

WORKING PAPER · NO. 2024-07

Blinded Versus Unblinded Review: A Field Study Comparing the Equity of Peer-Review

Timothy J. Pleskac, Ellie Kyung, Gretchen Chapman, and Oleg Urminsky

JANUARY 2024

Blinded versus unblinded review: A field study comparing the equity of peer-review

Timothy J. Pleskac

University of Kansas

Ellie Kyung

Babson College

Gretchen Chapman

Carnegie Mellon University

Oleg Urminsky

University of Chicago

Abstract

Blinded review is often proposed as a solution to potential inequities in merit-based evaluations. Yet, an equitable merit-based evaluation should also be reliable and valid. We evaluated the equity of blinded and unblinded review in a high-stakes field experiment during peer-review of submissions for an academic conference. Each conference submission received double-blind peer-reviews with reviewer and author identities withheld and single-blind peer-reviews with only reviewer identities withheld. We found that both systems yielded evaluations with moderate reliability after averaging across three or more independent reviewers. This level of reliability resulted in the review systems agreeing on 40% of the top submissions. This agreement level suggests that some differences between review systems emerge from noisy human judgment. The comparison also revealed that single-blind reviews favored senior coauthors and disfavored Asian (versus White) authors more than double-blind review. Double-blind reviews slightly favored male authors more than single-blind reviews. For submissions selected for presentation, neither author characteristics nor the review process consistently predicted judged talk quality or popularity. Single- and double-blind review scores similarly predicted subsequent publication. These results imply that preferences based on authors' characteristics during single-blind review were not based on well-calibrated beliefs, and there was no strong evidence to suggest that knowing the authors' characteristics would improve peer-review. The results also suggest that double-blind review may also reduce the ability of reviewers to take into consideration when submissions are from under-represented groups. We discuss ways blind review may be adapted to create an equitable merit-based evaluation.

Keywords: peer review, double blind, single blind, field study, metascience

Introduction

Blind review—be it for job, credit or loan applications, legal procedures, or scientific peer review—is often seen as an effective means to help create an equitable merit-based evaluation system (Robertson & Kesselheim, 2016; MacCoun, 2020). Does blinding help achieve this goal? For hiring decisions, blind review sometimes seems to create a more equitable selection process, helping women and minorities advance in the interview process (Åslund & Skans, 2012; Goldin & Rouse, 2000). In the context of peer review, *double-blind review*—where both reviewer and author identities are withheld—has been associated with an increase in the representation of women (Budden et al., 2008) and scientists from less prestigious institutions or organizations (Crane, 1967; Tomkins et al., 2017). In other cases, however, blinding has not resulted in broader representation. In some job selection settings, female applicants fared *worse* with blinded applications compared to non-anonymous applications (Behaghel et al., 2015; Krause et al., 2012). Similarly, in the domain of peer review, double-blind review does not always promote better representation compared to single-blind review (Cox & Montgomerie, 2019; Hammerschmidt et al., 2008; Whittaker, 2008). These mixed results are also present in the few randomized experiments on blind review (Blank, 1991; Okike et al., 2016).

The degree to which a merit-based evaluation differentially favors certain groups or individuals is only one part of the equity question concerning peer-review or any merit-based evaluation system.¹ We contend an equitable merit-based evaluation should also be a valid and reliable predictor of merit. Does anonymous evaluation come at the cost of less accurate predictions of quality? Indeed, in discussions about whether to blind reviewers, some may argue that partiality towards specific identities might be justified if those author characteristics are valid proxies of merit. Such an argument would be consistent with a belief- or statistical-based model of discrimination (Aigner & Cain, 1977; Arrow, 1973; Bohren et al., 2019; H. Fang & Moro, 2011; Phelps, 1972). In this case, if the beliefs used to make the evaluation reflect accurate statistical associations, then blinding judges could make merit-based evaluations less valid and potentially less reliable. Alternatively,

¹The construct of equity can have a broader scope, including whether people from different backgrounds may face different barriers, obstacles, and opportunities. Here, we focus on the narrower question of peer-review equity, focusing on whether or not a particular group is advantaged or disadvantaged due to the disclosure of author characteristics during the review process.

The preregistration for the study can be found on OSF [here](#). De-identified survey data and analyses are at the Open Science Framework: [OSF Anonymous Peer Review Link](#). Due to identifiability concerns, we can not make the submission data publicly available. We thank Matthew Paronto for helping administer the study, Bob Burnham for programming the submission system, Abigail Bergman for helping collect the publication data, Deepti Poluru for assistance with coding, the anonymous faculty, post-docs, and graduate students who evaluated the talks, and the undergraduate research assistants that diligently collected data on the talks. We also thank the Executive Board of the Society for Judgment and Decision Making for the permission to run this study. Our gratitude goes to Drs. Donna Ginther, Ralph Hertwig, and Thomas Wallsten for critical feedback on early drafts of this paper. This project was funded by the National Science Foundation MMS 1824259 to E.K. & T.J.P. Author contributions: Conceptualization: E.K., G.C., O.U., & T.J.P.; Investigation: E.K., G.C., O.U., & T.J.P.; Formal analysis: T.J.P.; Writing: E.K., G.C., O.U., & T.J.P. Correspondence concerning this article should be addressed to Timothy J. Pleskac (e-mail: pleskac@ku.edu). *Version 7.0, 03/07/2022.*

when the beliefs are inaccurate (Bordalo et al., 2016; Fryer et al., 2008; Judd & Park, 1993), blinding reviewers should not reduce the validity and reliability of the evaluations. Blinding reviewers might even yield evaluations with greater validity and reliability if reviewers then put more weight on more accurate cues. A third possibility is that partiality towards particular characteristics is preference-based, arising from in-group preferences or unfavorable attitudes towards people with a particular characteristic (Becker, 1957). Here again, blinding—assuming preference is not related to a valid statistical association—should not affect the validity and reliability of the evaluations. In sum, a complete assessment of the equity of a merit-based evaluation system requires three components: an evaluation of the reliability of the system, an evaluation of the degree to which individuals and groups are treated differently (differential impact of characteristics), and an evaluation of the validity of the system. In the context of deciding whether to blind reviewers or not, there are no published results on the impact blinding has on the reliability and validity of merit-based evaluations. We set out to address this limitation.

We report a unique large-scale pre-registered field experiment with conference submissions for the Society for Judgment and Decision Making’s (SJDM) 39th Annual Conference. The setting allowed us to use a unique experimental design, in which each reviewer was randomly assigned to review submissions in either a single-blind review system or a double-blind review system. Each submission was then randomly assigned to reviewers such that at least three double-blind and three single-blind reviewers reviewed each submission. This repeated-measures review design allowed us to control for author and submission characteristics in the single- vs. double-blind comparison. We compared single- and double-blind review of the submissions on three criteria: reliability, the differential impact of author characteristics, and validity. A head-to-head comparison between these different review systems in situ using the data we collected is informative regarding all three criteria. Establishing the differential impact of review type on specific author types would suggest that author characteristics potentially influence evaluations directly (in single-blind) or indirectly (in double-blind). Using relevant outcome data, including the judged quality and popularity of the conference talk and whether the submission was ultimately published, we can test whether either review system better predicts relevant outcomes.

Our multi-faceted comparison provides a more comprehensive assessment of the equity of single- vs. double-blind review systems than prior research. If single-blind review differentially favors a particular author characteristic, but is more accurate in predicting relevant outcomes than double-blind review, then this would imply the characteristic is a valid proxy for merit. Such a result would suggest a legitimate reason for favoring single-blind review. Alternatively, if a particular characteristic is favored in single-blind compared to double-blind review and there is no difference in the accuracy of the evaluations from the two review systems, then this would be evidence of inequities from single-blind review. It is also useful to assess the degree to which any author characteristics may predict the relevant outcomes. If, for instance, none of the author characteristics are predictive, there is little justification for making individuating information available during the review process. Testing whether author characteristics are predictive will also help establish the degree to which any differential treatment of author characteristics may reflect accurate beliefs about the statistical association between author characteristics and the submission’s merit.

The setting of our study also makes our comparison of single- and double-blind re-

view particularly informative. The SJDM organization is an international society of about 1,800 psychologists, economists, organizational researchers, decision analysts, and others dedicated to studying judgment and decision making. Thus, the reviewers' expertise and the domain make it a near-ideal context in which to examine the effect blinding has on expert judgment. At the time of the study, the conference used single-blind review to select submissions for talks, but anecdotally it seemed that many members preferred double-blind review. We confirmed this preference in a survey of society members that we conducted (see Supplementary Material Section ?? for details on the survey). Respondents rated their preference for single- vs. double-blind review on a 1 - 7 scale, where ratings greater than 4 indicated a preference for double-blind review. The preference for double-blind review was strongest when respondents took the perspective of an author ($M = 5.8$, $[5.5, 6.1]$) vs. a conference attendee ($M = 5.4$, $[5.1, 5.7]$ on a 7-point scale; $\beta = -0.23$, $[-0.36, -0.10]$), and only slightly stronger as an author vs. a reviewer ($M = 5.6$, $[5.3, 5.8]$, $\beta = -0.11$, $[-0.24, 0.02]$) (Table ??). Less experienced reviewers also had a stronger preference for double-blind review ($\beta = -0.21$, $[-0.36, -0.05]$). Thus, our study took place in a setting in which there was a *practice-preference gap* between the single-blind institutional practice and institutional members' preference for double-blind review. We suspect a similar gap exists for many other institutions and organizations that use merit-based evaluations. At least in the case of SJDM, part of the reason for the practice-preference gap seemed to be a disagreement as to whether the additional author information might make single-blind review more effective than double-blind review. We designed our study to address this unresolved question by collecting a suite of relevant outcome data and implementing a method to assess the reliability of the different review systems.

The domain of peer review for conference submissions is an informative setting to study blinding more generally. Merit-based reviews are used in a variety of managerially relevant contexts (MacCoun, 2020; Robertson & Kesselheim, 2016). Although our field study took place in the specific context of an academic society's conference, many of the features in this setting are present in other contexts in which peer review is used. These features include a group of peer subject-matter experts who evaluate others' work, such as during performance evaluations or project reviews within large businesses or non-profit organizations. Another common feature is that the peer group comes from a relatively large community with many interconnected relations. Thus, non-anonymized peer review often involves lay beliefs about demographic characteristics inferred from a name and the reviewer's impressions of the specific individuals, based on personal interactions. This property occurs in many different management contexts. For example, when an organization solicits outside proposals for a project, such as developing marketing materials, its marketing personnel are likely to have professional connections with the outside agencies and their employees. The impact of race, ethnicity, and gender on peer evaluations is of broad interest, both in academia and managerial contexts. Furthermore, our study also included factors such as reputation, professional stature, and prestige. These factors and their impact on peer evaluations are also important but have been less studied in these contexts. Thus, we designed the current study to extract generalizable principles about the strengths and weaknesses of employing either single- or double-blind review in contexts in which evaluation is conducted. In the discussion, we review these strengths and weaknesses and make evidence-based recommendations for organizations that want to employ peer review

for merit-based evaluations, particularly but not exclusively peer-review of scientific work.

