which the reviewer seemed to agree with the criteria (labeled by column) using only the comment that the reviewer left for the paper. Raters evaluated on a 5-point scale, from 1 (strongly disagree) through 5 (strongly agree), or "unclear" if the criteria was not mentioned. Those unmentioned are imputed the mean and Table A34 puts the indicator for unclear as the outcome variable. Observations are weighed by the inverse number of total raters assigned to a given paper-reviewer, so that each paper-reviewer has equal weight. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

Putting enumerated reviewer beliefs on the left hand side of equation 1, I find that blinding significantly reduces student disparities in reviewer beliefs (Table A32).[39] Non-blind reviewers are significantly more likely to suggest that a submission by a non-student advances the state of the field, is novel, is written clearly, and is relevant to the conference, than a submission by a student, these gaps are essentially gone under blind reviewers. These are not driven by reviewers changing whether they mention criteria (Table A34) but rather their agreement with the criteria. This suggests that changes in reviewer beliefs of paper content and fit with the conference are important mechanisms for explaining blinding effects on student disparities. To quantify the extent to which changes in beliefs can explain total blinding effects, I isolate the variation in reviewer scores that are predicted by reviewer comment beliefs using the regression in column 1 of Table A31. Regressing the predicted values from this regression on the covariates of Equation 1 reproduces almost 2/3

[39]Table A34 shows that blinding did not significantly affect whether a criteria is mentioned in a comment.

of the blinding effects on the student score gap (Table A33), reinforcing the prominence of beliefs as a mechanism.

Table A33: Effects of Blinding Explained by Reviewer Comments

	Actual Score		Predicted Score	
	(1)	(2)	(3)	(4)
Student	-0.48***		-0.25***	
	(0.11)		(0.06)	
Lower Rank Inst.	-0.75***		-0.24***	
	(0.12)		(0.06)	
Female	-0.19		0.01	
	(0.14)		(0.07)	
Has Female PI	-0.30**		-0.03	
	(0.15)		(0.07)	
Blind	-0.46***		-0.14*	
	(0.15)		(0.08)	
Student × Blind	0.33**	0.32***	0.22**	0.20***
	(0.16)	(0.12)	(0.09)	(0.07)
Lower Rank Inst. × Blind	0.20	0.26*	-0.03	0.03
	(0.16)	(0.14)	(0.09)	(0.07)
Female × Blind	0.12	0.26*	-0.05	-0.03
	(0.19)	(0.15)	(0.10)	(0.08)
PI Female × Blind	-0.03	-0.04	-0.10	-0.06
	(0.21)	(0.18)	(0.11)	(0.09)
Rater FE	×	×	×	×
Paper FE		×		×
Reviewer FE		×		×
N	6369	6369	6369	6369
N Paper-Reviewer	2590	2590	2590	2590
N Clusters	245	245	245	245
N Papers	657	657	657	657
R^2	0.04	0.57	0.04	0.42

Notes. Observations are weighed by the inverse number of total raters assigned to a given paper-reviewer, so that each paper-reviewer has equal weight. Observations are at the paper-reviewer-rater level. Heteroskedastic robust standard errors in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A34: Effect of blinding on comments (extensive margin)

	Advances		Novel		Clear		Relevant	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student	0.005		-0.012		-0.012		0.015	
	(0.010)		(0.012)		(0.014)		(0.014)	
Lower Rank Inst.	0.014		0.007		-0.012		0.012	
	(0.011)		(0.013)		(0.014)		(0.015)	
Female	-0.004		0.006		-0.015		-0.010	
	(0.011)		(0.015)		(0.015)		(0.016)	
Has Female PI	-0.007		0.009		0.022		0.018	
	(0.012)		(0.017)		(0.019)		(0.019)	
Blind	-0.002		-0.011		-0.020		-0.018	
	(0.013)		(0.017)		(0.020)		(0.019)	
Student × Blind	-0.010	-0.014	0.002	0.001	-0.013	-0.011	-0.017	-0.026
	(0.013)	(0.013)	(0.017)	(0.016)	(0.019)	(0.018)	(0.019)	(0.016)
Lower Rank Inst. × Blind	0.008	0.003	0.023	0.013	0.016	0.015	0.033	0.010
	(0.014)	(0.015)	(0.019)	(0.017)	(0.020)	(0.018)	(0.021)	(0.018)
Female × Blind	0.004	0.004	0.009	0.013	0.012	0.014	0.014	0.011
	(0.015)	(0.016)	(0.021)	(0.019)	(0.022)	(0.021)	(0.022)	(0.020)
PI Female × Blind	0.020	0.024	-0.007	-0.029	-0.014	-0.015	-0.002	-0.008
	(0.017)	(0.017)	(0.022)	(0.021)	(0.027)	(0.024)	(0.027)	(0.025)
Paper FE		×		×		×		×
Reviewer FE		×		×		×		×
N	6382	6382	6382	6382	6382	6382	6382	6382
N Clusters	245	245	245	245	245	245	245	245
N Papers	657	657	657	657	657	657	657	657
R^2	0.11	0.27	0.16	0.36	0.10	0.32	0.28	0.49

Notes. This table shows the effects of blinding on reviewer comments, as in Table A32, but on whether a given criteria is mentioned in the comment. Observations are at the paper-reviewer-rater level, where a rater evaluated whether the criteria was not mentioned. Observations are weighed by the inverse number of total raters assigned to a given paper-reviewer, so that each paper-reviewer has equal weight. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

E Second Experiment: Directly Eliciting Reviewer Beliefs

The followup experiment was conducted during the review process for the 2025 conference which occurred October - early December of 2024. The main purpose of the second experiment was to elicit reviewer beliefs of submission outcomes.

Since 2020, the conference had completely switched to blind review, so that all reviewers did not receive author names and all applicants were assigned to three blind reviewers. Relative to this, the main changes that were made for the second experiment were to (1) include 2 additional questions during the review process to elicit reviewer beliefs about submissions' future citations and publication status:

1. What do you think will be the status of this submitted project in 5 years? If you think it will be published, which journal do you think is the best match?
   (a) Not published in peer-reviewed journal, and not available as pre-print
   (b) Not published in peer-reviewed journal, but available as pre-print
   (c) Published in peer-reviewed journal
   Please name journal: [text box]
2. Among past submissions sent to Cosyne, the average submission (or related preprint, or journal article) had 26 citations after 5 years, the median had 8, and the 90th percentile had 67. How many citations do you think this submission will have, if any, in 5 years?

and (2) invite all reviewers to complete two additional reviews after the blind reviews were finished. This was conducted on the same platform as the blind reviews. The inclusion of the additional questions and the additional reviews were explained to reviewers directly on the review platform as reviewers first completed the blind reviews:

> *To better understand our review process, we have added 2 questions for each submission you review this year. Your responses to these questions will not be sent to the authors, nor used to penalize any submissions.*
>
> *This year, we will conduct a supplemental review to gather more information. You will be assigned two additional reviews after the initial review deadline. Upon completing all of your reviews, we will email $100 to you or donate the same amount to a charity of your choice, depending on which you prefer. Your completion of these additional reviews is greatly appreciated.*

Once the blind review deadline passed, the conference directly contacted reviewers regarding the additional reviews with the following message:

> *Thank you very much for submitting your reviews for the 2025 Cosyne conference.*
>
> *To better understand how we can improve the review process, we ask for your participation in one additional step. In the following page, you will be assigned two additional submissions to review and asked your opinion about them. Your responses will be considered in the future design of Cosyne and acceptance decisions.*
>
> *We greatly appreciate your time. Upon completing all of your reviews by [date, 1 week after blind reviews are due], we will email $100 to you or donate the same amount to a charity of your choice, depending on which you prefer.*

The two additional reviews were non-blind, so that reviewers could see author names, exactly as in past years (2020 and prior) when the conference used non-blind review. The non-blind component gave the same instructions and asked the same questions as the blind review (including Overall score, publication status, and citations), but included three additional questions aimed at eliciting reviewer perceptions of alternative objectives associated with submissions:

1. If this submission were accepted for a talk, how engaging do you think the talk will be for the Cosyne audience?
    - 1-5 scale, with 1-not engaging and 5-extremely engaging
2. To what extent would the author benefit from presenting a talk at the conference?
    - 1-5 scale, with 1-would not benefit and 5-benefit a lot
3. Aside from the talk and/or poster, to what extent would Cosyne benefit from having this author team attend?
    - 1-5 scale, with 1-would not benefit and 5-benefit a lot

Table A35: Author Traits

	Mean	SD
Applicant Traits		
Student: Yes (%)	56.6	49.6
Student: No (%)	41.6	49.3
Student: Unknown (%)	1.8	13.3
Has Institution Rank (%)	76.3	42.6
Institution Rank | Have Rank	82.3	160.8
Institution Rank: Top 20 (%)	27.0	44.4
Institution Rank: 20+ (%)	49.3	50.0
Institution Rank: Not University (%)	19.6	39.7
Institution Unknown (%)	4.1	19.8
Gender: Female (%)	29.8	45.8
Gender: Male (%)	68.1	46.6
Gender: Unknown (%)	2.1	14.3
PI Traits		
Gender: Female (%)	19.5	39.7
Gender: Male (%)	80.1	40.0
Gender: Unknown (%)	0.4	6.3
Other		
Number of Authors	4.4	2.9
Solo Author (%)	2.0	14.0
Observations (Papers)	1003	

Notes. This table provides descriptive statistics for the full sample of papers submitted to the conference in 2025, 5 years after the main experiment. Observations are at the paper level.

223 of the 328 (68%) Reviewers complete non-blind review (408 of the 1,003 papers). Reviewers who completed non-blind reviews tended to be more male, student, and affiliated with worse-ranked institutions.

Table A36: Reviewers in second experiment

	(1)	(2)	(3)	(4)
Variable	All	No Non-Blind	Yes Non-Blind	Diff
Gender: Female	0.33	0.41	0.29	-0.12**
	(0.47)	(0.50)	(0.46)	(0.04)
Gender: Male	0.67	0.59	0.71	0.12**
	(0.47)	(0.50)	(0.46)	(0.04)
Student	0.03	0.00	0.04	0.04*
	(0.18)	(0.00)	(0.20)	(0.06)
Inst: Top 20	0.31	0.38	0.30	-0.08
	(0.46)	(0.49)	(0.46)	(0.16)
Inst: 20+	0.39	0.31	0.43	0.12*
	(0.49)	(0.47)	(0.50)	(0.06)
Inst: Not Uni	0.24	0.26	0.22	-0.04
	(0.43)	(0.44)	(0.41)	(0.40)
Inst: Rank Unknown	0.05	0.05	0.06	0.01
	(0.23)	(0.21)	(0.23)	(0.67)
Completed Nonblind review	0.72	0.00	1.00	1.00
	(0.45)	(0.00)	(0.00)	()
Observations	328	87	223	328

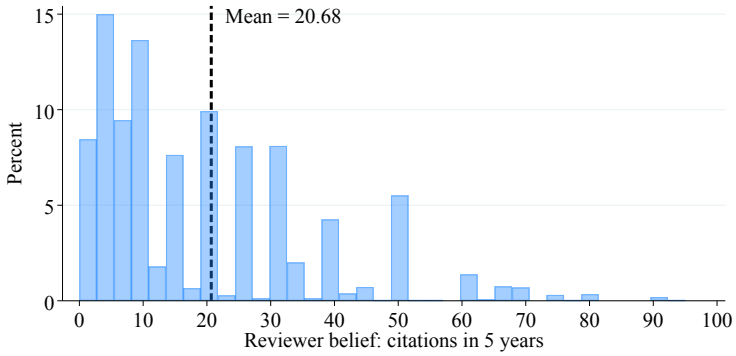
Notes. This table shows summary statistics for reviewers in the second experiment, and separately for reviewers who completed their additional non-blind reviews and those who did not. The final column shows the mean difference from single hypothesis t-tests between the means of reviewers who completed their additional non-blind reviews and those who did not. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A37: Selection into non-blind review

	OLS			Probit		
	(1)	(2)	(3)	(4)	(5)	(6)
	_	_	_	completed	completed	completed
Reviewer: student	0.14*	0.15**	0.17**	0.52*	0.55*	0.59*
	(0.07)	(0.08)	(0.08)	(0.30)	(0.30)	(0.31)
Reviewer: Lower rank inst.	0.10**	0.10**	0.11**	0.31**	0.31**	0.32**
	(0.04)	(0.04)	(0.04)	(0.13)	(0.13)	(0.13)
Reviewer: female	-0.12***	-0.12***	-0.12***	-0.34***	-0.35***	-0.36***
	(0.04)	(0.04)	(0.04)	(0.11)	(0.11)	(0.11)
Student		-0.04	-0.04		-0.12	-0.11
		(0.04)	(0.04)		(0.11)	(0.11)
Lower Rank Inst.		0.04	0.03		0.11	0.09
		(0.05)	(0.05)		(0.13)	(0.13)
Female		-0.01	-0.02		-0.02	-0.07
		(0.04)	(0.04)		(0.11)	(0.12)
Has Female PI		-0.00	0.01		-0.00	0.03
		(0.05)	(0.05)		(0.13)	(0.13)
Subfield FE			×			×
N	648	648	648	648	648	648

Notes. Dependent variable is an indicator for whether the reviewer completed the assigned submission. Observation is at the paper-reviewer level: each reviewer shows up twice because each reviewer was assigned to two additional reviews. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

Figure A21: Reviewers' Citation Beliefs



Notes. Observation at the paper-reviewer level, pooling both blind and non-blind reviewers. Citation counts above the 99th percentile are trimmed.