An important factor in our setting is that reviewing is often done under time pressure and with limited relevant information. For example, in the current study, each reviewer evaluated approximately thirty 600-word submissions within three weeks as an additional task on top of their regular workload. Under these circumstances, we might expect reviewers to rely on heuristic cues like identities and credentials to infer the quality of the work (Hertwig et al., 2019). Another relevant aspect of the setting is the interdisciplinary nature of the conference. International scholars with backgrounds in economics, management, marketing, psychology, and other related areas participate in the conference. Moreover, some conference topics, such as probability theory, utility theory, statistics, and methodology, are mathematically intensive, a characteristic often associated with disparities by gender and race (Ceci et al., 2014). Thus, we expect results from our comparison of single- and double-blind review to generalize to a wide range of related domains, including those historically associated with disparities.

The conference setting also represented a high-stakes environment. With $N = 530$ submissions for only 108 speaking slots (thus a 20% acceptance rate), competition for a talk, which can have important career implications, was tight. Presenting a paper at a high-impact conference can be associated with an increased chance of publishing the work (Gorodnichenko et al., 2021) and, for published work, ultimately lead to more citations (de Leon & McQuillin, 2020). Moreover, many submissions were from authors on the job market, for whom being selected for a talk can be an important signal.

Finally, the setting also allowed us to address a key methodological challenge in comparing single- and double-blind review: the variability between submissions. Instead of trying to match submissions between single- and double-blind review conditions, as might be done in a typical audit study (Bertrand & Duflo, 2017), or creating artificially matched submissions, as might be done in a correspondence test (Jowell & Prescott-Clarke, 1970), we leveraged the fact that multiple reviewers evaluated each submission. We randomly assigned each reviewer to conduct all of their reviews in either the single-blind condition or the double-blind condition. We then randomly assigned each submission to at least three single- and three double-blind reviewers. Altogether this approach permitted us to directly control for heterogeneity across the submissions while employing random assignment to isolate the causal effects of single- and double-blind review.

Methods

The SJDM Executive Board and the XXXX Institutional Review Board approved the study. The preregistration for the study is listed on OSF [here](#). The supplementary materials (SM) document the deviations from the preregistration. De-identified survey data and analyses are available at the Open Science Framework: [Blinded OSF Anonymous Peer Review Link](#). More detailed information on the methods, including the measures used and collected and the procedures are in the SM.

Submissions

There were $N = 530$ submissions for possible presentation at the meeting. Submissions had between one and six authors, with an average of 2.8 ($SD = 1.0$). A total of 26

(5%) submissions had one author, while two-author submissions were the most prevalent, at 218 (41%) submissions. Table ?? summarizes the characteristics of submitting authors. In the past, no information was collected beyond authors' names and institutions, so authors submitting work were not expecting to have to enter additional information. In order to maintain a submission process that was as similar to the process of previous years and not raise suspicions of authors, we asked for detailed demographic information of only the corresponding author. For the corresponding authors, we asked for information on gender, race, citizenship, year of Ph.D., area of Ph.D., and position. For the co-authors, we asked for gender and position, and we also have the institution for each author. Asking a corresponding author to submit all of this information on all co-authors would have been a substantial change from previous years and we were concerned about raising suspicion. The supplementary materials contain a complete listing of all measures and observations that were collected at the time of the submission (Section ??).

Reviews

A total of $N = 113$ reviewers, selected to be representative of the society's faculty membership, were recruited to evaluate the submissions (see Table ??). One reviewer did not complete their reviews, leaving $N = 112$ reviewers. The society sent its standard review invitation, which contained no information about the study. Because a larger than usual set of reviewers was recruited, many reviewers had not previously reviewed for the conference ($N = 50$). Prior reviewers had mostly reviewed between 1 and 2 times previously for the conference ($M = 1.5$; $SD = 1.7$). Reviewers were randomly assigned to either the single- or double-blind condition. Each reviewer was asked to review 30 randomly assigned submissions within three weeks of the assignment. Consistent with precedent at the society, in both conditions, reviewers rated each abstract on a nine-point scale from 1 (Poor) to 9 (Excellent). For all analyses, we standardized ratings within each reviewer. The supplementary materials contain a complete listing of all measures and observations collected from the reviewers (Section ??).

Talk Ratings

We recruited $N = 18$ faculty and $N = 12$ doctoral students and post-doctoral researchers (henceforth students) to rate the $N = 108$ conference talks. We attempted to recruit talk raters who were representative of the society membership (see Table ??). The conference had nine sessions, each with three tracks of four talks per session. Our goal was to have two faculty and two student raters attend each of the talks, giving us full coverage of all the talks. Each faculty member was randomly assigned to rate three sessions to achieve this goal. Specifically, we randomly assigned faculty to one of the three tracks in three of the nine sessions during the conference. Students and post-doctoral researchers were randomly assigned to rate the talks in one of the tracks, in either the first five sessions or the last four sessions of the conference. The supplementary materials contain a complete listing of all measures and observations collected by the raters (Section ??).

Procedure

We designed the study to keep the submission process consistent with the typical process for the annual meeting (e.g., same abstract length, call for submissions, and review criteria and instructions), adding only the step of randomly assigning half the reviewers to double-blind review. To maintain the experiment's integrity, only the authors of this paper, the SJDM executive board, the NSF program officers, and the NSF grant reviewers were aware of the field experiment. The field experiment was revealed to attendees on the last day of the conference during the morning business meeting.

The call for submissions was issued on Monday, June 4, 2018, with a deadline of June 22, 2018. Reviewers were recruited soon after the submission deadline. We assigned submissions to reviewers on July 6, with a deadline of July 26, 2018 to submit their reviews. We randomly assigned submissions to reviewers so that at least three double-blind and three single-blind reviewers reviewed each submission. To assess reliability in the different review processes, we randomly assigned $N = 53$ of the submissions (10%) to three additional double-blind and three additional single-blind reviewers. This design resulted in reviewers rating an average of $M = 31$ ($SD = 1.6$) submissions.

The conference chair (XXXX) selected 108 talks for presentation at the annual conference based primarily on the average rating (across the single- and double-blind ratings). For the subset of submissions with two sets of single- and double-blind reviews, the chair randomly selected one set of three single-blind and one set of three double-blind reviews to include in the average. Consistent with past conferences, the talks occurred during nine sessions, each consisting of three parallel tracks, composed of four talks each. We assigned two students and two faculty raters to each talk. They independently rated each talk on its significance, methods, results, innovation, and uniqueness, which we, in turn, used to form a composite measure of overall judged quality. We also had two research assistants assigned to each talk. They counted attendance (at five minutes into the talk), the number of hands raised to ask a question throughout the talk, and the number of questions answered.

Finally, two years after the conference, we sought to identify published submissions. We did this via a Google Scholar search and by emailing contact authors for submissions. For each submission, we identified whether a submission had been published in a peer-reviewed journal, and if it had, the impact factor of the journal at that time.

For more details on the procedures and all the measures collected, please see the supplementary materials (Section 2).

Analysis

In general, we take a quantitative approach focusing on the effect sizes for all the collected variables. Unless otherwise noted, the predictors and outcome variables used in all the regressions were standardized to facilitate interpretation and comparison between regression coefficients. When appropriate, we also report the effects in non-standardized units to help the reader appreciate the practical significance of the effect. We used Bayesian estimation methods for data analysis. We report the posterior predicted mean of the parameter of interest and, in brackets, the 95% Credible Interval (CI). The CI summarizes the posterior distribution, excluding the 2.5% of the distribution at each tail. The values in the CI are the probable values of the effect conditional on the observed data and thus

convey the uncertainty around the effects. To highlight particular results, we use the term *credible* when the 95% CI excludes 0.

We also went further than conventional Bayesian data analysis approaches and evaluated the improvement that an independent variable brings to predicting an outcome variable. We evaluated the predictive improvement via a model comparison approach using leave-one-out cross-validation (LOO). The model comparison compared the difference in the expected log pointwise predictive density for a new dataset (elpd_{loo}) between including a variable (e.g., author seniority) or not. This statistic quantifies the difference between the models in predicting a held-out observation. Generally, we consider $|\Delta \text{elpd}_{\text{loo}}| > 1 \text{ SE}$ to be evidence for a particular model. Thus, for instance, if a model with the interaction between a measured variable and the reviewer condition is favored over a model that does not have the interaction term, then the LOO implies that the interaction term helps predict the difference between single- and double-blind review.

In all cases, we used the weakly informative default priors. That is, the priors carried a little information into the analysis, only moderately biasing the parameter of interest toward a particular value, typically 0. This property (i.e., priors skeptical of extreme parameter values) provides moderate regularization of the estimates, which reduces the risk of over-fitting. We inspected MCMC chains for representation and accuracy and sought to have all reported parameters based on an effective sample size of 10,000. Because posterior distributions are independent of the number of planned or unplanned tests, there is no need to correct for the number of posterior distribution tests (Gelman et al., 2012; Kruschke & Liddell, 2018a).

Further details on the analysis are available in the Supplementary Materials (Section ??). All scripts and output from our final analyses are posted on the Open Science Framework.

Power

As we stated, our goal in the field study was a quantitative approach focusing on the effect sizes for all the collected variables. This quantitative approach combined with our Bayesian methods implies we are not relying on null hypothesis significance testing but focusing on the posterior distribution of effects of interest. To highlight particular results, we used the term *credible* when the 95% CI excludes 0. Thus, a typical null hypothesis significance testing power analysis where the probability of obtaining a p-value less than .05 is not appropriate. Nevertheless, one can estimate the probability of achieving a particular goal in a Bayesian generalized power-analysis (Kruschke & Liddell, 2018b). We did this focusing on our lowest-powered statistic: the Pearson correlation r . Our analysis asked what the probability of a 95% credible interval for a posterior distribution of over the Pearson r excluding 0 would be for a given sample size and a given generating correlation ρ . We took this approach because a power analysis with multi-level regression is not straightforward and entails many assumptions. All of our primary analyses and conclusions are largely based on Bayesian multi-level regressions, which are more powerful. Thus, the power analysis with the Pearson correlations represents a lower bound.

The Supplementary Material provide a full description of the power analyses we conducted for the Pearson correlations and the difference in the Pearson correlations (see Supplementary Material Section 2.1). The latter is relevant as we were interested in the dif-

ferences between single- and double-blind review. Our power analyses revealed that for the analyses focused on the larger set of submissions (e.g., the relationship between a particular author characteristic and the reviewer ratings), samples of data generated with correlations of $\rho \approx 0.15$ had approximately an 80% probability or larger of being labeled as credible. For analyses focused on the submissions given as a talk where the sample size was $n = 108$, the corresponding correlation was $\rho \approx 0.27$. Thus, our field study was adequately powered to detect small to moderate associations with our least powerful statistic.

Many of our questions and hypotheses focused on the difference between single- and double-blind review. With the Pearson correlation, this comparison is at the level of the difference between correlations ($\Delta\rho$). Our power analyses revealed that for the analyses focused on the larger set of submissions, samples of data generated with differences in correlations of $\Delta\rho \approx 0.20$ had approximately an 80% probability or larger of being labeled as credible. For analyses focused on the submissions given as a talk where the sample size was $n = 108$ the corresponding correlation was $\Delta\rho \approx 0.35$. Thus, our field study was adequately powered to detect what we might consider small to moderate differences between correlations with our least powerful statistic (see Supplementary Material Section 2.1).

Most of our analyses and conclusions are based on more powerful regressions that increase the power of detecting if a variable was credibly associated with a dependent variable of interest. There were several aspects by which our regressions increased the power. First, we used simultaneous regressions controlling for 12 author and submission characteristics and 5 reviewer characteristics. Second, each submission received at least 3 single-blind and 3 double-blind reviews. This property created a within-submission comparison of single- and double-blind review. In our statistical models, we accounted for this via multi-level regressions with submission as a random effect (a random intercept). Because each reviewer reviewed approximately 30 submissions, we also could treat the reviewer as a random intercept. Similarly, in the outcome data, the models accounted for random effects of the session and other variables. Third, we used vague but informed priors to regularize our regressions and results to help guard against overly sensitive analyses of extreme data. Interaction effects in multiple regressions are essentially differences in correlations, so these statements are true for the interaction effects. For more details on the power analyses please see the Supplementary Materials (Section 2).

Results

The first author was also the corresponding author for $N = 456$ of the 530 submissions. Because we only have data about the corresponding author for many of the variables (including nationality, race, and area of Ph.D.), our primary analyses focus on submissions where the first author was the corresponding author. Table ?? lists summary statistics of the authors. Next, we compare single- and double-blind review in terms of the review rating reliability, the differential impact of author and submission characteristics, and the predictive validity for relevant outcomes, including the quality and popularity of the presentation and subsequent publication.

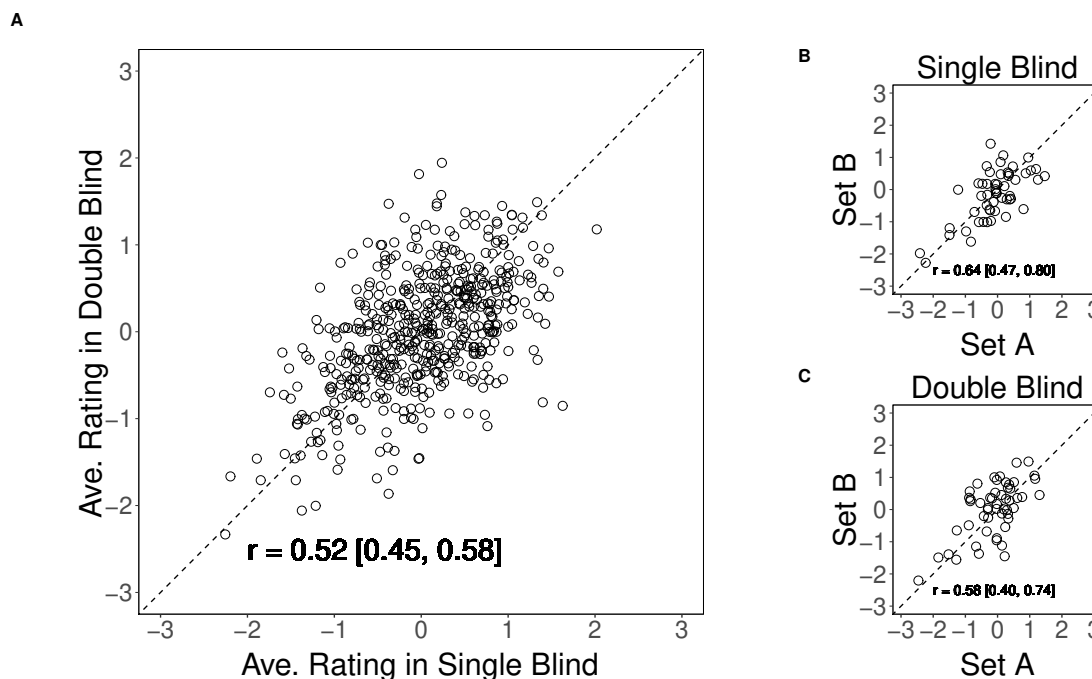


Figure 1. A: Correlation between ratings from the two review processes. B: Relationship between the reviews for the subset of 53 submissions reviewed by two teams of reviewers in the single-blind condition (inter-rater reliability). C: Relationship between the reviews for the subset of 53 submissions (same set as in B) that were reviewed by two teams of reviewers in the double-blind condition (inter-rater reliability). There is no credible difference between the two conditions.

Reliability

Reliability of the Average Ratings. We first analyzed reliability at the submission level, quantifying agreement both within and between single- and double-blind review ratings. Recall that per procedures of the society, reviewers were randomly assigned to a subset of submissions, reviewing on average 31 ($SD = 1.6$) out of the $N = 530$ submissions. This design meant that each submission received at least three single-blind and three double-blind reviews. Per procedures in the society, submission decisions were based on the average rating across reviewers assigned to each submission (after standardizing the ratings for each rater). Thus, we first look at the review rating reliability based on the average overall rating for each submission.

Between Condition. The correlation between the average double-blind and average single-blind review across all 530 submissions was $r = 0.52$ [0.45, 0.58], conventionally a moderate level of agreement (Figure 1). However, the level of agreement depended on the rated average quality of the submission. A median split by the single-blind rating reveals more agreement between single- and double-blind review for the bottom half of the submissions ($r = 0.45$, [0.35, 0.54]) than the top half ($r = 0.19$, [0.07, 0.29]; $\beta = -0.40$, [-0.65, -0.16]; see Table ??). The low level of agreement at the top is consequential.

Among the top 108 submissions (i.e., the number accepted for talks), only 43 submissions (40%, [31, 49]) would have been recommended for talks under both single- and double-blind review. These results suggest that the conference program would feature a different talk line-up if it switched from single-blind to double-blind review. In comparison, there was more agreement among the bottom 108 submissions (58 submissions in common, 54%, [44, 63]; $M_{\text{diff}} = 14\%$ [1, 27]), and therefore more agreement about what constitutes a weak submission than a strong one.

Within Condition. Recall that 10% of the submissions ($N = 53$) were randomly selected for review by six single-blind reviewers and six double-blind reviewers. We designed the study in this way to evaluate the aggregate-level reliability within each review-type condition. The aggregate-level reliability is of interest because the standing procedure for the conference—which was also used for this conference—has been to use the average rating across three reviewers to recommend submissions for talks. For this subset of submissions, for both single- and double-blind review, we divided the six reviews into two sets of three reviews (Set A and Set B). For each set of three reviews, we calculated the average rating. Figures 1B and C plot the relationship between the average ratings given to these submissions between the sets of reviewers. Regressing the average reviews for each submission onto the review condition (single- or double-blind), set (A or B), and the interaction between these two revealed a small but non-credible interaction ($\beta = 0.20$, $[-0.06, 0.45]$). This result implies the two review processes do not credibly differ in reliability.

Crucially, just as the moderate level of agreement between review systems means that fewer than half of the submissions recommended for a talk with one review system would be recommended with a different review system, the moderate level of agreement within conditions means the same would be true if the same review system was used but with a different set of reviewers. To see this, we simulated 530 pairs of normally distributed random variables correlated at $r = .64$ (the level of agreement we observed in the single-blind condition) and at $r = .58$ (the level of agreement we observed in the double-blind condition). Then, we calculated the proportion of times the same pair was in the top 108 values for each variable. We repeated this simulation 10,000 times. On average, we found agreement for 49% of the pairs (i.e., both values were in the top 108 values) in our simulation of single-blind review, and 46% of the pairs when simulating double-blind review. Given the moderate level of agreement within review systems, the conference program would thus feature quite different talks if the same review system was used, but the conference switched from one randomly-selected subset of reviewers to another. We discuss the implications of this moderate level of agreement in the Discussion.

Reliability of the Individual Reviewer Ratings. The reliability analyses so far have focused on averaging the ratings for each submission. Because each reviewer was randomly assigned to a subset of the $N = 530$ submissions, we have limited information to estimate the within-system reliability at the individual level meaningfully. Specifically, the number of submissions common to any two reviewers ranged from 0 (the median value) to 29. Nevertheless, to better understand the reliability at the individual reviewer level, we did the following. First, within each review condition (single- and double-blind) for each submission, we randomly selected two reviews, creating $n = 530$ pairs of reviews for a given review condition. Then, we calculated the correlation between these pairs of reviews. Finally, we repeated this entire procedure 100 times, recording the posterior distribution

of the correlation for each iteration. We ran this procedure for each review condition. For the single-blind review condition, the average correlation between individual review scores was $M = 0.34$, and the average 95% CI was $[0.26, 0.41]$. For the double-blind review condition, the average correlation between reviews was $M = 0.28$, and the average 95% CI was $[0.20, 0.35]$. For each iteration, we also calculated the difference in the posterior distributions of the correlations between the single- and double-blind condition. On average, the single-blind review condition had a slightly higher correlation between reviews, but it was not a credible difference, with the average difference being $M = 0.06$ and the average 95% CI for that difference being $[-0.05, 0.17]$.

Summary. The reviewer-average ratings typically used to evaluate submissions to the conference had moderate levels of reliability in both review-type conditions. At the individual level, review ratings had low reliability in both review types. We did not find credible differences in reliability between single- and double-blind conditions. However, there was more agreement between review conditions regarding lower-rated than higher-rated submissions, suggesting more agreement when evaluating weak submissions than when evaluating strong submissions.

Differential Impact of Author Characteristics

Our next set of analyses examined the degree to which author characteristics differentially impacted review ratings. This analysis addressed the question that most people may be primarily concerned with when considering the equity of the different review processes: the outcomes of historically majority or over-represented authors vs. authors from other backgrounds. Do women score better with a single- or double-blind review process? Do people who identify as White score better under single- or double-blind review processes than other ethnic groups? How is author seniority related to the reviews one receives? To answer these questions, we collected data for up to twelve different characteristics for each submission (Table 1). One set of characteristics consisted of information related to author identities: the authors' gender, race, position, institution, first author nationality, years since Ph.D., and Ph.D. area.

The second set of characteristics focused on the submitted abstract's characteristics. These variables consisted of information extracted from the submission text via natural language processing tools (see the Supplementary Material Section ?? for details). Briefly, one characteristic was the sentiment (-1 (negative) to 1 (positive)) estimated using sentiment analysis. The other characteristic was the topic of the text, estimated via a topic model from the extensive *Wiley-Blackwell Handbook of Judgment and Decision Making* (Keren & Wu, 2015). The topic model identified three topics: (1) theories, (2) empirical phenomena, and (3) psychology of judgment and decision making. We assigned each abstract the topic with the highest posterior probability. We included these characteristics because double-blind reviewers' lack of information on authors' characteristics could result in relying more on features of the submission itself. If so, the submission topic might be more strongly associated with the review score in the double-blind than in the single-blind condition. Moreover, including these variables controlled for differences in topic and sentiment, that might otherwise be confounded with author characteristics.²

²Our preregistration plan was to include submitted keywords representing submission characteristics.