Table A38: Reviewer Beliefs

	Blind			Non-blind					
	Cites	Published	Pub (Weight)	Cites	Published	Pub (Weight)	Talk	Conference	Author
All	20.24	0.64	5.26	23.69	0.70	5.97	2.91	3.65	3.37
	(22.31)	(0.48)	(6.78)	(22.53)	(0.46)	(6.70)	(1.21)	(1.15)	(1.25)
By Applicant Student Status									
Student	19.18	0.63	4.90	21.26	0.70	6.07	2.74	3.54	3.28
	(19.48)	(0.48)	(6.53)	(20.23)	(0.46)	(6.88)	(1.18)	(1.18)	(1.27)
Not Student	21.81	0.66	5.86	26.98	0.71	5.91	3.13	3.78	3.50
	(25.80)	(0.47)	(7.14)	(24.98)	(0.46)	(6.48)	(1.22)	(1.09)	(1.22)
Difference	-2.63***	-0.03*	-0.96***	-5.72**	-0.01	0.16	-0.39***	-0.24**	-0.22*
	[0.86]	[0.02]	[0.28]	[2.23]	[0.04]	[0.73]	[0.12]	[0.11]	[0.12]
By Applicant Institution Rank									
Lower Ranked	19.28	0.63	4.94	22.11	0.70	6.06	2.81	3.61	3.39
	(19.94)	(0.48)	(6.40)	(21.08)	(0.46)	(6.85)	(1.17)	(1.19)	(1.24)
Top 20	24.01	0.69	6.43	25.94	0.72	6.73	3.05	3.69	3.43
	(28.74)	(0.46)	(7.34)	(21.43)	(0.45)	(7.24)	(1.19)	(1.14)	(1.19)
Difference	-4.73***	-0.07***	-1.49***	-3.83	-0.03	-0.67	-0.23*	-0.08	-0.04
	[1.05]	[0.02]	[0.33]	[2.59]	[0.05]	[0.92]	[0.14]	[0.14]	[0.14]
By Applicant Gender									
Female	20.43	0.66	5.59	24.66	0.72	5.69	2.87	3.69	3.49
	(18.58)	(0.48)	(6.96)	(23.76)	(0.45)	(5.53)	(1.19)	(1.13)	(1.19)
Male	20.27	0.64	5.19	23.68	0.71	6.32	2.95	3.68	3.34
	(23.94)	(0.48)	(6.75)	(22.15)	(0.46)	(7.21)	(1.22)	(1.13)	(1.26)
Difference	0.16	0.02	0.40	0.99	0.02	-0.63	-0.08	0.02	0.15
	[0.92]	[0.02]	[0.30]	[2.42]	[0.05]	[0.79]	[0.13]	[0.12]	[0.13]
By PI Gender									
Female	20.18	0.62	5.18	20.36	0.62	4.60	2.98	3.71	3.24
	(20.35)	(0.49)	(6.72)	(19.01)	(0.49)	(5.39)	(1.11)	(1.12)	(1.28)
Male	20.17	0.65	5.29	24.55	0.72	6.34	2.89	3.63	3.41
	(22.31)	(0.48)	(6.80)	(23.30)	(0.45)	(6.97)	(1.24)	(1.16)	(1.25)
Difference	0.01	-0.03	-0.11	-4.19	-0.10*	-1.74**	0.09	0.07	-0.17
	[1.04]	[0.02]	[0.35]	[2.74]	[0.05]	[0.87]	[0.14]	[0.13]	[0.15]
N Paper-reviewer	2825	3009	2428	415	441	349	440	440	440
N Paper	1002	1003	996	383	408	323	407	407	407

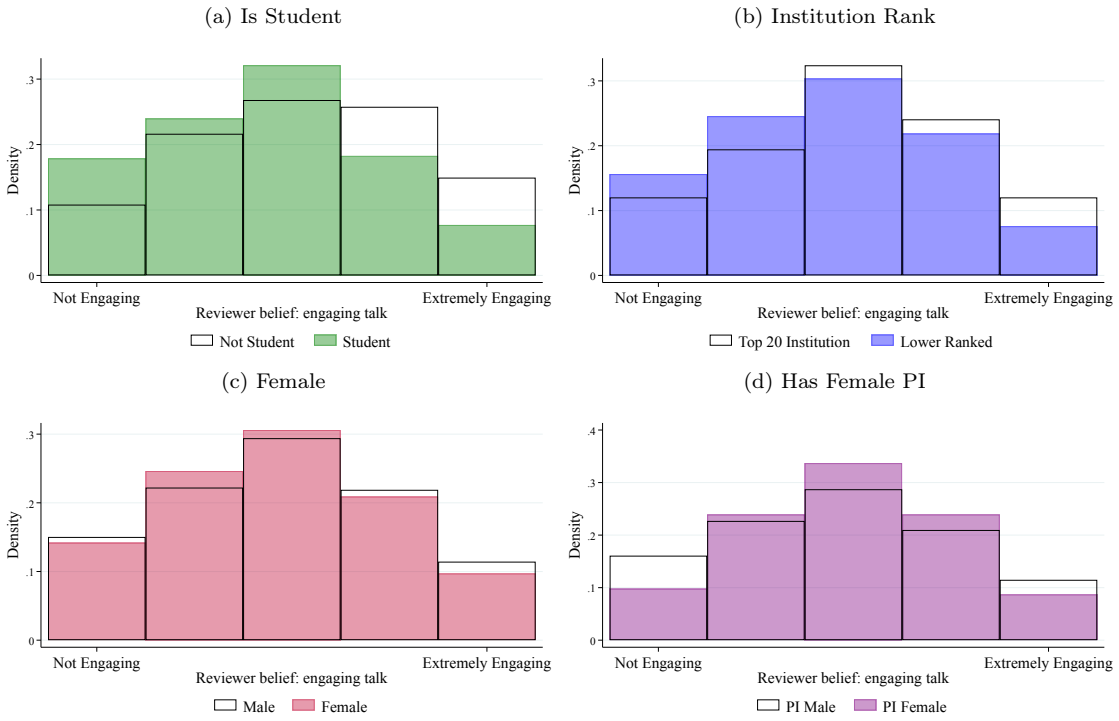
Notes. This table shows summary statistics for reviewers' predictions of citation and publication statuses five years after the experiment. Observations are at the paper-reviewer level. Each column captures summary statistics for a unique outcome: number of citations (including zeros), whether the paper is available online, whether the paper is published, journal-weighted publication status. 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 predicted to be 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. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A39: Blinding effects on reviewer scores and beliefs in second experiment

	Overall		N Cites		Published	
	(1)	(2)	(3)	(4)	(5)	(6)
Blind	-0.17	-0.16	-1.52	3.69	0.02	0.00
	(0.24)	(0.24)	(2.56)	(5.12)	(0.05)	(0.06)
Student	-0.27		-4.73**		0.02	
	(0.19)		(2.24)		(0.04)	
Lower Rank Inst.	-0.15		-1.60		-0.02	
	(0.22)		(2.37)		(0.05)	
Female	-0.14		-0.12		0.03	
	(0.19)		(2.06)		(0.04)	
Has Female PI	-0.02		-2.09		-0.02	
	(0.25)		(1.99)		(0.05)	
Student × Blind	0.09	0.30	1.97	0.79	-0.06	-0.02
	(0.21)	(0.20)	(2.35)	(2.70)	(0.05)	(0.05)
Lower Rank Inst. × Blind	-0.51**	-0.61**	-2.98	-6.25	-0.07	-0.09
	(0.24)	(0.24)	(2.59)	(4.04)	(0.05)	(0.06)
Female × Blind	0.22	0.39*	0.49	0.16	-0.01	0.05
	(0.21)	(0.21)	(2.28)	(2.69)	(0.05)	(0.05)
Has Female PI × Blind	-0.02	-0.09	1.57	-0.24	-0.00	-0.03
	(0.27)	(0.25)	(2.22)	(2.80)	(0.06)	(0.06)
Paper FE		×		×		×
Reviewer FE	×	×	×	×	×	×
N	3450	3450	3240	3232	3450	3450
N Clusters	328	328	313	313	328	328
N Papers	1003	1003	1002	994	1003	1003
R^2	0.18	0.65	0.29	0.60	0.28	0.60

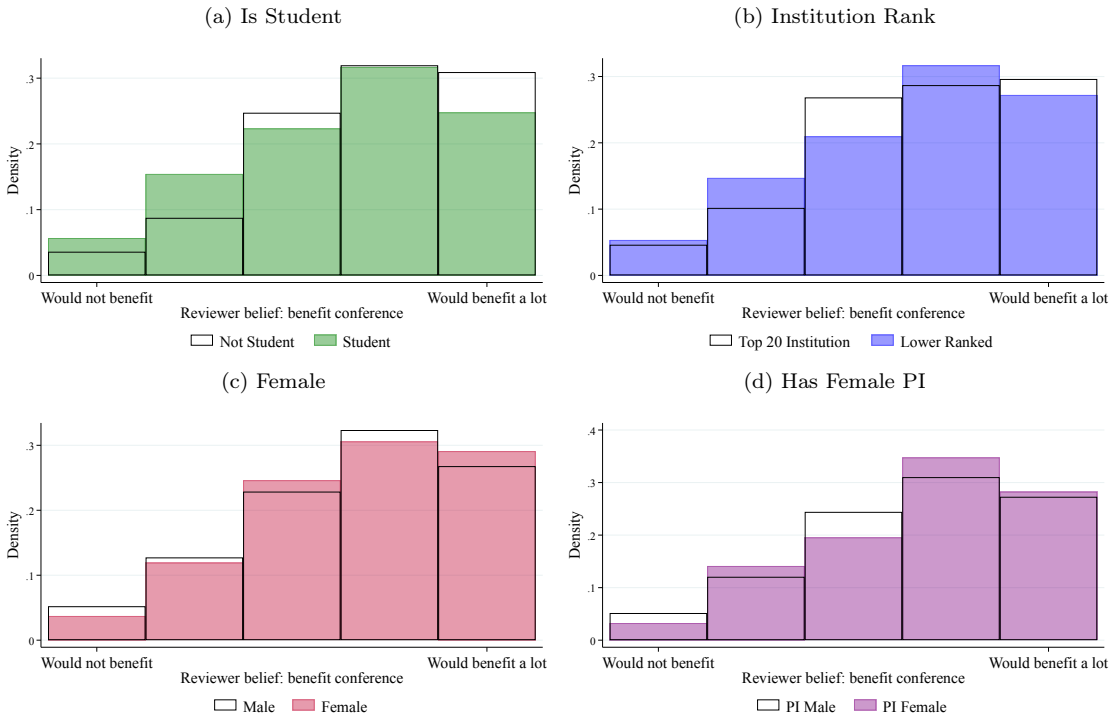
Notes. This table shows the effects of blinding on (columns 1-2) reviewer scores, and reviewer beliefs (columns 3-4) of citations and (columns 5-6) publication 5 years later. Observations are at the paper-reviewer level. Blind is an indicator for whether the reviewer was randomly assigned to score papers 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. * $p < .1$, ** $p < .05$, *** $p < .01$