Table 1

Predictors used in comparing single- vs. double-blind reviews.

Characteristic	Variable	Description
Condition	Single- or double-blind review	A binary variable indicating single- (0) or double-blind (1) review.
First Author	Gender of first author	Binary variable indicating whether the first author was female (0) or male (1).
	Years since Ph.D.	Years from first author's Ph.D. to 2018.
	Domestic or international	Whether the first author was a U.S. citizen (0) or not (1).
	Race and Ethnicity	Race and ethnicity of the first author (Asian, Black, White, Hispanic, Native American, Other, Not available).
	Ave. number of citations for JDM-relevant departments at first author's institutions	Average number of citations for top papers from Psychology, Economics and Business, and Social Science departments at the institution of the focal author as determined by the Web of Science (August 29, 2019). We use this variable as a proxy for institutional prestige.
	Ph.D. area of the first author	Ph.D. area of the first author (Economics, Management, Marketing, Other, Psychology).
	Position of the first author	Position of the first author (student, undergraduate student, master's student, Ph.D. student, Post-doc, research scientist, practitioner, assistant professor, associate professor, full professor).
Coauthors	Proportion of coauthors who were men	Proportion of coauthors excluding first author who were male.
	Ave. number of citations for JDM-relevant departments at coauthors' institutions	Average number of citations for top papers from Psychology, Economics and Business, and Social Science departments at the institution of the coauthors. See ave. number of citations JDM-relevant departments at first author's institutions.
	Ave. position rank of coauthor(s)	Average position rank of coauthor(s) excluding focal author where there are 10 positions (see Position of first author).
Text	Topic	Topic was calculated from a topic model of the 35 chapter Blackwell-Wiley Handbook of Judgment and Decision Making Keren & Wu (2015). Three core topics emerged from the handbook: theories of judgment and decision making, empirical phenomena of judgment and decision making, and psychology of judgment and decision making.
	Sentiment	Sentiment score was estimated from a sentiment classifier where sentiment ranged from -1 (negative) to 1 (positive).
Reviewer	Reviewer gender	Binary variable indicating whether reviewer was female (0) or male (1).
	Reviewer years since Ph.D.	Years since 2018 that reviewer obtained Ph.D.
	Ave. number of citations JDM-relevant departments at reviewer's institutions	The average number of citations for top papers from Psychology, Economics and Business, and Social Science departments at the institution of the coauthors. See ave. number of citations JDM-relevant departments at first author's institutions.
	Ph.D. area of the reviewer	Ph.D. area of the reviewer (Economics, Management, Marketing, Other, Psychology).
	Position of the reviewer	Position of the reviewer (research scientist, practitioner, assistant professor, associate professor, full professor).
	Reviewer experience past 7 years	Number of times reviewed for the SJDM Annual Meeting in the past 7 years.

In the regressions, this statistic was sometimes expanded to include all authors to not exclude single-author submissions. See Footnote 3.

Correlations Between Author Characteristics and Review Ratings. Figure 2 lists the correlations between the characteristics and the review ratings. Inspecting the bottom two rows of Figure 2 reveals similar correlational patterns between the review ratings and the individual characteristics for the two review systems. The only credible difference was that the correlation between the average coauthor rank (i.e., seniority) and review ratings was credibly higher in the single-blind condition than in the double-blind condition (diff = 0.19 [0.07, 0.33]). This pattern suggests that single-blind reviewers based their evaluations partly on coauthor seniority, which was not available to double-blind reviewers.

Simultaneous Regression Analyses. To further assess the degree to which the characteristics differentially impacted single- and double-blind review ratings, we entered review type (single vs. double), all twelve author and submission characteristics, and the

However, the variability in the keywords proved too extensive, and we had more confidence in the topic modeling and estimated sentiment. See Supplementary Material for details on the Preregistration and deviations from the Preregistered plan (1)

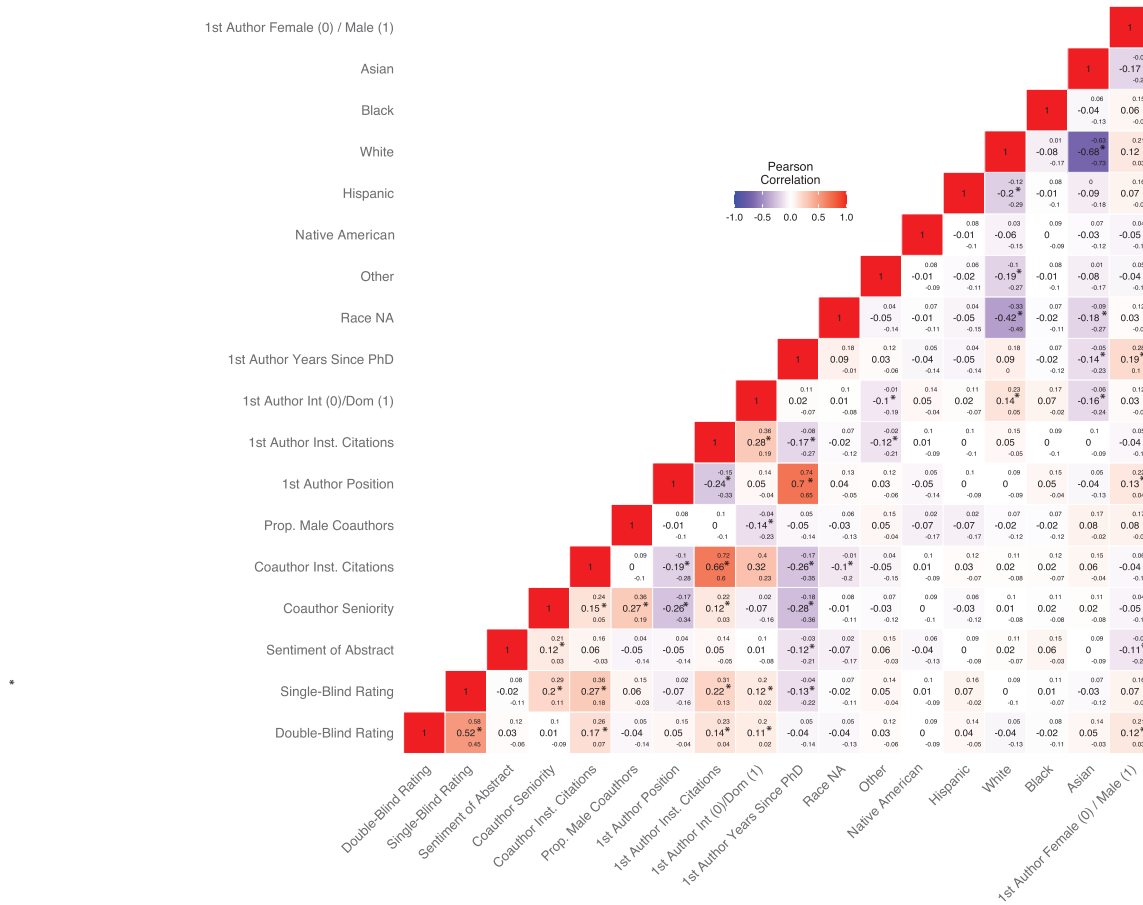


Figure 2. Correlations between author and text characteristics and single- vs. double-blind review ratings. The central value is the mean posterior estimate. The values in the upper and lower right-hand corners are the estimates of the upper and lower bounds of the 95% credible interval. Starred values indicate a credible interval that excludes 0. Academic position of the first author was treated as a numerical variable in these correlations but as a categorical variable in the regressions. Correlations for race and ethnicity (Asian, Black, White, Hispanic, Native American, Other, and No Answer) were calculated by comparing each group to all others. In the regressions, race and ethnicity coefficients represent comparisons to White. Area of Ph.D. and submission topic (from a topic model) are categorical variables and are not in these matrices but were entered into the regressions. Table 1 describes each of these characteristics. See also Tables ??-??.

interaction between each characteristic and review type simultaneously in a Bayesian hierarchical regression model, with the rating of each submission by each reviewer as the outcome variable. The regressions modeled the variability between submission and reviewers as random intercepts. We also controlled for differences between reviewers by including six potentially relevant reviewer characteristics (e.g., gender, position, and experience) in

the regressions as covariates (Table 1).³ In total, $N = 394$ submissions had a complete set of characteristics for the analysis. By definition, the $N = 26$ single-author submissions did not have characteristics for coauthors. Therefore to include them in our analyses, we replaced the coauthor variables with variables for all authors (e.g., the proportion of male authors instead of male coauthors) for all the submissions.

Evidence for Overall Differential Impact Across Author Characteristics.

We first asked if there was evidence for overall differential impact across all twelve characteristics. To do this we compared a model with the interaction terms between each of the twelve characteristics (Differential Impact Model; **DI Model**) and the review condition to a model without these interaction terms (No Differential Impact Model; **No DI Model**). The model comparison was done by comparing the LOO statistic for the **DI Model** and the **No DI Model**. This is a strong test, as it asks whether it is necessary to jointly account for differential impact in all twelve characteristics to predict the differences between single- and double-blind review. The difference between the **No DI Model** and **DI Model** was ($\Delta elpd_{loo} = 6.659$; $SE = 6.922$; ??) favoring the **No DI Model** but well within the margin of error. Thus, we find little support for overall differential impact of single- vs double-blind review across the full set of author and submission characteristics. However, even if there is insufficient evidence of difference across all the characteristics, differences between single- and double-blind review may occur for some individual characteristics, which we test next.

Differential Impact of Specific Author Characteristics. We next tested for the differential impact of review type on specific individual or submission characteristics, controlling for the other characteristics. The key statistics of interest were the interactions between the review conditions and each characteristic. We identified three author characteristics that were differentially related to the reviews in the two review systems: first-author gender, first-author race, and co-author seniority.

In terms of first-author gender, we found a credible interaction between first-author gender and the review condition ($\beta = 0.255$, $[0.066, 0.445]$; Table ??). To better appreciate the strength of this interaction, we compared the No DI Model, where only submission characteristics were entered as predictors, and a second model that added (stepwise) the first-author gender by review condition interaction term to the No DI Model. This model comparison did not provide strong evidence in either direction, as the comparison was within the margin of error ($\Delta elpd_{loo} = -0.110$; $SE = 1.738$; Table ??). Taken together, the credible interaction, on the one hand, and the inconclusive support from the stepwise model comparison of the predictive improvement, on the other hand, warrant caution in interpreting the interaction. With this cautionary note, to better understand this weak interaction, we regressed first-author gender onto review ratings separately for single- and double-blind review (Tables ??). These regressions revealed that submissions with male first-authors received a 16.3% $[3.5, 28.6]$ higher rating than those with female first-authors in the double-blind condition, as compared to a weaker, non-credible 9.4% $[-4.6, 23.0]$ higher rating for submissions with male first-authors in the single-blind condition (Figure 2, Tables ?? & ??). Notably, the regression analyses suggest these differences in review ratings are unlikely to

³We tested whether the effects of review type differed by reviewer characteristics. None of the reviewer characteristics showed credible interactions with the review condition (Table ??). Moreover, including these interactions as well as potential three-way interactions with author-specific characteristics, led to a substantially worse prediction by the regression model (via leave-one-out cross-validation) (Table ??).

be due to differences in submission sentiment, research topic, or any other author characteristic that might be correlated with author gender, as these are controlled for in the analysis. Moreover, our regression analyses reveal that this interaction between first-author gender and review system is present when controlling for reviewer characteristics (e.g., gender of the reviewer), suggesting reviewer characteristics cannot explain these results. We did not anticipate this slight boost in ratings for male-authored submissions in double-blind reviews and return to this topic in the Discussion.