Figure A22: Distribution of Reviewer Beliefs: Extent to which talk is engaging



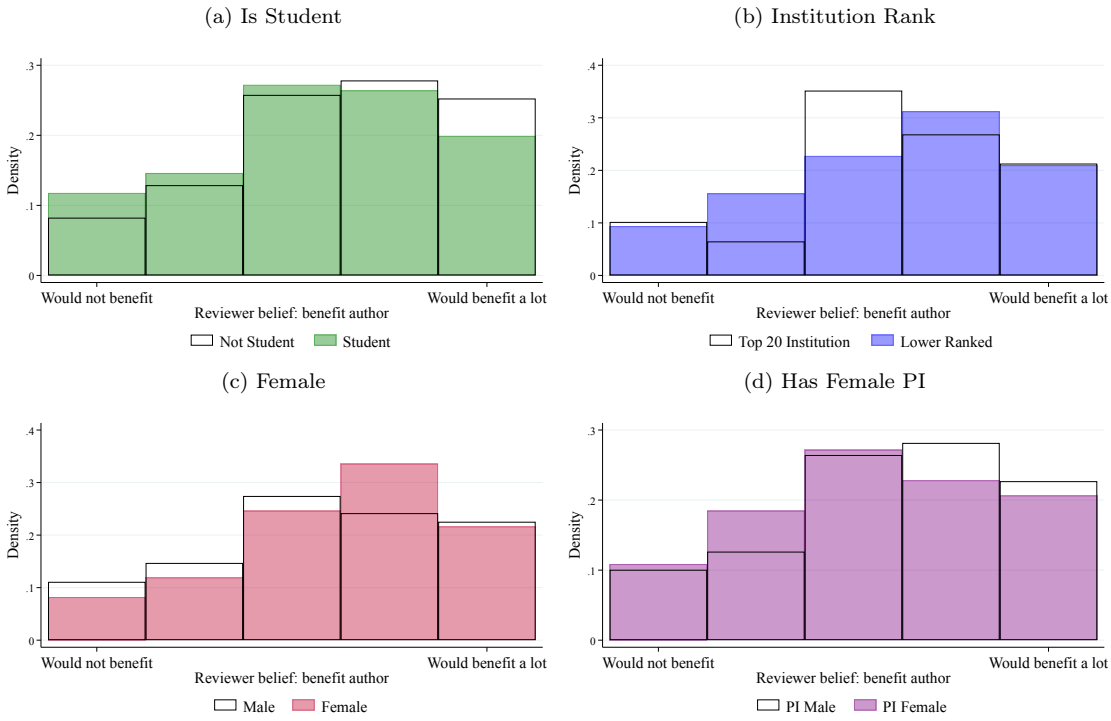
Notes. "If this submission were accepted for a talk, how engaging do you think the talk will be for the Cosyne audience?"

Figure A23: Distribution of Reviewer Beliefs: Extent to which conference would benefit from having author attend



Notes. "Aside from the talk and/or poster, to what extent would Cosyne benefit from having this author team attend?"

Figure A24: Distribution of Reviewer Beliefs: Extent to which author would benefit



Notes. "To what extent would the author benefit from presenting a talk at the conference?"

Table A40: Reviewer heterogeneity in non-blind beliefs

	Panel A: By Reviewer Gender											
	N Cites		Is Online		Is Published		Talk Engaging		Benefit Conf		Benefit Author	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Student	-6.94**		-0.01		0.02		-0.48***		-0.24*		-0.40**	
	(2.99)		(0.04)		(0.06)		(0.15)		(0.14)		(0.15)	
Lower Rank Inst.	-1.52		0.01		-0.03		-0.14		0.00		0.01	
	(3.14)		(0.03)		(0.06)		(0.17)		(0.17)		(0.19)	
Female	-0.07		0.01		0.06		-0.12		0.04		0.13	
	(2.94)		(0.03)		(0.06)		(0.14)		(0.14)		(0.15)	
Has Female PI	-1.98		-0.07		-0.14**		0.16		-0.01		-0.30*	
	(2.94)		(0.05)		(0.07)		(0.16)		(0.16)		(0.17)	
Reviewer: female	-3.55		-0.02		0.02		-0.20		-0.01		0.04	
	(6.62)		(0.07)		(0.13)		(0.35)		(0.31)		(0.31)	
Student × Reviewer Female	9.87*	4.77	-0.01	0.08	-0.01	0.07	0.41	0.24	0.20	0.19	0.44*	0.23
	(5.55)	(4.33)	(0.06)	(0.05)	(0.11)	(0.11)	(0.27)	(0.28)	(0.25)	(0.25)	(0.25)	(0.22)
Lower Rank Inst. × Reviewer Female	-6.27	-3.27	0.04	0.03	-0.06	-0.20	-0.06	-0.17	-0.03	-0.10	-0.17	-0.06
	(6.33)	(6.53)	(0.06)	(0.08)	(0.12)	(0.12)	(0.31)	(0.33)	(0.29)	(0.26)	(0.28)	(0.29)
Female × Reviewer Female	3.28	7.37	0.01	-0.03	-0.14	-0.10	0.16	0.41	-0.04	0.22	0.09	0.09
	(4.89)	(4.55)	(0.06)	(0.04)	(0.11)	(0.11)	(0.26)	(0.30)	(0.28)	(0.29)	(0.26)	(0.25)
Has Female PI × Reviewer Female	-5.07	-0.59	0.15**	0.11	0.19	0.15	0.10	0.37	0.21	0.04	0.89***	0.01
	(5.58)	(4.06)	(0.06)	(0.09)	(0.14)	(0.18)	(0.32)	(0.34)	(0.33)	(0.31)	(0.27)	(0.22)
Reviewer FE		×		×		×		×		×		×
N	415	408	441	436	441	436	440	434	440	434	440	434
N Clusters	211	204	223	218	223	218	223	217	223	217	223	217
R^2	0.09	0.73	0.09	0.66	0.09	0.72	0.09	0.58	0.11	0.67	0.11	0.76
	Panel B: By Reviewer Institution Rank											
	N Cites		Is Online		Is Published		Talk Engaging		Benefit Conf		Benefit Author	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Student	-4.10		-0.01		0.09		-0.03		0.05		0.13	
	(4.25)		(0.04)		(0.08)		(0.23)		(0.23)		(0.23)	
Lower Rank Inst.	-7.43		0.05		-0.02		-0.13		0.06		-0.02	
	(4.62)		(0.04)		(0.08)		(0.21)		(0.23)		(0.22)	
Female	1.38		-0.03		-0.05		-0.17		-0.08		-0.26	
	(3.64)		(0.04)		(0.08)		(0.20)		(0.21)		(0.21)	
Has Female PI	2.07		0.01		0.05		0.37		0.45*		0.41	
	(5.15)		(0.05)		(0.09)		(0.24)		(0.26)		(0.27)	
Reviewer: Lower rank inst.	-3.55		0.04		0.05		0.42		0.35		0.10	
	(6.07)		(0.07)		(0.12)		(0.29)		(0.28)		(0.29)	
Student × Reviewer Lower Rank Inst	-0.07	-9.94**	-0.06	-0.08	-0.13	-0.09	-0.56*	-0.76***	-0.40	-0.42**	-0.58**	-0.55***
	(5.68)	(4.71)	(0.06)	(0.06)	(0.11)	(0.08)	(0.29)	(0.27)	(0.30)	(0.21)	(0.28)	(0.18)
Lower Rank Inst. × Reviewer Lower Rank Inst	4.06	-3.70	-0.04	0.06	-0.03	-0.01	-0.04	-0.07	-0.17	-0.13	0.17	0.14
	(5.78)	(4.47)	(0.06)	(0.04)	(0.11)	(0.08)	(0.29)	(0.28)	(0.29)	(0.24)	(0.29)	(0.20)
Female × Reviewer Lower Rank Inst	1.96	0.92	0.09	0.12**	0.10	0.03	0.12	0.36	0.02	0.28	0.77**	0.29
	(5.48)	(5.43)	(0.06)	(0.06)	(0.12)	(0.10)	(0.28)	(0.25)	(0.29)	(0.25)	(0.30)	(0.22)
Has Female PI × Reviewer Lower Rank Inst	-8.48	2.96	-0.04	0.10	-0.27**	0.04	-0.33	0.40	-0.71**	0.04	-0.84**	-0.08
	(6.24)	(4.19)	(0.07)	(0.09)	(0.13)	(0.11)	(0.31)	(0.29)	(0.34)	(0.27)	(0.36)	(0.23)
Reviewer FE		×		×		×		×		×		×
N	415	408	441	436	441	436	440	434	440	434	440	434
N Clusters	211	204	223	218	223	218	223	217	223	217	223	217
R^2	0.09	0.76	0.13	0.72	0.11	0.74	0.10	0.61	0.15	0.70	0.12	0.78

Notes. This table presents disparities in reviewers' predictions of paper quality outcomes and the alternative objectives, as in Table 6, but including interactions for reviewer characteristics. Alternative objectives include: the extent to which the talk would be engaging if it were accepted (1-5 scale), extent to which the author benefits from presenting a talk at the conference (1-5 scale), and the extent to which the conference would benefit from having the author team attend (1-5 scale). Observations are at the paper-reviewer level. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A41: Reviewer heterogeneity in the non-blind scoring equation

	(1)	(2)	(3)	(4)	(5)
Reviewer belief: N citations	0.01*	0.01**	0.02***	0.02***	0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Reviewer belief: paper is online	0.45	0.51	0.80**	0.75*	0.72
	(0.49)	(0.48)	(0.41)	(0.40)	(0.49)
Reviewer belief: paper is published	0.71**	0.85***	0.54**	0.52**	0.72**
	(0.28)	(0.29)	(0.24)	(0.24)	(0.31)
Reviewer belief: engaging talk	0.68***	0.58***	0.71***	0.68***	0.57***
	(0.10)	(0.11)	(0.10)	(0.11)	(0.12)
Reviewer belief: benefit conference	0.56***	0.47***	0.54***	0.64***	0.57***
	(0.10)	(0.13)	(0.10)	(0.10)	(0.12)
Reviewer belief: benefit author	0.15*	0.22**	0.14	0.16	0.22
	(0.09)	(0.10)	(0.09)	(0.13)	(0.13)
Reviewer belief: N citations × Reviewer Female	0.00	-0.02			-0.01
	(0.01)	(0.01)			(0.01)
Reviewer belief: paper is online × Reviewer Female	0.09	-0.19			-0.43
	(0.73)	(0.80)			(0.82)
Reviewer belief: paper is published × Reviewer Female	-0.21	-0.36			-0.30
	(0.42)	(0.40)			(0.41)
Reviewer belief: engaging talk × Reviewer Female		0.40*			0.38
		(0.24)			(0.24)
Reviewer belief: benefit conference × Reviewer Female		0.18			0.14
		(0.24)			(0.25)
Reviewer belief: benefit author × Reviewer Female		-0.11			-0.13
		(0.17)			(0.18)
Reviewer belief: N citations × Reviewer Lower Rank Inst			-0.02**	-0.01	-0.01
			(0.01)	(0.01)	(0.01)
Reviewer belief: paper is online × Reviewer Lower Rank Inst			-0.75	-0.71	-0.43
			(0.77)	(0.77)	(0.81)
Reviewer belief: paper is published × Reviewer Lower Rank Inst			0.16	0.28	0.22
			(0.42)	(0.46)	(0.46)
Reviewer belief: engaging talk × Reviewer Lower Rank Inst				0.07	0.06
				(0.22)	(0.21)
Reviewer belief: benefit conference × Reviewer Lower Rank Inst				-0.23	-0.24
				(0.23)	(0.24)
Reviewer belief: benefit author × Reviewer Lower Rank Inst				-0.02	0.02
				(0.18)	(0.17)
Reviewer FE	×	×	×	×	×
N	434	434	434	434	434
N Clusters	217	217	217	217	217
R^2	0.90	0.90	0.90	0.90	0.90

Notes. Dependent variable is the score that a paper receives from a reviewer, among non-blind reviews who were asked about talk quality and benefitting others. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A42: Non-Blind Disparities in Reviewer Beliefs (weighted)

	N Citations		Is Online		Is Published		Talk Quality		Benefit Conference		Benefit Author	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Student	-3.94*	-4.50*	-0.01	-0.01	0.02	0.03	-0.33***	-0.31**	-0.18	-0.16	-0.22*	-0.21*
	(2.39)	(2.45)	(0.03)	(0.03)	(0.05)	(0.05)	(0.13)	(0.13)	(0.12)	(0.12)	(0.13)	(0.13)
Lower Rank Inst.	-3.77	1.04	0.03	0.06	-0.05	0.01	-0.13	0.18	0.02	0.24*	0.02	0.19
	(2.77)	(2.77)	(0.03)	(0.04)	(0.05)	(0.05)	(0.15)	(0.15)	(0.14)	(0.15)	(0.15)	(0.16)
Female	0.60	-1.48	0.01	-0.00	0.01	-0.01	-0.08	-0.19	0.03	-0.04	0.12	0.07
	(2.46)	(2.56)	(0.03)	(0.03)	(0.05)	(0.05)	(0.13)	(0.13)	(0.12)	(0.12)	(0.13)	(0.13)
Has Female PI	-3.37	-0.65	-0.04	-0.02	-0.09	-0.06	0.11	0.31**	0.01	0.14	-0.13	-0.02
	(2.50)	(2.55)	(0.04)	(0.04)	(0.06)	(0.06)	(0.15)	(0.14)	(0.14)	(0.13)	(0.16)	(0.16)
Submission Content		8.39***		0.06***		0.11***		0.57***		0.40***		0.31***
		(1.75)		(0.02)		(0.03)		(0.08)		(0.08)		(0.09)
IPW	×	×	×	×	×	×	×	×	×	×	×	×
Subfield FE	×	×	×	×	×	×	×	×	×	×	×	×
Outcome mean	23.67	23.67	0.91	0.91	0.70	0.70	2.91	2.91	3.65	3.65	3.37	3.37
First stage F-stat	-	75.93	-	95.34	-	95.34	-	97.57	-	97.57	-	97.57
N Paper-reviewers	415	415	441	441	441	441	440	440	440	440	440	440
N Papers	383	383	408	408	408	408	407	407	407	407	407	407

Notes. This table presents disparities in reviewers' predictions of paper quality outcomes and the alternative objectives, as in Table 6, except that paper-reviewer level observations are weighted by the probability that the non-blind assignment was completed, as predicted by reviewer and paper traits. IPW indicates that observations are weighted by the inverse of predicted probabilities as estimated from the final column in Table A37. Alternative objectives include: the extent to which the talk would be engaging if it were accepted (1-5 scale), extent to which the author benefits from presenting a talk at the conference (1-5 scale), and the extent to which the conference would benefit from having the author team attend (1-5 scale). Observations for odd-numbered columns are at the paper-level, with heteroskedastic robust standard errors in parentheses. Even-numbered columns show subgroup differences conditional on submission content as proxied by a paper's blind scores: to account for noise in blind scores, a given blind review score for a paper is instrumented by the average score that the paper's other blind reviewers gave, as described in Section 5.1. Standard errors are clustered at the paper level. For even numbered columns, standard errors are clustered at the paper-level. * $p < .1$, ** $p < .05$, *** $p < .01$

F Model Details

F.1 Non-linearities in the Model

The main text uses linear approximations (e.g. equation 7). However, the approach can be adapted to allow for more flexible functional forms in the expectation functions. As an example, consider including in the paper quality expectation function a quadratic term in the paper's expected blind score:

$$\mathbb{E}[Q|x, S^B(v_p)] = \beta_0 + \beta_1 S^B(v_p) + \beta_2 x_p + \beta_3 S^B(v_p)^2 \tag{A4}$$

β_3 is not identified by instrumenting for the observed blind score and its square from a given reviewer, by the observed blind score and its square from the paper's other reviewer (in other words, instrument $\tilde{S}_{p,r}$ and $(\tilde{S}_{p,r})^2$ with $\tilde{S}_{p,r'}$ and $(\tilde{S}_{p,r'})^2$ for reviewers $r \neq r'$). This is because unlike the case of instrumental variables when the endogenous variable is observed, because $S^B(v_p)$ is unobserved so that squaring the noise-containing proxy variable $(\tilde{S}_{p,r})$ introduces squares of the noise term.

As a simplified example to illustrate, suppose we are interested in identifying the parameters from the following regression:

$$Y_p = \beta_1 v_p + \beta_2 v_p^2 + \varepsilon_p \tag{A5}$$

where ε_p is an error term. Going forward I remove p subscripts for simplicity. v is not observed, but the data contain two proxies of it:

$$s_1 = v + u_1\,,\ s_2 = v + u_2 \tag{A6}$$

where u_1 and u_2 are independent and identically distributed error terms with mean zero ($\mathbb{E}[u] \equiv \mathbb{E}[u_1] = \mathbb{E}[u_2]$). Rewriting equation A5 in terms of s_1:

$$Y = \beta_1(s_1 - u_1) + \beta_2(s_1 - u_1)^2 + \varepsilon \tag{A7}$$

$$= \beta_1 s_1 + \beta s_1^2 + \underbrace{\varepsilon - \beta_1 u_1 - 2\beta_2 s_1 u_1 + \beta_2 u_1^2}_{\equiv e} \tag{A8}$$

Using s_2 as an instrument for s_1 and s_2^2 as an instrument for s_1^2 requres satisfying exogeneity. In particular, evaluating exogeneity for s_2^2:

$$\mathbb{E}[s_2^2 e] = \mathbb{E}[s_2^2 \varepsilon] - \beta_1\,\mathbb{E}[s_2^2 u_1] - 2\beta_2\,\mathbb{E}[s_2^2 s_1 u_1] + \beta_2\,\mathbb{E}[s_2^2 u_1^2] \tag{A9}$$

$$= -2\beta_2\,\mathbb{E}[s_2^2 s_1 u_1] + \beta_2\,\mathbb{E}[s_2^2 u_1^2] \tag{A10}$$

$$= -\beta_2\,\mathbb{E}[v^2 u^2] - \beta_2\,\mathbb{E}[u_2^2 u_1^2] \tag{A11}$$

where the second equality is due to independence of error terms, third equality is because $\mathbb{E}[s_2^2 s_1 u_1] = \mathbb{E}[(v^2 + 2vu_2 + u_2^2)(vu_1 + u_1^2)] = \mathbb{E}[v^2 u_1^2 + u_2^2 u_1^2]$ and $\mathbb{E}[s_2^2 u_1^2] = \mathbb{E}[(v^2 + 2vu_2 + u_2^2)u_1^2] = \mathbb{E}[v^2 u_1^2 + u_2^2 u_1^2]$. In other words, exogeneity only holds if $\beta_2 = 0$ or $\mathbb{E}[u^2] = 0$.

Instead, the paramters of equation A5 can be identified using alternative approaches. Specifically, moments of the unobserved latent variable can be identified using moments of the proxy measures (Hausman et al., 1991, 1995). Schennach (2016) provides a review.

F.2 Additional Model Estimates

Table A43: Decomposing Non-Blind Score Gaps

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	-0.28	-0.19	-0.11	0.13
	[-0.51,0.00]	[-0.44,0.23]	[-0.39,0.23]	[-0.18,0.57]
	Panel B: Decomposition components			
Submission content	-0.17	-0.54	-0.05	-0.32
	[-0.44,0.05]	[-1.02,-0.30]	[-0.37,0.23]	[-0.73,-0.05]
Accurate statistical discrimination	0.08	0.00	-0.18	0.09
	[-0.05,0.26]	[-0.14,0.22]	[-0.41,0.01]	[-0.04,0.31]
Inaccurate statistical discrimination	-0.13	0.06	0.16	-0.13
	[-0.34,0.04]	[-0.19,0.27]	[-0.04,0.40]	[-0.41,0.05]
Alternative objectives	-0.33	0.25	-0.12	0.34
	[-0.59,-0.07]	[-0.07,0.58]	[-0.41,0.14]	[0.08,0.67]
Unexplained	0.10	-0.49	0.03	-0.18
	[-0.30,0.54]	[-0.95,0.07]	[-0.39,0.55]	[-0.72,0.36]

Notes. This table decomposes disparities in non-blind scores. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5) and subfield. Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level. Sample is papers with at least 2 blind scores (98% of the full sample). Table A44 shows the decomposition for each sub-component.

Table A44: Decomposing Non-Blind Score Gaps: By Components

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	-0.28	-0.19	-0.11	0.13
	[-0.51,0.002]	[-0.44,0.23]	[-0.39,0.23]	[-0.18,0.57]
	Panel B: Decomposition components			
Submission content	-0.17	-0.54	-0.05	-0.32
	[-0.44,0.05]	[-1.02,-0.30]	[-0.37,0.23]	[-0.73,-0.05]
Accurate statistical discrimination	0.08	0.001	-0.18	0.09
	[-0.05,0.26]	[-0.14,0.22]	[-0.41,0.01]	[-0.04,0.31]
N Citations	-0.005	-0.02	-0.11	0.02
	[-0.09,0.08]	[-0.13,0.10]	[-0.28,0.02]	[-0.06,0.15]
Online	0.03	-0.004	-0.01	0.03
	[-0.03,0.15]	[-0.05,0.06]	[-0.09,0.03]	[-0.03,0.13]
Published	0.05	0.03	-0.05	0.04
	[0.002,0.15]	[-0.02,0.12]	[-0.14,0.01]	[-0.01,0.14]
Inaccurate statistical discrimination	-0.13	0.06	0.16	-0.13
	[-0.34,0.04]	[-0.19,0.27]	[-0.04,0.40]	[-0.41,0.05]
N Citations	-0.05	0.04	0.10	-0.02
	[-0.17,0.04]	[-0.09,0.18]	[-0.02,0.26]	[-0.15,0.08]
Online	-0.04	0.03	0.02	-0.03
	[-0.18,0.04]	[-0.04,0.14]	[-0.05,0.11]	[-0.16,0.04]
Published	-0.04	-0.02	0.05	-0.08
	[-0.14,0.03]	[-0.14,0.07]	[-0.04,0.16]	[-0.24,0.01]
Alternative objectives	-0.33	0.25	-0.12	0.34
	[-0.59,-0.07]	[-0.07,0.58]	[-0.41,0.14]	[0.08,0.67]
Talk quality	-0.22	0.09	-0.12	0.24
	[-0.39,-0.06]	[-0.10,0.28]	[-0.30,0.02]	[0.09,0.46]
Author benefit	-0.03	0.02	0.01	0.003
	[-0.10,0.02]	[-0.03,0.09]	[-0.03,0.06]	[-0.05,0.06]
Conference benefit	-0.08	0.14	-0.02	0.09
	[-0.21,0.03]	[-0.002,0.31]	[-0.14,0.11]	[-0.02,0.26]
Unexplained	0.10	-0.49	0.03	-0.18
	[-0.30,0.54]	[-0.95,0.07]	[-0.39,0.55]	[-0.72,0.36]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but dis-aggregates each of the outcome measures. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5) and subfield. Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. Sample is papers with at least 2 blind scores (98% of the full sample). 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A45: Decomposing the Non-Blind Score Gap: Without Conditioning on Submission Content