To test differences based on the first-authors race/ethnicity, this variable was coded into seven categories: Asian, Black, White, Hispanic, Native American, Other, and Race Not Available. In the regressions, race was entered as a categorical variable with each category compared to White. The analysis revealed a credible interaction between the review condition and whether the first author identified as Asian (compared to White) ($\beta = 0.180$, $[0.010, 0.351]$; Tables ??).⁴ That is, controlling for other author and submission characteristics, including whether the author was a U.S. citizen or not, the interaction coefficient implies there was an 18% gap in the ratings of submissions with Asian first-authors relative to those with White first-authors in double- vs. single-blind review. Regressing first-author race onto single- and double-blind reviews separately revealed that this credible interaction was driven by Asian first authors receiving an 8.6% $[-7.3, 24.3]$ *higher* rating than White first-authors in the double-blind condition. In contrast, they received a 3.5% $[-12.8, 20.2]$ *lower* rating than White first-authors in the single-blind condition.

Lastly, we also found a difference between single- and double-blind review based on coauthor seniority. Recall that the correlations indicated a credible difference between single- and double-blind review based on coauthor seniority (Figure 2A). However, this result did not persist in the regressions in which single-author submissions were included. It appears that replacing the coauthor variables with variables for all authors—in this case, the average seniority of all authors instead of the average seniority of the co-authors—and thus losing the distinction between first-author and coauthors mitigated the effect (Table ??). Therefore, we conducted another set of regressions, using only the multiple-author submissions ($N = 370$).

This multi-author regression revealed a credible interaction between the review condition and the average seniority of coauthors ($\beta = -0.090$, $[-0.168, -0.011]$) (Table ??). Model comparisons via the stepwise addition of this interaction term to the No DI Model (where only submission characteristics were entered as predictors), revealed that including this term improved the model’s predictive accuracy ($\Delta elpd_{loo} = 7.820$; $SE = 4.270$; Table ??). To better understand this interaction, we regressed coauthor seniority on review ratings separately for single- and double-blind review. These regressions revealed that in single-blind review, for every position increase in average coauthor rank (e.g., from associate to full professor), there was an 8.2% $[4.5, 11.8]$ increase in the review rating. However, there was no credible comparable change for double-blind review ($M = 0.3\%$, $[-3.2, 4.0]$) (based on a 1.91 standard deviation of coauthor rank; see Table ??; see also Figure 2A & Table ??). These differences in review ratings are unlikely to be due to differences in

⁴Model comparisons via the stepwise addition of the race interaction term onto the No DI Model revealed that including the first author’s race did not improve the model’s predictive accuracy ($\Delta elpd_{loo} = 5.223$; $SE = 1.826$; Table ??). However, this model comparison was based on including the race of all individuals.

Table 2

Coefficients from regressing submission outcomes onto single- and double-blind ratings.

	Single (Alone)	Double (Alone)	Rating x Condition
Talk Rating	0.05 [-0.08, 0.18]	-0.02 [-0.14, 0.10]	-0.07 [-0.31, 0.17]
Attendance	-0.11 [-0.42, 0.21]	-0.11 [-0.39, 0.17]	-0.04 [-0.61, 0.52]
Num. Questions	-0.01 [-0.36, 0.34]	0.16 [-0.15, 0.47]	0.11 [-0.53, 0.75]
Talk/Poster			
Rating	0.35 [0.14, 0.56]	0.23 [0.01, 0.43]	-0.18 [-0.75, 0.41]
Published	0.29 [0.08, 0.52]	0.42 [0.19, 0.66]	0.28 [-1.60, 2.21]
Impact Factor	0.24 [0.04, 0.43]	0.19 [-0.01, 0.40]	-0.03 [-0.36, 0.29]

Bolded values indicated that credible intervals exclude 0. Num. Questions are the number of hands raised for questions during the talk. All continuous variables were standardized, and the impact factor was log-transformed. When analyzing conference outcomes, the conference session was entered as a covariate. When analyzing publication outcomes, conference presentation (1 or 0) was entered as a covariate. The coefficient for the Rating $\times$ Condition Interaction is from a model where the outcome was regressed onto the review rating, the condition, and the interaction. The regressions predicting “published” as an outcome are logistic regressions, while all other regressions use a normal link function.

other author characteristics, submission sentiment, research topic, or reviewer characteristics, which are controlled for in the analysis. We return to this result in the Discussion as well. Still, the higher ratings for submissions with senior coauthors, only in single-blind review, suggest reviewers were using this characteristic as a cue.

Summary. Comparing single- and double-blind review revealed that although, overall, across all twelve author and submission characteristics, there was no evidence for these characteristics being treated differently between single- and double-blind review, there was evidence suggesting that a subset of the characteristics were treated differently. Specifically, single-blind review favored senior coauthors more than double-blind review, while double-blind review favored male authors more than single-blind review. We also found that Asian first authors were more likely than White first authors to be rated lower in single- vs. double-blind review. These results suggest that reviewers are influenced by some of the author identity cues that are only available in single-blind review. However, the analyses thus far do not determine whether reliance on these cues makes evaluations less or more accurate in identifying high-quality submissions. Next, we assess to what degree and how both single- and double-blind review ratings, as well as the characteristics available in single-blind review, are related to relevant outcomes: judged talk quality, talk popularity, and ultimately publication.

Validity

Predictive Validity of Review Ratings. Our final set of analyses focused on predicting a set of outcome variables for each submission which measured the judged quality of the conference talk and future publication potential of the work. Specifically, we collected talk attendance, number of questions asked, expert quality judgments of the full conference talk, and publication status of the submissions two years later. Our first analyses tested whether the submission reviews themselves were predictive of talk outcomes. After excluding one cancelled talk, our analyses were based on 107 accepted submissions given as talks at the conference. Among the accepted talks, submission review ratings (single-

or double-blinded) did not credibly discriminate between higher and lower quality talks, in terms of differing on expert judgments of the presented talk, attendance, or the number of potential questions (see Table 2 and the bottom two rows of Figure 2B).⁵

However, because the selection of talks for the conference was based primarily on the review ratings, our ability to assess predictive validity is limited by a range restriction problem. The correlations in Figure 2 statistically account for range restriction, suggesting that the range restriction does not account for the lack of predictive validity (see Mendoza & Mumford, 1987, for methods to correct for restriction of range). As an additional test, we extended our analyses to include additional submissions ($N = 56$) that were not accepted as talks but were presented as posters and were judged by experts, on a similar scale, in a student-poster competition. We created an overall score, similar to our talk ratings, to represent the judged quality of the poster presentation.⁶ Combining the expert judgments of posters with expert judgments of the spoken presentations revealed that the single- and double-blind ratings both credibly predicted these judgments. Moreover, the validity did not differ by review type (Table 2).

We also assessed the submissions' publication status two years after the conference. We identified $N = 184$ of the 530 submissions (35%) as published in a peer-reviewed journal. Controlling for presentation at the conference, the review ratings from both single- and double-blind reviews credibly predicted the submissions' publication status. However, the ability of the reviews to predict future publication did not differ by review type (Table 2). Among the published papers, we also looked at whether review scores predicted Web of Science's (logged) journal impact factors. The single-blind review ratings credibly predicted the impact factor, while double-blind ratings had a smaller, non-credible relationship with impact factors. However, there was no credible difference between single- and double-blind scores in predicting the impact factor for published work. In sum, we did not find a credible difference between the validity of single- and double-blind review scores for any of the outcomes.

Predictive Validity of Author Characteristics. A second analysis tested whether any author or submission characteristics revealed to reviewers in single-blind review predicted the outcomes. A common argument for single-blind review is that author identities or credentials may provide valuable predictive information. Perhaps, for instance, submissions with more senior coauthors involve better research and result in better talks; therefore, factoring this characteristic into reviews improves conference quality. No single author characteristic showed a strong, consistent predictive relationship across the measured outcomes (see top 16 rows of Figure 3 and Table ??).

However, some author characteristics did have a relationship with specific outcomes worth highlighting. The first relationship was an association between judged talk quality and whether the first author identified as Asian vs. White. Submissions with Asian first authors scored 35.8% [3.5, 68.0] higher on judged talk quality than submissions with White

⁵Talk quality, as judged by experts on-site, was credibly associated with estimated attendance ($\beta = 0.188$, [0.015, 0.363]), after controlling for the conference session: for every one standard deviation increase in the quality rating, there were 6.5 [.3, 12.6] more attendees (on average, there were 93.9 ($SD = 34.9$) attendees at each talk). Such a relationship suggests some signal in the use of on-site ratings of the quality of the talks as an outcome measure.

⁶Our assumption here is that the scales were used in the same way when judging the talks and posters. Conditional on this assumption, the analysis helps address the restriction of range limitation.