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
	Panel B: Decomposition components			
Accurate statistical discrimination	0.03	-0.19	-0.19	-0.02
	[-0.12,0.16]	[-0.43,-0.01]	[-0.43,-0.02]	[-0.19,0.14]
Inaccurate statistical discrimination	-0.07	0.17	0.22	-0.07
	[-0.25,0.09]	[-0.05,0.42]	[0.02,0.49]	[-0.30,0.11]
Alternative objectives	-0.37	-0.04	0.01	0.15
	[-0.65,-0.11]	[-0.37,0.25]	[-0.25,0.29]	[-0.12,0.45]
Unexplained	-0.03	-0.66	-0.20	-0.25
	[-0.42,0.35]	[-1.06,-0.21]	[-0.60,0.22]	[-0.70,0.21]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but without proxying for submission content using blind scores. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. Sample is papers with at least 2 blind scores (98% of the full sample). 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A46: Decomposing the Non-Blind Score Gap: By reviewer institution

(a) Top 20 ranked institution

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.33	-0.48	0.45	-0.63
	[-0.66,-0.02]	[-0.82,-0.15]	[0.08,0.83]	[-1.00,-0.21]
Conditional on submission content	-0.27	-0.01	0.00	-0.17
	[-0.67,0.24]	[-0.40,0.91]	[-0.77,0.43]	[-0.68,0.75]
	Panel B: Decomposition components			
Submission content	-0.06	-0.47	0.46	-0.46
	[-0.53,0.28]	[-1.48,-0.13]	[0.07,1.26]	[-1.25,-0.06]
Accurate statistical discrimination	0.03	-0.05	-0.27	0.09
	[-0.35,0.44]	[-0.55,0.41]	[-1.02,0.35]	[-0.27,0.58]
Inaccurate statistical discrimination	-0.10	0.12	0.05	0.06
	[-0.75,0.54]	[-0.61,1.00]	[-0.70,0.70]	[-0.61,0.73]
Alternative objectives	0.07	0.55	-0.66	0.86
	[-0.69,0.72]	[-0.10,1.79]	[-1.67,-0.10]	[0.15,2.06]
Unexplained	-0.27	-0.64	0.88	-1.18
	[-1.27,1.00]	[-2.14,0.82]	[-0.19,2.23]	[-2.63,0.14]

(b) Lower ranked institution

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.44	-0.69	-0.32	0.00
	[-0.71,-0.16]	[-1.01,-0.39]	[-0.69,0.05]	[-0.39,0.38]
Conditional on submission content	-0.19	-0.13	0.02	0.35
	[-0.48,0.29]	[-0.48,0.60]	[-0.38,0.65]	[-0.06,0.97]
	Panel B: Decomposition components			
Submission content	-0.25	-0.56	-0.34	-0.35
	[-0.72,0.03]	[-1.30,-0.24]	[-1.03,0.02]	[-1.01,0.02]
Accurate statistical discrimination	0.07	0.03	-0.07	0.05
	[-0.12,0.34]	[-0.18,0.29]	[-0.43,0.29]	[-0.16,0.34]
Inaccurate statistical discrimination	-0.04	-0.02	0.08	-0.20
	[-0.50,0.35]	[-0.46,0.43]	[-0.47,0.66]	[-0.80,0.25]
Alternative objectives	-0.48	0.10	-0.01	0.21
	[-0.95,0.06]	[-0.38,0.75]	[-0.53,0.49]	[-0.27,1.00]
Unexplained	0.26	-0.24	0.02	0.29
	[-0.48,1.00]	[-1.15,0.80]	[-0.77,0.97]	[-0.63,1.26]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but subsetting non-blind evaluations by reviewer institution rank. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A47: Decomposing the Non-Blind Score Gap: By reviewer gender

(a) Male reviewer

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.33	-0.70	-0.09	-0.25
	[-0.58,-0.07]	[-1.00,-0.46]	[-0.42,0.25]	[-0.58,0.09]
Conditional on submission content	-0.15	-0.11	-0.17	0.01
	[-0.43,0.26]	[-0.44,0.43]	[-0.56,0.21]	[-0.37,0.48]
	Panel B: Decomposition components			
Submission content	-0.18	-0.59	0.08	-0.26
	[-0.56,0.10]	[-1.15,-0.30]	[-0.31,0.48]	[-0.72,0.13]
Accurate statistical discrimination	0.09	0.00	-0.21	0.10
	[-0.10,0.30]	[-0.20,0.27]	[-0.49,0.02]	[-0.07,0.39]
Inaccurate statistical discrimination	-0.20	0.16	0.26	-0.18
	[-0.52,0.04]	[-0.15,0.54]	[-0.03,0.61]	[-0.56,0.11]
Alternative objectives	-0.47	0.33	-0.04	0.28
	[-0.84,-0.17]	[-0.04,0.77]	[-0.36,0.26]	[-0.02,0.67]
Unexplained	0.44	-0.61	-0.18	-0.19
	[-0.07,1.06]	[-1.25,0.09]	[-0.77,0.40]	[-0.86,0.47]

(b) Female reviewer

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.55	-0.69	-0.25	-0.06
	[-0.86,-0.23]	[-1.03,-0.35]	[-0.59,0.11]	[-0.50,0.33]
Conditional on submission content	-0.45	-0.18	0.04	0.37
	[-0.77,0.04]	[-0.59,0.67]	[-0.36,0.74]	[-0.10,1.24]
	Panel B: Decomposition components			
Submission content	-0.10	-0.51	-0.29	-0.43
	[-0.57,0.18]	[-1.41,-0.22]	[-0.93,0.04]	[-1.23,-0.08]
Accurate statistical discrimination	0.06	0.05	0.05	0.02
	[-0.34,0.46]	[-0.44,0.46]	[-0.84,0.72]	[-0.41,0.53]
Inaccurate statistical discrimination	-0.05	-0.03	-0.14	0.06
	[-0.90,1.04]	[-1.28,0.87]	[-1.14,0.90]	[-0.91,0.89]
Alternative objectives	0.22	-0.25	-0.50	0.58
	[-0.44,1.91]	[-1.25,1.06]	[-2.77,0.27]	[-0.76,1.89]
Unexplained	-0.68	0.05	0.64	-0.30
	[-2.80,0.27]	[-1.52,1.93]	[-0.32,3.08]	[-1.79,1.88]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but subsetting non-blind evaluations by reviewer gender. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A48: Decomposing the Non-Blind Score Gap: Adjusting Citations for Traditionally Low-Scoring groups

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.41	-0.72	-0.12	-0.21
	[-0.61,-0.21]	[-0.94,-0.51]	[-0.36,0.14]	[-0.47,0.06]
Conditional on submission content	-0.28	-0.19	-0.07	0.12
	[-0.51,-0.01]	[-0.44,0.21]	[-0.35,0.26]	[-0.19,0.55]
	Panel B: Decomposition components			
Submission content	-0.13	-0.53	-0.05	-0.33
	[-0.40,0.10]	[-0.98,-0.29]	[-0.38,0.23]	[-0.73,-0.05]
Accurate statistical discrimination	0.11	0.05	-0.12	0.13
	[-0.02,0.30]	[-0.10,0.29]	[-0.32,0.04]	[-0.02,0.36]
Inaccurate statistical discrimination	-0.16	0.01	0.11	-0.16
	[-0.39,0.02]	[-0.26,0.21]	[-0.07,0.31]	[-0.46,0.02]
Alternative objectives	-0.33	0.25	-0.12	0.34
	[-0.59,-0.07]	[-0.07,0.58]	[-0.41,0.14]	[0.08,0.67]
Unexplained	0.09	-0.50	0.07	-0.18
	[-0.31,0.53]	[-0.95,0.06]	[-0.35,0.57]	[-0.74,0.36]

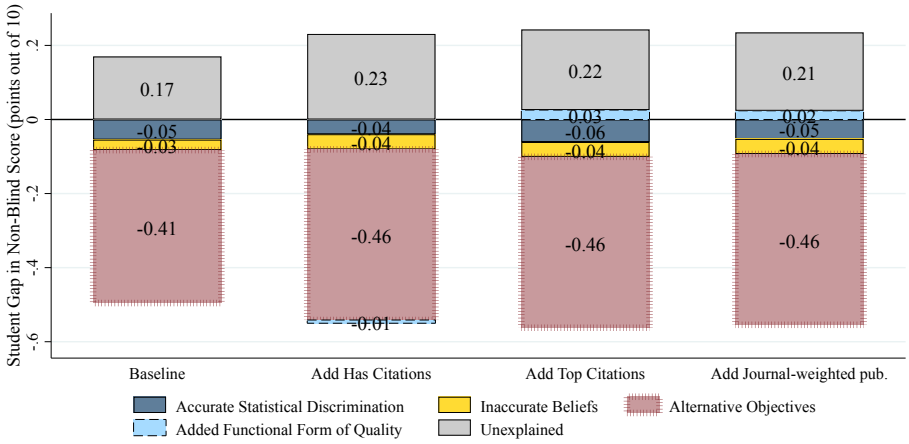
Notes. This table decomposes disparities in non-blind scores, as in Table A43, but considering the case when citations for traditionally lower-scoring groups (students, lower rank institutions, female applicants and PIs) are deflated by 10 percentage points relative to true paper quality, by inflating 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, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

To assess whether non-blind reviewers scoring behavior is consistent with rewarding the highest-quality papers or penalizing the lowest-quality ones, and the extent to which this affects the conclusions of the decomposition, I test whether additional functional forms of citations helps explain the non-blind score disparities beyond the components considered in the decomposition of the main text. To do so, I decompose the student score gap by estimating the realized quality expectation functions using the data from the original experiment, and then using the data in my second experiment to generate predicted values of realized outcomes, and compare whether conclusions change with the inclusion of additional quality measures. I use the data from this second experiment because I did not elicit reviewer beliefs over these alternative functional forms and therefore cannot estimate the coefficients of the score equation in the main experimental data that incorporates these measures. For instance, note that even if I observe beliefs over expected citations ($\mathcal{E}[NCites|x,v]$), I cannot back out beliefs over a submission's likelihood of having at least one citation without imposing additional assumptions on the distribution of beliefs because due to Jensen's inequality:

$\mathbb{1}\{\mathcal{E}[NCites|x,v] > 0\}$ need not equal $\mathcal{E}[\mathbb{1}\{NCites > 0\}|x,v]$. Collecting beliefs data data allows me to instead directly test the extent to which accurate statistical discrimination over these alternative quality measures explain non-blind score gaps. In other words, I fit a regression of realized quality on observables from the original experiment and use it to generate predictions of quality for the sample of the second experiment. I repeat the process for various functional forms of the realized quality, and use these predictions as regressors in the decomposition so that there is no reliance on beliefs over these outcomes.

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. To test reviewers seek to prevent the lowest-quality submissions from being accepted, I use an indicator for whether the paper has at least one citation. Figure A25 summarizes. First, I find that decomposing the non-blind score gap with only the data from the second experiment reproduces the conclusion that alternative objectives explain almost the entirety of the student score gap (left-most bar in Figure A25). Second, including additional functional forms of quality measures do not explain additional variation beyond the measures used in the main analyses (Figure A25). The indicator for having at least one citation explains around 0.01 additional points of the gap, and both the indicator for being in the top decile of citations and journal-weighted publications contribute to the gap in the opposite direction.

Figure A25: Incorporating Additional Functional Forms of Quality



Notes. This figure decomposes the student - non-student disparity in non-blind scores into the contribution that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Observation is at the paper-level, using the sample of papers from the second experiment.

Table A49: Interacting submission content with author traits (student and institution rank)

	Panel A: Student									
	Non-Blind	Realized Quality			Quality Beliefs			Alternative Objectives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Score	Cites	Online	Pub	Cites	Online	Pub	Talk	BConf	BAuth
Submission Content	1.47***	33.09**	0.26**	0.22**	10.25***	0.07**	0.05	0.68***	0.50***	0.46***
	(0.52)	(16.66)	(0.11)	(0.10)	(3.33)	(0.03)	(0.05)	(0.15)	(0.14)	(0.15)
Submission Content × Student	-0.64	-22.34	-0.16	-0.07	-3.17	-0.02	0.08	-0.23	-0.23	-0.29*
	(0.54)	(17.05)	(0.11)	(0.11)	(3.59)	(0.04)	(0.06)	(0.16)	(0.16)	(0.17)
Student	3.50	130.59	0.99	0.52	13.58	0.11	-0.44	1.04	1.21	1.49
	(3.17)	(99.38)	(0.67)	(0.67)	(21.17)	(0.24)	(0.35)	(0.97)	(0.97)	(1.01)
Lower Rank Inst.	-0.09	0.92	0.01	0.05	2.15	0.06*	0.01	0.17	0.28*	0.21
	(0.23)	(6.61)	(0.05)	(0.06)	(2.92)	(0.04)	(0.06)	(0.15)	(0.15)	(0.16)
Female	-0.10	-11.33***	-0.03	-0.08	-0.79	0.00	-0.01	-0.16	-0.01	0.12
	(0.17)	(4.29)	(0.05)	(0.05)	(2.73)	(0.03)	(0.05)	(0.13)	(0.13)	(0.13)
Has Female PI	0.12	2.00	0.05	0.06	0.52	-0.01	-0.06	0.36***	0.17	0.02
	(0.20)	(4.78)	(0.04)	(0.06)	(2.66)	(0.04)	(0.06)	(0.14)	(0.13)	(0.15)
Subfield FE	×	×	×	×	×	×	×	×	×	×
Outcome mean		25.79	0.79	0.55	23.61	0.91	0.70	2.93	3.67	3.40
First stage F-stat	3.27	3.30	3.30	3.30	17.32	17.90	17.90	17.73	17.73	17.73
N	2530	1290	1290	1290	1151	1226	1226	1224	1224	1224
N Papers	645	645	645	645	383	408	408	407	407	407
	Panel B: Institution Rank									
	Non-Blind	Realized Quality			Quality Beliefs			Alternative Objectives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Score	Cites	Online	Pub	Cites	Online	Pub	Talk	BConf	BAuth
Submission Content	1.16***	29.12*	0.23**	0.24**	11.41***	0.13***	0.13**	0.64***	0.41***	0.41***
	(0.44)	(15.06)	(0.10)	(0.10)	(2.97)	(0.04)	(0.05)	(0.14)	(0.14)	(0.14)
Submission Content × Lower Ranked Inst.	-0.17	-17.25	-0.12	-0.10	-5.09	-0.12***	-0.06	-0.16	-0.08	-0.20
	(0.48)	(16.02)	(0.10)	(0.11)	(3.72)	(0.04)	(0.06)	(0.17)	(0.16)	(0.17)
Student	-0.26*	0.77	0.08**	0.09**	-4.62*	-0.01	0.02	-0.30**	-0.15	-0.20
	(0.16)	(4.47)	(0.04)	(0.05)	(2.39)	(0.03)	(0.05)	(0.12)	(0.12)	(0.13)
Lower Rank Inst.	0.82	99.57	0.70	0.63	32.77	0.78***	0.36	1.10	0.72	1.41
	(2.93)	(95.16)	(0.64)	(0.70)	(22.82)	(0.29)	(0.40)	(1.03)	(1.03)	(1.04)
Female	-0.10	-11.14***	-0.03	-0.07	-0.88	0.01	-0.00	-0.17	-0.03	0.10
	(0.17)	(4.20)	(0.05)	(0.05)	(2.78)	(0.03)	(0.05)	(0.13)	(0.12)	(0.13)
Has Female PI	0.14	3.16	0.06	0.07	-0.12	-0.02	-0.06	0.34**	0.16	0.01
	(0.20)	(5.26)	(0.05)	(0.06)	(2.70)	(0.04)	(0.06)	(0.13)	(0.13)	(0.16)
Subfield FE	×	×	×	×	×	×	×	×	×	×
Outcome mean		25.79	0.79	0.55	23.61	0.91	0.70	2.93	3.67	3.40
First stage F-stat	3.11	3.34	3.34	3.34	14.71	18.09	18.09	19.01	19.01	19.01
N	2530	1290	1290	1290	1151	1226	1226	1224	1224	1224
N Papers	645	645	645	645	383	408	408	407	407	407

Notes. This table shows subgroup differences in realized quality and reviewers' beliefs about paper quality and alternative objectives as in Table 6, but adds interaction terms between submission content and applicant student status and institution rank. See Table A50 for interactions with applicant and PI gender. Observations for odd-numbered columns are at the paper-level, with heteroskedastic robust standard errors in parentheses. Submission content is proxied by a paper's blind scores: to account for noise in blind scores, a given blind review score for a paper is instrumented by the average score that the paper's other blind reviewers gave, as described in Section 5.1. Standard errors are clustered at the paper level. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A50: Interacting submission content with author traits (applicant and PI gender)

	Panel C: Applicant Gender									
	Non-Blind	Realized Quality			Quality Beliefs			Alternative Objectives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Score	Cites	Online	Pub	Cites	Online	Pub	Talk	BConf	BAuth
Submission Content	1.05***	20.80***	0.17***	0.19***	10.66***	0.06**	0.09**	0.61***	0.44***	0.32***
	(0.22)	(7.47)	(0.05)	(0.05)	(2.46)	(0.03)	(0.04)	(0.11)	(0.11)	(0.12)
Submission Content × Applicant Female	-0.02	-9.14	-0.05	-0.04	-5.60	0.01	0.02	-0.17	-0.18	-0.09
	(0.40)	(8.60)	(0.08)	(0.10)	(3.72)	(0.04)	(0.06)	(0.16)	(0.16)	(0.16)
Student	-0.28*	-0.47	0.07**	0.08**	-4.83**	-0.01	0.02	-0.30**	-0.14	-0.20
	(0.14)	(4.01)	(0.03)	(0.04)	(2.39)	(0.03)	(0.05)	(0.12)	(0.12)	(0.13)
Lower Rank Inst.	-0.19	-1.88	-0.01	0.05	2.31	0.06	0.02	0.15	0.27*	0.17
	(0.18)	(5.17)	(0.04)	(0.05)	(2.94)	(0.04)	(0.06)	(0.15)	(0.15)	(0.16)
Female	0.01	41.82	0.26	0.15	32.70	-0.05	-0.15	0.86	1.04	0.61
	(2.34)	(49.33)	(0.47)	(0.57)	(22.41)	(0.25)	(0.36)	(0.93)	(0.96)	(0.94)
Has Female PI	0.13	2.18	0.06	0.07	0.49	-0.01	-0.06	0.35***	0.17	0.02
	(0.19)	(4.64)	(0.04)	(0.06)	(2.69)	(0.04)	(0.06)	(0.14)	(0.13)	(0.15)
Subfield FE	×	×	×	×	×	×	×	×	×	×
Outcome mean		25.79	0.79	0.55	23.61	0.91	0.70	2.93	3.67	3.40
First stage F-stat	9.94	9.91	9.91	9.91	19.95	24.89	24.89	25.76	25.76	25.76
N	2530	1290	1290	1290	1151	1226	1226	1224	1224	1224
N Papers	645	645	645	645	383	408	408	407	407	407
	Panel D: PI Gender									
	Non-Blind	Realized Quality			Quality Beliefs			Alternative Objectives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Score	Cites	Online	Pub	Cites	Online	Pub	Talk	BConf	BAuth
Submission Content	1.02***	16.79***	0.13***	0.17***	8.72***	0.06***	0.11***	0.53***	0.32***	0.25**
	(0.20)	(6.15)	(0.04)	(0.05)	(2.10)	(0.02)	(0.03)	(0.09)	(0.09)	(0.10)
Submission Content × PI Female	0.21	11.25	0.18	0.01	-1.61	0.02	-0.06	0.15	0.34*	0.27
	(0.63)	(15.96)	(0.17)	(0.13)	(4.20)	(0.05)	(0.09)	(0.19)	(0.19)	(0.18)
Student	-0.27*	-0.11	0.08**	0.08**	-5.07**	-0.01	0.02	-0.32**	-0.16	-0.21*
	(0.15)	(4.12)	(0.04)	(0.04)	(2.44)	(0.03)	(0.05)	(0.12)	(0.12)	(0.13)
Lower Rank Inst.	-0.18	-2.23	-0.01	0.04	1.39	0.06*	0.01	0.15	0.28*	0.19
	(0.18)	(5.09)	(0.04)	(0.05)	(2.85)	(0.04)	(0.06)	(0.15)	(0.15)	(0.16)
Female	-0.11	-11.34***	-0.03	-0.08	-0.99	0.00	-0.01	-0.17	-0.02	0.10
	(0.16)	(3.81)	(0.04)	(0.05)	(2.76)	(0.03)	(0.05)	(0.13)	(0.12)	(0.13)
Has Female PI	-1.05	-61.16	-0.94	0.01	9.91	-0.14	0.28	-0.49	-1.83	-1.56
	(3.56)	(89.25)	(0.94)	(0.74)	(24.25)	(0.30)	(0.50)	(1.10)	(1.13)	(1.05)
Subfield FE	×	×	×	×	×	×	×	×	×	×
Outcome mean		25.79	0.79	0.55	23.61	0.91	0.70	2.93	3.67	3.40
First stage F-stat	2.09	3.34	3.34	3.34	14.25	13.61	13.61	17.52	17.52	17.52
N	2530	1290	1290	1290	1151	1226	1226	1224	1224	1224
N Papers	645	645	645	645	383	408	408	407	407	407

Notes. This table shows subgroup differences in realized quality and reviewers' beliefs about paper quality and alternative objectives as in Table 6, but adds interaction terms between submission content and applicant/PI gender. See Table A49 for interactions with applicant student status and institution rank. Observations for odd-numbered columns are at the paper-level, with heteroskedastic robust standard errors in parentheses. Submission content is proxied by a paper's blind scores: to account for noise in blind scores, a given blind review score for a paper is instrumented by the average score that the paper's other blind reviewers gave, as described in Section 5.1. Standard errors are clustered at the paper level. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A51: Decomposing Non-Blind Score Gaps: Interactions

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	-0.26	-0.17	-0.11	0.17
	[-0.51,0.19]	[-0.49,0.77]	[-0.42,0.27]	[-0.36,1.20]
	Panel B: Decomposition components			
Submission content	-0.19	-0.56	-0.05	-0.36
	[-0.71,0.07]	[-1.52,-0.28]	[-0.41,0.26]	[-1.39,0.14]
Accurate statistical discrimination	0.09	0.02	-0.18	0.13
	[-0.05,0.35]	[-0.14,0.51]	[-0.41,0.01]	[-0.11,0.57]
Inaccurate statistical discrimination	-0.14	0.07	0.17	-0.17
	[-0.44,0.04]	[-0.43,0.36]	[-0.04,0.42]	[-0.67,0.10]
Alternative objectives	-0.32	0.30	-0.09	0.36
	[-0.60,-0.03]	[-0.04,0.69]	[-0.39,0.17]	[0.07,0.74]
Unexplained	0.11	-0.56	-0.01	-0.15
	[-0.34,0.70]	[-1.18,0.40]	[-0.47,0.54]	[-0.93,0.89]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but including interaction terms between the submission trait of interest and submission content (Estimation regressions are in Tables A49 and A50). The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A52: Decomposing the Non-Blind Score Gap: Without Proxying for Submission Content