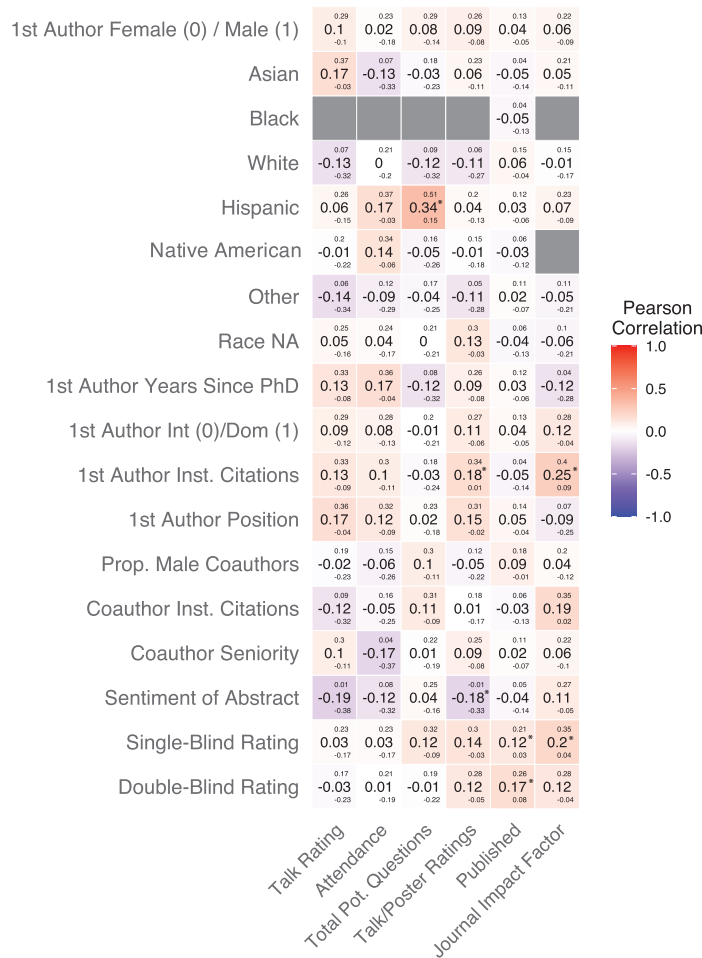


Figure 3. Correlation between author and text characteristics and relevant outcomes. The central value is the mean posterior estimate. The values in the upper and lower right-hand corners are the estimates of the upper and lower bounds of the 95% credible interval. Starred values indicate the credible interval excludes 0. The correlations with overall talk rating, attendance, and total potential questions were corrected for indirect restriction of range (Mendoza & Mumford, 1987). Journal impact factor was log-transformed before computing the correlation. Academic position was treated as a numerical variable in these correlations but as a categorical variable in the regressions. Correlations for race and ethnicity (Asian, Black, White, Hispanic, Native American, Other, and No Answer) were calculated by comparing to the particular group to all others. In the regressions, race and ethnicity coefficients represent comparisons to White. Area of Ph.D. and topic from a topic model analyses are categorical variables and are not in these matrices but were entered into the regressions. Empty gray cells indicate that no observations were available to calculate the correlation. Table 1 describes each of these characteristics. Figure ?? lists the correlations between the outcome variables. See also Tables ??-??.

first authors. Likewise, the judged talk/poster ratings revealed that submissions with Asian first authors scored 69.9% [14.4,126.4] higher than submissions with White first authors.

The higher judged talk and talk/poster quality ratings for submissions with Asian first authors are notable given that single-blind review tended to disfavor submissions with Asian first authors relative to White first authors. However, a higher judged quality rating at the conference did not translate into a higher likelihood of being published or of being published in a higher-impact journal.

The second author characteristic was the prestige of the first author’s institution, as indexed by the average number of citations per paper for JDM-relevant departments at the first author’s institute. For an increase of one in the average number of citations for JDM-relevant departments, there was, on average, a 7.9% [1.4,14.4] increase in the talk rating (based on a 3.56 standard deviation in the average number of citations to papers in JDM-relevant departments). The prestige of the first author’s institution was also associated with attendance, such that for every additional citation in the average number of citations, there were, on average, 26.1 [10.1, 41.8] more attendees. However, this apparent predictive validity of the prestige of the first author’s institution did not translate into an increased likelihood of being published or published in higher-impact journals.

The third characteristic was author seniority. For every increase in the average coauthor rank (e.g., associate to full professor), there were 5.7 [1.1, 10.3] *fewer* attendees (based on a 1.91 standard deviation in coauthor rank; see Table ??). This decrease in attendance for talks with more senior coauthors is notable given that single-blind review favored submissions with more senior coauthors.

The fourth characteristic was the gender of the authors. Presentations with male first-authors tended to have somewhat lower attendance. Though not credible, there were, on average, 21.1 fewer attendees [−0.3, 42.5] for talks with a male vs. female first-author (Table ??). At the same time, the proportion of male authors predicted publication status such that for every 33% increase in the proportion of male authors, the submission was 1.69 [1.17, 2.45] times more likely to be published. These last two effects are notable in light of our finding that male first authors received a higher rating in double-blind than in single-blind review. These results suggest a small advantage associated with submissions from male first-authors (thus higher ratings in double-blind and ultimately a greater likelihood of publication) that was not reflected in reviewers’ and conference participants’ evaluations.

Discussion

Blind review is often discussed as a mechanism to level the playing field, creating more equitable merit-based evaluations across a range of domains, including employment, legal cases, and science (Robertson & Kesselheim, 2016). Past studies examining the effectiveness of blind review in achieving this goal have yielded mixed results, including in the few randomized experiments on the topic. Moreover, none of the past studies tested the impact of blinding on the validity and reliability of merit-based evaluations. Our high-stakes field experiment addressed these shortcomings by comparing single- and double-blind review head-to-head in a real-world context. Doing so provided a unique opportunity to evaluate submissions under both review systems, randomly assigning reviewers to the different systems and submissions. This structure allowed us to control for the heterogeneity between submissions that have limited conclusions in past comparisons and to identify the similarities and differences between the two review processes in situ.

Similarities and Differences Between Single- and Double-blind Review

Overall, our field study revealed that single- and double-blind review were similar in many respects. Ratings from both review systems were similarly valid, predicting some quality-related outcomes. We found both review systems had similar levels of reliability. At the level of the individual reviewers, the reliability was low, in line with some past studies (Bornmann et al., 2010; Fiske & Fogg, 1992; Lawrence, 2015) but not zero or close to zero as others have suggested (Pier et al., 2018). However, recognizing that averaging judgments improves reliability, the society uses the average of three ratings for submissions. Nevertheless, the average ratings within both review-type conditions showed similar levels of moderate reliability.

Even with the boost in reliability from averaging ratings, the moderate level of reliability has important consequences. Single- and double-blind reviews were moderately correlated ($r = 0.52$), such that the two review systems agreed on 40% of the recommendations for acceptance as a presentation. The agreement within review systems was slightly higher ($r = .58$ to $r = .64$). This means that if we reran the review process with the same review system (e.g., single-blind review) but different reviewers, between 46% to 49% of the same submissions would be recommended both times for presentation. Noise impacts many aspects of human judgment (Kahneman et al., 2021), sometimes in ways that make the direct consequences difficult to appreciate. In the context of merit-based evaluations, the consequences are apparent. Due entirely to noise in human judgment, either switching from single- to double-blind review or redoing the review process using the same system would result in over half the presentations at the conference being different. It is worth considering how differences in conference acceptance, arising from noise in human judgments, have been misattributed to systematic differences in review processes. We believe this result illustrates the critical need to establish and evaluate the reliability of any merit-based evaluation, and to develop methods to improve reliability further (e.g., Arkes et al., 2006).

The low to moderate overall reliability may also help explain why we see sufficient evidence for only a few differences between single- and double-blind review. That being said, we did find credible evidence of systematic differences between single- and double-blind review in the differential impact of author characteristics on reviews. These differences manifest in how individual characteristics were used or associated with the reviews. These differences, taken together with the relevant outcomes we collected, help identify the limitations of each system and some of the causes of these limitations.

One variable that had differential impact on single- vs. double-blind review was coauthor seniority. In single-blind review, we found that review ratings were higher for submissions with more senior coauthors (on average, an 8% increase in the rating for each 1-step increase in average rank). This relationship was not credibly present in double-blind review. Notably, talks with more senior coauthors had lower attendance, were not judged as better talks, and were not more likely to be published. Given that conference-goers did not favor these talks, the partiality among single-blind reviewers is not necessarily preference-based but may instead reflect a miscalibrated belief about the demand for and likely eventual quality of full conference talks from more senior coauthors. These results suggest that single-blind review may help advance a rich-get-richer dynamic known as a *Matthew Effect* for coauthors: submissions with more eminent scientists receive more credit

relative to an unknown researcher (Merton, 1968). This favoritism towards established scholars could potentially contribute to the demographic inertia of fields in which some groups are under-represented among senior scholars (Hargens & Long, 2002) and impede steps to foster diversity without improving conference talk quality. However, because our test only involves a single field study in a specific context, more research on this question is needed before drawing strong conclusions.

We also uncovered some evidence that double-blind review might help address some other potential inequities. Controlling for 11 other author and submission characteristics, including whether the author was a U.S. citizen or not, Asian first authors received an 18% higher review rating than submissions with White first authors in double-blind review compared to single-blind review. These results are consistent with other evidence that scientific submissions by Asian authors may be disadvantaged in unblinded evaluations. For instance, grant applications to the National Institutes of Health with Asian investigators were 3.9 percentage points less likely to be funded than those by White investigators (Ginther et al., 2011). Similarly, hypothetical prospective doctoral students with Chinese or Indian names were substantially less likely to get a response and/or a meeting from a perspective faculty mentor, especially when requested at a later date (Milkman et al., 2012). A strength of our study is that we also have outcome data for at least some of the submissions, allowing us to understand better the potential mechanisms behind these differences and the consequences. In our study, when submissions with Asian authors were ultimately presented, the talks also tended to be judged higher in quality. This pattern of results implies that single-blind reviews may be impacted by miscalibrated beliefs or a dispreference for submissions by Asian first-authors, while double-blind review may result in more equitable merit-based evaluation.

However, more data is needed to assess the consequences of review type by race more thoroughly. Our initial pre-registration did not include race and ethnicity as a predictor, because we did not expect to have enough diversity in submissions to examine this question. Indeed, there were too few submissions at this conference from authors identified as Black, Hispanic, or Native American to be able to evaluate differential impact of review type (Table ??). The posterior distributions of the effect sizes reflect these low numbers for these race and ethnicity groups (see for example Tables ??, ??). These low numbers of submissions reveal a potentially deeper issue, the limited diversity of the people actively engaged in the science of judgment and decision making. This is a general problem throughout science, sometimes referred to as the pipeline problem (Allen-Ramdial & Campbell, 2014; Brown et al., 2016; Gibbs Jr et al., 2014). The results of our study suggest that double-blind review may help by reducing inequities that may occur in single-blind peer review, and because people may perceive double-blind review as fairer (as our survey results suggest), which can promote engagement with an organization.

Our results also suggest that the anonymity of double-blind review may not be a cure-all for the potential inequities in merit-based evaluations. We found that male first-authors had higher ratings than females in double-blind review, as compared to single-blind review. We did not anticipate this specific difference between single- and double-blind review and, as we established, it is a small effect. The result may also be specific to our context, an organization that had historically been predominantly male but had become more gender balanced in recent years. Nevertheless, it is notable that a similar pattern has been reported

previously, where women did worse when employers use anonymous review of job applicants' résumés (Behaghel et al., 2015; Krause et al., 2012). This potential for male submissions to score higher in double-blind review suggests that potential gender differences, at least in the fields related to judgment and decision making, may be due to something other than belief- or preference-based discrimination. Because double-blind reviewers cannot access author characteristics, we know they are not using author gender as an explicit cue. Instead, they must be reacting to some cues in the abstract that correlate with the author's gender. Our analyses control for only a limited number of factors, including submission sentiment, three research topics, and reviewer characteristics. It would be useful for future research to identify the relevant gender-related cues. Moreover, as we established, submissions with male first authors were also more likely to be published, which seems consistent with review ratings.