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
	Panel B: Decomposition components			
Accurate statistical discrimination	0.03	-0.19	-0.19	-0.02
	[-0.12,0.16]	[-0.43,-0.01]	[-0.43,-0.02]	[-0.19,0.14]
Inaccurate statistical discrimination	-0.07	0.17	0.22	-0.07
	[-0.25,0.09]	[-0.05,0.42]	[0.02,0.49]	[-0.30,0.11]
Alternative objectives	-0.37	-0.04	0.01	0.15
	[-0.65,-0.11]	[-0.37,0.25]	[-0.25,0.29]	[-0.12,0.45]
Unexplained	-0.03	-0.66	-0.20	-0.25
	[-0.42,0.35]	[-1.06,-0.21]	[-0.60,0.22]	[-0.70,0.21]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but conditional on median submission content. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A53: Decomposing the Non-Blind Score Gap: At median submission content

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	-0.36	-0.20	-0.11	0.20
	[-0.73,-0.01]	[-0.50,0.33]	[-0.42,0.35]	[-0.42,1.68]
	Panel B: Decomposition components			
Submission content	-0.09	-0.53	-0.05	-0.40
	[-0.42,0.30]	[-1.12,-0.25]	[-0.48,0.26]	[-2.00,0.20]
Accurate statistical discrimination	0.04	-0.03	-0.20	0.16
	[-0.19,0.24]	[-0.20,0.25]	[-0.45,0.01]	[-0.12,0.82]
Inaccurate statistical discrimination	-0.08	0.10	0.19	-0.21
	[-0.31,0.14]	[-0.21,0.36]	[-0.03,0.46]	[-0.91,0.14]
Alternative objectives	-0.38	0.27	-0.13	0.41
	[-0.67,-0.10]	[-0.08,0.65]	[-0.42,0.14]	[0.13,0.82]
Unexplained	0.06	-0.54	0.03	-0.16
	[-0.46,0.59]	[-1.10,0.12]	[-0.42,0.63]	[-1.14,1.25]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but conditional on median submission content. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A54: Decomposing the Non-Blind Score Gap: Above or Below Median in Submission Content

(a) Below median

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	1.04	0.17	-0.07	-0.25
	[-0.50,7.35]	[-1.51,6.91]	[-2.54,1.41]	[-7.58,3.08]
	Panel B: Decomposition components			
Submission content	-1.49	-0.90	-0.09	0.06
	[-7.83,0.06]	[-7.72,0.76]	[-1.60,2.36]	[-3.29,7.40]
Accurate statistical discrimination	0.78	0.61	0.10	-0.27
	[-0.06,4.32]	[-0.19,4.27]	[-0.67,0.96]	[-3.47,1.01]
Inaccurate statistical discrimination	-0.84	-0.23	-0.03	0.32
	[-4.30,0.07]	[-3.69,0.77]	[-0.98,0.91]	[-1.23,3.51]
Alternative objectives	0.34	0.67	0.38	-0.31
	[-0.38,1.45]	[-0.07,1.88]	[-0.46,1.29]	[-1.64,0.57]
Unexplained	0.76	-0.88	-0.51	0.02
	[-1.38,7.07]	[-3.51,6.34]	[-3.01,1.52]	[-7.78,4.67]

(b) Above median

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.65,-0.24]	[-0.96,-0.52]	[-0.39,0.09]	[-0.46,0.08]
Conditional on submission content	-1.50	-0.50	-0.14	0.57
	[-7.03,0.00]	[-5.47,1.11]	[-1.59,2.11]	[-2.80,8.88]
	Panel B: Decomposition components			
Submission content	1.05	-0.23	-0.01	-0.77
	[-0.42,6.65]	[-1.80,4.67]	[-2.35,1.44]	[-9.10,2.57]
Accurate statistical discrimination	-0.56	-0.54	-0.44	0.52
	[-3.64,0.20]	[-3.16,0.22]	[-1.45,0.37]	[-0.74,4.23]
Inaccurate statistical discrimination	0.54	0.36	0.37	-0.64
	[-0.27,3.28]	[-0.46,2.77]	[-0.43,1.33]	[-4.26,1.00]
Alternative objectives	-0.96	-0.06	-0.54	1.00
	[-1.98,-0.29]	[-0.95,0.53]	[-1.36,0.17]	[0.22,2.28]
Unexplained	-0.51	-0.26	0.47	-0.31
	[-5.96,1.53]	[-5.08,1.79]	[-1.37,3.08]	[-6.07,8.23]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but subsetting non-blind evaluations by whether submission content is (a) below or (b) above median. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

G Submission Text

In this section, I consider the observables and text in the 300 word description and 2 page document that applicants submit. I start first with the characteristics of the 2-page document associated with the sample of submissions from the main sample, to test whether the relationship between scores and particular submission traits changed due to blinding. I first compute a measure of text difficulty, the Flesch-Kincaid readability measure (Flesch, 1948; Kincaid et al., 1975) using the textstat python package (see for example Hengel (2022) who use this measure for economics papers to test whether women are held to higher writing quality standards than men). This measure calculates the number of words per sentence and syllables per word, and is intended to capture the grade-level necessary to understand the text, meaning that higher values correspond to harder-to-read text. Table A55a gives summary statistics, and Table A55b presents correlations with author traits. Table A56 shows the differential impacts of blinding along these dimensions.

Table A55: Submission Traits

(a) Submission characteristics

	(1)
Variable	All
Text Difficulty	12.40
	(2.63)
N Figures in submission	1.96
	(1.19)
N Words in submission	1078.97
	(362.94)
Observations	657

(b) Correlation With Author Traits

	Text Difficulty	N Figures	N Words
	(1)	(2)	(3)
Student	-0.14	-0.27***	7.17
	(0.21)	(0.09)	(29.10)
Lower Rank Inst.	-0.04	-0.12	-34.29
	(0.23)	(0.10)	(31.24)
Female	-0.04	-0.16	19.39
	(0.25)	(0.11)	(34.20)
Has Female PI	0.60**	0.01	-24.09
	(0.28)	(0.12)	(38.18)
Sample Mean	12.40	1.96	1078.97
N	656	656	656
R^2	0.03	0.07	0.03

Notes. This table presents (a) the average characteristics across all submissions' 2-page documents from the main experiment, and (b) correlates author traits with submission observables. Dependent variable indicated by column headers. I (b), text difficulty and word count are standardized to be mean zero and standard deviation one. Regressions control for subfield.

Table A56: Accounting for submission observables

	Non-blind	Blind	All
	(1)	(2)	(3)
Student	-0.47***	-0.13	
	(0.11)	(0.12)	
Lower Rank Inst.	-0.75***	-0.55***	
	(0.12)	(0.13)	
Female	-0.25	-0.02	
	(0.16)	(0.13)	
Has Female PI	-0.27*	-0.32**	
	(0.15)	(0.15)	
Text Difficulty	-0.02	-0.02	
	(0.02)	(0.02)	
N Figures in submission	0.09	0.17***	
	(0.06)	(0.06)	
N Words in submission (std)	0.13	0.24***	
	(0.09)	(0.08)	
Student × Blind			0.34***
			(0.13)
Lower Rank Inst. × Blind			0.30**
			(0.15)
Female × Blind			0.27*
			(0.16)
PI Female × Blind			-0.06
			(0.20)
Text Difficulty × Blind			0.02
			(0.03)
N Figures × Blind			0.07
			(0.07)
N Words × Blind			0.08
			(0.07)
N	1287	1300	2587
R^2	0.19	0.17	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. Regressions control for subfield. Standard errors in parentheses, clustered at the reviewer level. * $p < .1$, ** $p < .05$, *** $p < .01$

To understand whether these text-based information affect conclusions from the decomposition, I turn to the text from abstracts, which is available for both the main experiment and the second. I again calculate text difficulty, and I implement tools from the natural language processing literature to generate additional controls for submission content beyond the blind scores that I use in the decomposition of the main text. This approach of comparing disaparities after including text embedding controls mirrors Avivi (2024) who tests for gender bias in patent examiners by conditioning on controls generated from patent texts.

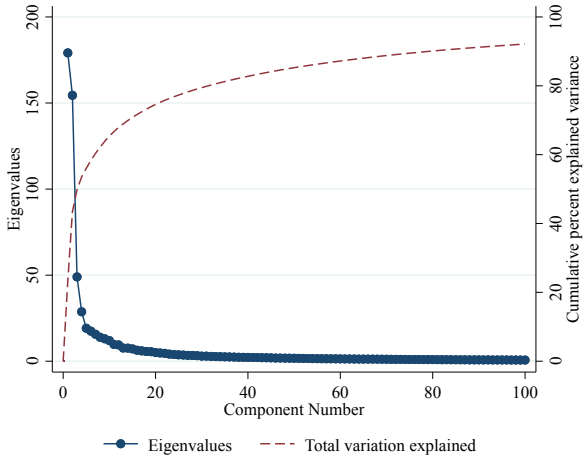
I use a pretrained sentence-embedding model to generate a 768-dimensional representation of each

abstract. Extracting embedding vectors from the text data is preferable over other traditional methods such as a bag-of-words approaches based on word counts, because the former extracts meaning from relationships among sequences of words rather than relying solely on frequncies (see Gentzkow et al. (2019) and Ash and Hansen (2023) for reviews). To generate these text embeddings, I use the SPECTER2 transformer-based text embedding model[40] that is designed specifically for scientific documents,[41] and trained on document data spanning 23 academic fields, including computer science and biology (Singh et al., 2022; Kinney et al., 2023). I use the SPECTER2 regression adapter, which produces encodings that are intended for use as linear regressors.

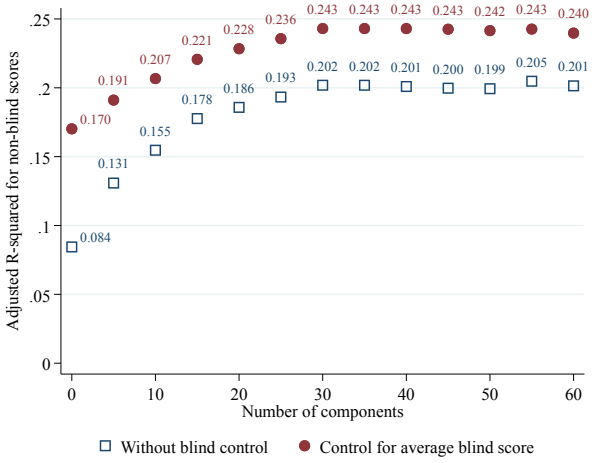
The langugage model encodes each abstract into a 768-dimensional vector. To avoid colinearity or overfitting, I use a dimension-reduction approach to address this high-dimensionality. I use principal component analysis to construct orthgonal linear combinations of the original text embeddings. This approach is attractive because it allows me to retain the principal components that capture most of the variation in the text data while including a smaller set of controls to include in each step of the decomposition. Figure A26a shows a scree plot, which illustrate the eigenvalues associated with each component number and the total variance explained. Figure A26b shows the variation in non-blind scores explained with each component number added as controls, by whether a control for the paper's average blind score. Below I show results that use the first 30 principal components, but results are not sensitive to this choice.

Figure A26: Variation Explained by Text Embedding Principal Components

(a) Scree Plot



(b) Variation explained in non-blind scores



Notes. This figure shows the (a) eigenvalues and percent of total variation in text embeddings and (b) variation in non-blind scores explained by each of the principal components for text embeddings generated using abstract text. For (b), the baseline specification regresses, at the paper-reviewer level among non-blind reviewers, non-blind scores on author traits, subfield fixed effects, and reviewer fixed effects.

Table A57 shows the non-blind score gap conditional on abstract text embeddings. I find that relative to the model that only includes author traits and subfield fixed effects which explains around 7 percent of the

[40] https://github.com/allenai/SPECTER2

[41] https://allenai.org/blog/specter2-adapting-scientific-document-embeddings-to-multiple-fields-and-task-formats-c95686c06567

variation in non-blind scores, adding in the abstract text embeddings explains around another 9 percentage points.

Table A57: Non-Blind Score Gap Conditional on Submission Text

	(1)	(2)	(3)	(4)	(5)	(6)
Student	-0.46***	-0.45***	-0.45***	-0.41***	-0.41***	-0.41***
	(0.12)	(0.12)	(0.12)	(0.11)	(0.11)	(0.11)
Lower Rank Inst.	-0.73***	-0.58***	-0.57***	-0.55***	-0.47***	-0.46***
	(0.13)	(0.13)	(0.13)	(0.13)	(0.12)	(0.12)
Female	-0.18	-0.13	-0.13	-0.18	-0.15	-0.15
	(0.14)	(0.14)	(0.14)	(0.13)	(0.13)	(0.13)
Has Female PI	-0.26	-0.20	-0.20	-0.16	-0.13	-0.14
	(0.16)	(0.16)	(0.16)	(0.15)	(0.16)	(0.15)
Text difficulty			0.03			0.03
			(0.03)			(0.03)
Blind score (average)				0.35***	0.26***	0.26***
				(0.04)	(0.04)	(0.04)
Submission text		×	×		×	×
Subfield FE	×	×	×	×	×	×
N	1289	1289	1289	1289	1289	1289
N Papers	657	657	657	657	657	657
R^2	0.07	0.19	0.19	0.14	0.22	0.22

Notes. This table shows subgroup differences in non-blind scores. "Submission text" refers to inclusion of controls for the principal components of text embeddings generated using the abstract text. Observation is at the paper-reviewer level, among non-blind reviewers.

I then add these controls to each component of the decomposition. For example, the quality expectation function is estimated by regressing the quality measure of interest on author traits, the principal components of the text embeddings, and blind scores using the instrumental variables approach discussed in the main text. Table A58 shows the decomposition results. Overall, I find that the conclusions are largely unchanged: majority of the student score gap is explained by alternative objectives.

Table A58: Decomposing the Non-Blind Score Gap: Controlling for Submission Text Observables

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.46	-0.59	-0.08	-0.28
	[-0.65,-0.22]	[-0.97,-0.52]	[-0.41,0.10]	[-0.46,0.09]
Conditional on submission content	-0.25	-0.06	-0.10	0.09
	[-0.51,0.08]	[-0.41,0.48]	[-0.44,0.29]	[-0.17,0.70]
	Panel B: Decomposition components			
Submission content	-0.21	-0.53	0.01	-0.36
	[-0.52,0.07]	[-1.25,-0.35]	[-0.47,0.28]	[-0.90,-0.04]
Accurate statistical discrimination	0.11	0.09	-0.19	0.14
	[-0.06,0.29]	[-0.13,0.32]	[-0.41,0.03]	[-0.05,0.36]
Inaccurate statistical discrimination	-0.15	-0.05	0.17	-0.18
	[-0.37,0.05]	[-0.27,0.27]	[-0.07,0.41]	[-0.45,0.06]
Alternative objectives	-0.31	0.23	-0.22	0.43
	[-0.61,-0.05]	[-0.07,0.63]	[-0.44,0.15]	[0.10,0.74]
Unexplained	0.09	-0.33	0.14	-0.31
	[-0.34,0.61]	[-0.98,0.26]	[-0.42,0.62]	[-0.75,0.46]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but adding controls for submission text using the for the principal components of text embeddings generated using the abstract text. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

Table A59: Decomposing the Non-Blind Score Gap: Controlling for Submission Text Observables and not blind scores

	Magnitude in points (out of 10)			
	Student	Institution	Gender	PI Gender
	Panel A: Non-Blind score gap			
Unconditional	-0.45	-0.73	-0.16	-0.20
	[-0.64,-0.23]	[-0.97,-0.52]	[-0.40,0.08]	[-0.46,0.08]
Conditional on submission content	-0.46	-0.61	-0.11	-0.17
	[-0.65,-0.22]	[-0.97,-0.51]	[-0.40,0.10]	[-0.45,0.09]
	Panel B: Decomposition components			
Submission content	0.01	-0.12	-0.04	-0.02
	[-0.05,0.05]	[-0.05,0.05]	[-0.06,0.05]	[-0.07,0.06]
Accurate statistical discrimination	0.04	-0.14	-0.16	-0.04
	[-0.11,0.17]	[-0.45,-0.01]	[-0.43,-0.01]	[-0.19,0.15]
Inaccurate statistical discrimination	-0.07	0.11	0.20	-0.06
	[-0.25,0.09]	[-0.06,0.43]	[0.01,0.50]	[-0.32,0.12]
Alternative objectives	-0.31	0.04	0.01	0.24
	[-0.64,-0.08]	[-0.38,0.26]	[-0.25,0.30]	[-0.14,0.48]
Unexplained	-0.11	-0.63	-0.16	-0.32
	[-0.43,0.34]	[-1.08,-0.18]	[-0.64,0.22]	[-0.76,0.23]

Notes. This table decomposes disparities in non-blind scores, as in Table A43, but adding controls for submission text using the for the principal components of text embeddings generated using the abstract text. The decomposition shows the contribution in non-blind score gaps that is attributable to accurate statistical discrimination, inaccurate statistical discrimination, alternative objectives, and additional functional forms of quality, after controlling for submission content (see Equation 5). Estimation steps are described in Section 5.1. For a given author characteristic (a column), the sum of the components (Panel B estimates) equals the unconditional score gap. Accurate statistical discrimination summarizes differences in paper quality measured by number of citations, being available online, being published, 5 years later. Inaccurate statistical discrimination summarizes reviewer beliefs over the same measures. Alternative objectives captures differences in reviewer beliefs over how engaging the talk would be, the extent to which the conference would benefit from accepting the applicant, and the extent to which the authors would benefit from being accepted. 90% confidence intervals in brackets, based on 1000 clustered bootstrap replications that are drawn at the paper level.

H 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, although both knew that the conference would use both blind and non-blind review. Whether applicants' choices to submit change by knowing that an evaluation is blind or not is out of the scope of this paper (see Boring et al. (2025)). 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 A11). 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 A27 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 five 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 show in Section C.9 that this is not driving the results.

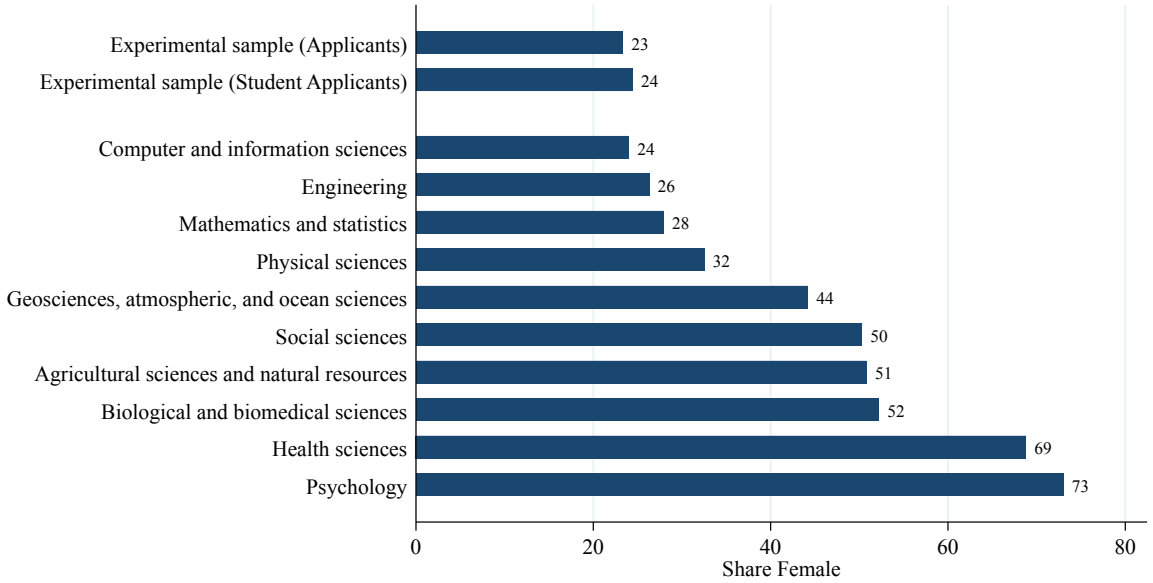
Concerning naturalness of the choice task, setting, and framing, I use a natural field experiment (Harrison and List, 2004), such that the task at hand for both applicants and reviewers was not artificial. Previous years of the conference used only non-blind review, making blinding a potentially novel 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 paper illustrates a method on disentangling sources of discrimination that can be carried over to settings beyond the one studied in this paper. Whether and how the results from this study translate to other contexts likely depends on numerous factors. For instance, one important moderator may be the gender composition of the setting: relative of other academic settings, the experimental sample appears to be more gender-imbalanced than other fields (Figure A27), and this may affect the underlying levels of subgroup quality differences (which determines the scope for accurate statistical discrimination), evaluator beliefs, 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.

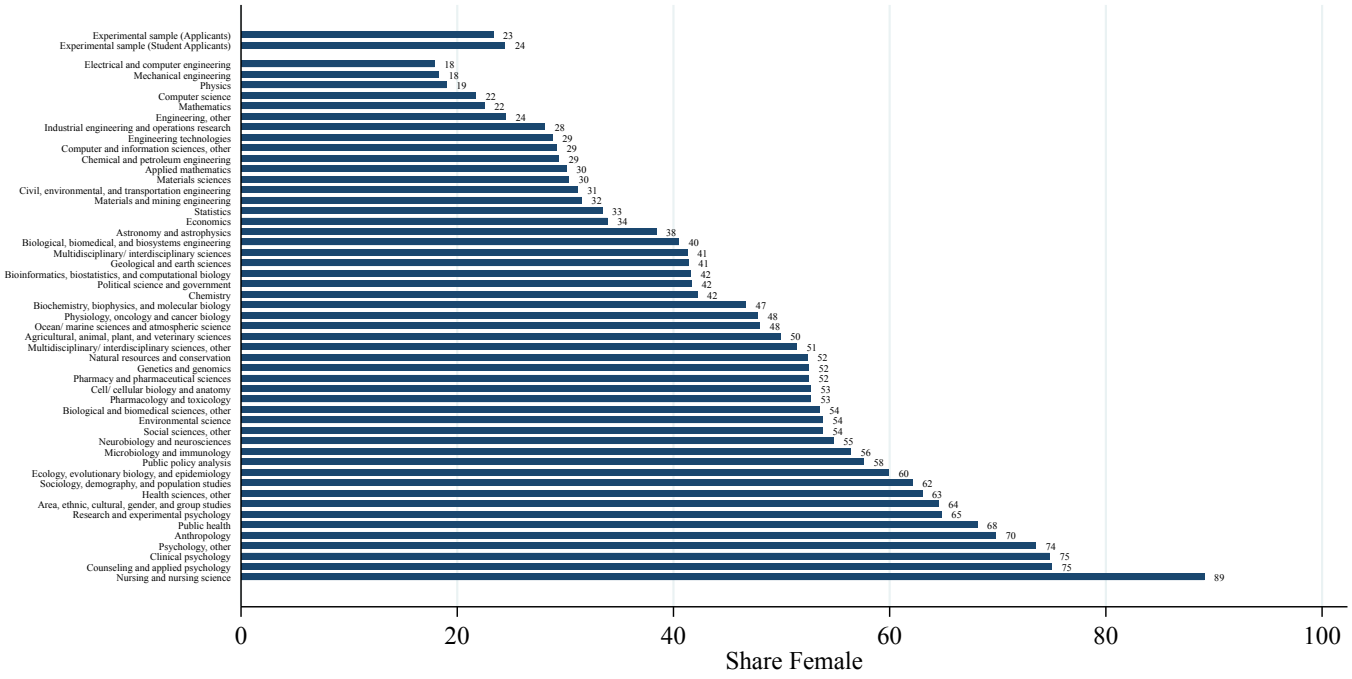
Ultimately, the goal of this paper is not to characterize disparities across all contexts but to offer a "WAVE1" insight, in the nomenclature of (List, 2020), to present a novel experimental design that can uncover the intricacies of a policy. Future work can extend this approach to other settings.

Figure A27: 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).