One possibility is that gender differences in personal and professional resources could lead to differences in the judged quality of the work (Ceci & Peters, 1984). For instance, there is evidence that gender asymmetries exist both in institutional resources (Duch et al., 2012), including time commitments and requests at work (Xie & Shauman, 2003), and in personal resources, as a result of differential family obligations (Ginther & Kahn, 2009). Another possibility may be that, although our analyses attempted to control for submission topic and some aspects of language, there may be differences in more specific topics, familiarity with research programs, the goals or positioning of the research, or in writing styles between male and female authors, with reviewers preferring those used by men. Further research is needed to resolve these differences. Whatever the cause, it seems possible that single-blind reviewers may have used their beliefs about the gender of the first authors to counteract inequities between men and women, in an attempt to ensure better representation (see also Williams & Ceci, 2015). If this was the case, then double-blind review may reduce the representation of certain perspectives or types of work that are correlated with gender, impeding attempts to improve equity. Our results suggest the need to carefully consider the potential consequences of how review system are designed and used.

Overall, we did not find differences in the validity of single- vs. double-blind review, in terms of outcomes for talks presented at the conference or eventual publication. The lack of higher validity for single-blind review is particularly striking, given that many of the outcomes (conference presentation and review processes at some journals) make the author identities public, which could result in a structural similarity, and even shared biases, between these outcomes and single-blind review.

Evidence-based Recommendations for Blinding Merit-based Evaluations

The results from our comparison of single- and double-blind review have important implications for the design and use of peer review for merit-based evaluations, not only of scientific work but in many other managerial situations as well. Arguably, our study's social context involved a relatively small network of rather densely connected peers. These social conditions are likely to be present in many other settings that require merit-based performance evaluations performed by peers with shared expertise, including employment, consumer purchases, and legal cases (Robertson & Kesselheim, 2016). Thus, we suspect our results are informative for these types of settings as well, although further research across different contexts is needed to assess generalizability.

There were two surprising sets of results that suggest reconsidering how peer review should be used in decision-making. First, we did not expect single- and double-blind review to be so similar in terms of their validity and reliability. Second, we also did not anticipate that double-blind reviews would give a slight edge to male-authored submissions. These results suggest the consequences of implementing blind-review, despite the greater perceived fairness, might be more nuanced than expected by its supporters. Some might even argue that these results support the use (or continued use) of non-anonymous (single-blind) review. Not only are single- and double-blind reviews similar, but double-blind review may impede reviewers' ability to regulate and counteract potential partiality towards submissions written by particular groups of people during merit-based evaluations, such as what appears to be a slight favoritism towards submissions written by men in our study.

We are skeptical of such arguments for single-blind review, for multiple reasons. First, some evidence suggests that the ability of reviewers to adjust for or counteract a particular partiality towards a characteristic may not occur when that characteristic is believed to be a valid predictor (e.g., coauthor seniority predicting merit), as correcting would seem counterintuitive (Rabinovitch et al., 2020). Second, single-blind review in our study favored submissions with more senior co-authors and disfavored Asian authors, unjustifiably so based on our outcome data. Third, even if single-blind review is used, an overwhelming majority of members of the society reported that they perceived single-blind review as less fair than double-blind review ($\beta = 0.54$, [0.42, 0.65]). Such perceptions of unfairness are likely to apply in other settings as well, and can have detrimental effects on organizational satisfaction, commitment, and citizenship behaviors (Konovsky, 2000). Conversely, merit-based evaluations that are perceived as fairer can attract people to organizations, which may help address some of the racial diversity issues we identified earlier in terms of the sparsity of the numbers (Branscombe et al., 1999; Chapman et al., 2005; Ryan & Ployhart, 2000).

Fourth, to the degree that single-blind review is being justified by its ability to allow reviewers to better correct for some unwanted partiality towards particular groups, there may be better methods for achieving such goals than to enable reviewers to implement ad-hoc adjustments. For instance, a more transparent process, such as invited submissions or speakers, can promote gender balance in a conference program. Finally, as we have shown, the author information available to reviewers during single-blind review did not show a strong, consistent predictive relationship across the measured outcomes. Thus, it is unclear why non-predictive information should be included, particularly when the individuating details create an opportunity for stereotypes or deference to eminence to enter the judgment (Biernat et al., 1991).

Our results also imply that double-blind review is not a cure-all for the problems of single-blind peer review. The tendency for male-authored submissions to score slightly higher with anonymous reviews illustrates the possibility that double-blind processes can leave systemic hurdles unaddressed and potentially disadvantage certain perspectives or types of work. This suggests that at the very least, organizations that implement double-blind review should continue to collect and monitor demographic data and, if possible, relevant outcome data. When an organization has the goal of promoting equity, not just equality, merit-based evaluations may be initially done via double-blinded review, with additional adjustments for particular individuals or groups who are disadvantaged by systemic

factors.

A novel approach to implementing double-blind review would be to use it in a way that reflects the reliability of the information in the merit-based evaluation itself. If the review systems exhibited no reliability, then it would seem justified to do away with peer review entirely and use a lottery to select submissions for scientific conferences (Thorngate et al., 2010). However, we found moderate reliability, which translated into the ability of the reviewer ratings to predict publication two years out. Moreover, we found greater agreement about what constitutes a weak submission than a strong one (see also Cortes & Lawrence, 2021). This pattern of results implies that there may be a way to blend double-blind peer review with lotteries: an *informed lottery*. In this approach, double-blind review would be used to separate between meritorious and non-meritorious submissions. This would leverage double-blind reviews' capability to distinguish the top from the bottom half of the submissions while excluding uninformative and potentially biasing details, such as personal characteristics and prior credentials. Acceptances would then be randomly selected from this meritorious set (for a similar suggestion see also F. C. Fang & Casadevall, 2016). The criterion used to split the submissions can be further adjusted in light of equity goals. In our study, for example, a median split on the reviewer rating largely minimized any advantage for male first authors. We contend that an informed lottery for scientific conferences 1) recognizes that reviews are most diagnostic for identifying weaker (as opposed to the strongest) research, 2) creates a more equitable and fair playing field for scholars, and 3) perhaps even fosters the evolution of science (Hull, 1988).

Conclusion

Adjusted double-blind review and an informed lottery certainly are not the only selection mechanisms for merit-based reviews. Other selection processes may involve statistical models, either alone or integrated with human judgment via hybrid modeling (Dawes et al., 1989). Nevertheless, we believe our study illustrates the value of moving away from default selection practices. Instead, the selection and design of merit-based evaluation practices should be evidence-based, using a head-to-head comparison of the alternatives, ideally involving experimentation in the context in which the practices are to be implemented (Goswami & Urminsky, 2020). Policy decisions should consider the validity, reliability, and differential impact of different characteristics of the evaluation procedures to fully remove unjustified impediments to advancing equity. Doing so requires a periodic examination of processes. Indeed, for the Society for Judgment and Decision Making, this investigation contributed to changes in the collection of demographic data for author submissions and society membership and a shift from single-blind to double-blind review for conference submissions. We encourage all organizations using merit-based review processes to conduct a similar evaluation of their own practices.

Our investigation of a highly structured and formal peer review process may be informative of processes in a wide range of less formal evaluation contexts in which interconnected peers attempt to assess future performance from partial information. Our results suggest that such processes are likely to be highly noisy and imprecise and, as a result, differences in evaluation outcomes that seem to arise from structural changes may instead be caused by low reliability. Our findings also point to the potential for very different simultaneous effects of non-anonymized reviews: over-favoring individuals and institutions with

stronger reputations, disfavoring some under-represented groups that may be mistakenly seen as lower in quality while simultaneously not disfavoring or even favoring groups for whom equitable treatment is a more salient goal. Furthermore, such differential treatment may be strong enough to not be justifiable even in relation to future non-anonymized but more information-based outcomes. Overall, our findings point to fundamental limitations in evaluations based on limited information: accuracy is likely to be low, but incorporating individuating information may often introduce bias without improving accuracy.

References

- Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *ILR Review*, 30(2), 175–187. doi: 10.1177/001979397703000204
- Allen-Ramdiel, S.-A. A., & Campbell, A. G. (2014). Reimagining the pipeline: Advancing stem diversity, persistence, and success. *BioScience*, 64(7), 612–618. doi: 10.1093/biosci/biu076
- Arkes, H. R., Shaffer, V. A., & Dawes, R. M. (2006). Comparing holistic and disaggregated ratings in the evaluation of scientific presentations. *Journal of Behavioral Decision Making*, 19(5), 429–439.
- Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton University Press.
- Åslund, O., & Skans, O. N. (2012). Do anonymous job application procedures level the playing field? *ILR Review*, 65(1), 82–107. doi: 10.1177/001979391206500105
- Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.
- Behaghel, L., Crépon, B., & Le Barbanchon, T. (2015). Unintended effects of anonymous resumes. *American Economic Journal: Applied Economics*, 7(3), 1–27. doi: 10.1257/app.20140185
- Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments* (Vol. 1, pp. 309–393). Elsevier. doi: 10.1016/bs.hefe.2016.08.004
- Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, 60(4), 485. doi: 10.1037/0022-3514.60.4.485
- Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *The American Economic Review*, 1041–1067.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

- Bohren, J. A., Imas, A., & Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10), 3395–3436. doi: 10.1257/aer.20171829
- Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4), 1753–1794. doi: 10.1093/qje/qjw029
- Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PloS one*, 5(12), e14331. doi: 10.1371/journal.pone.0014331
- Branscombe, N. R., Schmitt, M. T., & Harvey, R. D. (1999). Perceiving pervasive discrimination among african americans: Implications for group identification and well-being. *Journal of personality and social psychology*, 77(1), 135–149. doi: 10.1037/0022-3514.77.1.135
- Brown, B. A., Henderson, J. B., Gray, S., Donovan, B., Sullivan, S., Patterson, A., & Waggstaff, W. (2016). From description to explanation: An empirical exploration of the african-american pipeline problem in stem. *Journal of Research in Science Teaching*, 53(1), 146–177. doi: 10.1002/tea.21249
- Budden, A. E., Tregenza, T., Aarssen, L. W., Koricheva, J., Leimu, R., & Lortie, C. J. (2008). Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, 23(1), 4–6. doi: 10.1016/j.tree.2007.07.008
- Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15(3), 75–141. doi: 10.1177/1529100614541236
- Ceci, S. J., & Peters, D. (1984). How blind is blind review? *American Psychologist*, 39(12), 1491.
- Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: a meta-analytic review of the correlates of recruiting outcomes. *Journal of applied psychology*, 90(5), 928–944. doi: 10.1037/0021-9010.90.5.928
- Cortes, C., & Lawrence, N. D. (2021). *Inconsistency in conference peer review: Revisiting the 2014 neurips experiment*. Retrieved from <https://doi.org/10.48550/arXiv.2109.09774>
- Cox, A. R., & Montgomerie, R. (2019). The cases for and against double-blind reviews. *PeerJ*, 7, e6702. doi: 10.7717/peerj.6702
- Crane, D. (1967). The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 195–201.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668–1674. doi: 10.1126/science.2648573

- de Leon, F. L. L., & McQuillin, B. (2020). The role of conferences on the pathway to academic impact evidence from a natural experiment. *Journal of Human Resources*, 55(1), 164–193.
- Duch, J., Zeng, X. H. T., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K., & Amaral, L. A. N. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PloS one*, 7(12), e51332. doi: 10.1371/journal.pone.0051332
- Fang, F. C., & Casadevall, A. (2016). *Research funding: The case for a modified lottery*. American Society for Microbiology. doi: 10.1128/mBio.00422-16
- Fang, H., & Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. , 1, 133–200.
- Fiske, D. W., & Fogg, L. (1992). But the reviewers are making different criticisms of my paper! diversity and uniqueness in reviewer comments. *American Psychologist*, 591–598. doi: 10.1037/10109-048
- Fryer, R., Jackson, M. O., et al. (2008). A categorical model of cognition and biased decision-making. *Journal of Theoretical Economics*, 8(1), 1–42.
- Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189–211. doi: 10.1080/19345747.2011.618213
- Gibbs Jr, K. D., McGready, J., Bennett, J. C., & Griffin, K. (2014). Biomedical science ph. d. career interest patterns by race/ethnicity and gender. *PloS one*, 9(12), e114736. doi: 10.1371/journal.pone.0114736
- Ginther, D. K., & Kahn, S. (2009). Does science promote women? evidence from academia 1973-2001. In *Science and engineering careers in the united states: An analysis of markets and employment* (pp. 163–194). University of Chicago Press.
- Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and nih research awards. *Science*, 333(6045), 1015–1019. doi: 10.1126/science.1196783
- Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4), 715–741.
- Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Conference presentations and academic publishing. *Economic Modelling*, 95, 228–254. doi: 10.1016/j.econmod.2020.12.017
- Goswami, I., & Urminsky, O. (2020). No substitute for the real thing: The importance of in-context field experiments in fundraising. *Marketing Science*, 39(6), 1052–1070. doi: 10.1287/mksc.2020.1252

- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1), 5228–5235. doi: 10.1073/pnas.0307752101
- Hammerschmidt, K., Reinhardt, K., & Rolff, J. (2008). Does double-blind review favor female authors? *Frontiers in Ecology and the Environment*, 6(7), 354–354. doi: 1540-9295(2008)6[354a:DDRFFA]2.0.CO;2
- Hargens, L. L., & Long, J. S. (2002). Demographic inertia and women’s representation among faculty in higher education. *The Journal of Higher Education*, 73(4), 494–517. doi: 10.1080/00221546.2002.11777161
- Hertwig, R., Pleskac, T. J., Pachur, T., & the Center for Adaptive Rationality. (2019). *Taming uncertainty*. Cambridge, MA: MIT Press.
- Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In *Aaai* (Vol. 4, pp. 755–760).
- Hull, D. L. (1988). *Science as a process: An evolutionary account of the social and conceptual development of science*. Chicago, IL: University of Chicago Press.
- Jowell, R., & Prescott-Clarke, P. (1970). Racial discrimination and white-collar workers in britain. *Race*, 11(4), 397–417.
- Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological review*, 100(1), 109–128. doi: 10.1037/0033-295X.100.1.109
- Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little Brown.
- Keren, G., & Wu, G. (2015). *The Wiley-Blackwell handbook of judgment and decision making*. John Wiley & Sons, Ltd.
- Konovsky, M. A. (2000). Understanding procedural justice and its impact on business organizations. *Journal of Management*, 26(3), 489–511. doi: 10.1177/014920630002600306
- Krause, A., Rinne, U., & Zimmermann, K. F. (2012). Anonymous job applications in europe. *IZA Journal of European Labor Studies*, 1(1), 1–20. doi: 10.1186/2193-9012-1-5
- Kruschke, J. K., & Liddell, T. M. (2018a). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. doi: 10.3758/s13423-016-1221-4
- Kruschke, J. K., & Liddell, T. M. (2018b). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. doi: 10.3758/s13423-016-1221-4
- Lawrence, N. (2015, January 16). *Nips experiment analysis*. Retrieved from <https://inverseprobability.com/2015/01/16/blogs-on-the-nips-experiment>
- MacCoun, R. (2020). Blinding to remove biases in science and society. In R. Hertwig & C. Engel (Eds.), *Deliberate ignorance: Choosing not to know*. MIT Press.

- Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, 12(3), 282–293. doi: 10.3102/10769986012003282
- Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810), 56–63. doi: 10.1126/science.159.3810.56
- Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological science*, 23(7), 710–717.
- Okike, K., Hug, K. T., Kocher, M. S., & Leopold, S. S. (2016). Single-blind vs double-blind peer review in the setting of author prestige. *Jama*, 316(12), 1315–1316. doi: 10.1001/jama.2016.11014
- Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, 62(4), 659–661.
- Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., ... Carnes, M. (2018). Low agreement among reviewers evaluating the same nih grant applications. *Proceedings of the National Academy of Sciences*, 115(12), 2952–2957.
- Rabinovitch, H., Bereby-Meyer, Y., & Budescu, D. V. (2020). Achieving more with less: Intuitive correction in selection. *Psychological Science*, 31(4), 437–448. doi: 10.1177/0956797620903717
- Robertson, C. T., & Kesselheim, A. S. (2016). *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law*. Academic Press.
- Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of management*, 26(3), 565–606.
- Thorngate, W., Dawes, R. M., & Foddy, M. (2010). *Judging merit*. Psychology Press.
- Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48), 12708–12713.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5), 1413–1432. doi: 10.1007/s11222-016-9696-4
- Whittaker, R. J. (2008). Journal review and gender equality: a critical comment on Budden et al. *Trends in Ecology and Evolution*, 23(9), 478–479. doi: 10.1016/j.tree.2008.03.003
- Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2: 1 faculty preference for women on stem tenure track. *Proceedings of the National Academy of Sciences*, 112(17), 5360–5365. doi: 10.1073/pnas.1418878112
- Xie, Y., & Shauman, K. A. (2003). *Women in science: Career processes and outcomes*. Harvard University Press.

Supplementary Material for “Blinded versus unblinded review: A field study on the equity of peer-review processes”

Timothy J. Pleskac¹, Ellie Kyung², Gretchen Chapman³, and Oleg Urminsky⁴

¹University of Kansas

²Babson College

³Carnegie Mellon University

⁴University of Chicago

July 31, 2023

Contents

1	Preregistration	2
2	Methods	3
2.1	Power Analyses	3
2.2	Conference Materials	6
2.3	Submission Characteristics	7
2.4	Materials	7
2.4.1	Information Collected at Submission.	7
2.4.2	Conflicts of Interest and Selection Process.	8
2.4.3	Collected Variables on Reviewers at Agreement.	8
2.4.4	Collected Variables from Reviewers During and After Completing Reviews.	9
2.4.5	Intercorrelations Between Reviewer Characteristics.	9
2.5	Talk Evaluation and Statistics.	10
2.5.1	Characteristics of Raters.	10
2.5.2	Rated Dimensions.	10
2.5.3	Judged Overall Rating.	10
2.5.4	Collected Talk Statistics.	11
2.6	Poster Ratings.	11
2.7	Preference for Review process Survey Items.	11
2.8	Journal Publications.	12
2.9	Analyses.	12
3	Modeling the Text of the Long Abstracts	12
3.1	Text Corpus	12
3.2	Topics Modeling	13
3.3	Sentiment Analysis	14
4	Survey of SJDM Members’ Preferences	15
4.1	Characteristics of Respondents	16
4.2	Preference Ratings	18
4.2.1	Regression Coefficients for Predictors of Double- vs. Single-blind Preference.	18
4.3	Fairness Perceptions	19

5	Reliability of Review Systems	19
5.1	Comparing Bottom vs. Top Half of Submission	19
6	Differential Use Between Single- and Double-blind Review	19
6.1	Model Comparison	19
6.2	Regression Coefficients With All Characteristics Entered Simultaneously	21
6.2.1	All Submissions.	21
6.2.2	Multiauthor Submissions.	23
6.3	Regression Coefficients With Each Variable Entered Alone	25
6.4	Regression Coefficients with All Variables Simultaneously	27
7	Comparing Single- vs. Double-blind in the Top 108	29
8	Predictive Validity of Review Ratings and Submission Characteristics	31
8.1	Summary of Talk Outcomes Across sessions	31
8.2	Correlations Between Outcomes	31
8.3	Regression of Outcomes onto Single- and Double-blind Review Ratings	32
8.4	Regression of Outcomes Onto Author and Submission Characteristics	33
8.4.1	Regression with Standardized Outcome Values.	33
8.4.2	Regression with Raw Outcome Values.	35
8.4.3	Regression with Multi-author Submissions.	36
8.4.4	Regression with Raw outcome Values for Muti-author Submissions.	38

1 Preregistration

The study was preregistered. It can be found at the OSF at this link: [OSF Peer Review Link](#). There are seven deviations from the pre-registration that we should note here. The first deviation was in the analysis of the reviewer ratings. Our plan was to use the following abstract characteristics in the analyses: disciplinary area, number of studies included, word count, and keywords. We deviated from the plan as disciplinary area was not asked and it was not feasible to classify authors ex post. The number of studies was not a useful statistic as some studies had 0 and some had many and some abstracts were not empirical studies. In terms of word count, nearly all the abstracts were 600 words and thus there was little to no variability. Finally, in terms of keywords, although authors were given a list of keywords, authors often chose their own unique set of keywords. Moreover, to make use of the keywords we needed a method to group and classify them. To do this, we turned to methods from Natural Language Processing and once we did this it seemed a better measure of the topics was to model the text of the long abstract directly. Thus, we used the topics model of the abstract as an indicator of key topics (Section 3). We also included a measure of the sentiment of the abstract. The results reported in the paper are unchanged with the sentiment variable excluded from the models.

The second deviation was in the measure of the judged talk quality. Our initial plan was to have the talks rated on importance, methodological quality, persuasiveness, originality, organization, and degree of match between short and long abstract. As we finalized the talk rating instrument we modified the dimensions to the following set of dimensions: significance, methods, results, conclusion, innovation, and uniqueness. We found these dimensions to be more straightforward and easier to rate. Moreover, in the analysis, we averaged across these ratings and across all the raters to achieve a reliable measure of judged talk quality. By collapsing across raters we did not include rater characteristics into the ratings. But, all ratings were standardized within raters, which should help minimize differences.

The third deviation was that initially we did not plan to include race and ethnicity of the authors in the regressions as we did not expect sufficient numbers in each race and ethnicity category, though we did collect this information. Our expectation was due to the fact that our race and ethnicity categories had to reflect the international status of the society and thus there were many categories. After the first round of review of the paper, reviewers suggested that this variable be included. To include the variable, we grouped different categories together. For instance, we created one group we referred to as Asian or of Asian decent that included people that reported being “Arab, Caucasian, White or European”, or “Asian, Caucasian, White

or European”, “Asian”, “Asian, Caucasian, White or European, Other (please specify)”, or “Asian, Other (please specify)”. In total, we created 7 groups: “Asian”, “Black, African, or African American”, “Hispanic or Latino Origin”, “Native American or Alaska Native”, “Caucasian, White, or European”, “Other”, and “Prefer not to answer.”

The fourth deviation was our measure of the institutional prestige for all the authors. Our initial plan was to use the count of the number of talks in the previous 5 years of the SJDM conference for the respective institution. However, this variable was highly skewed with a few institutions having most of the talks and many other institutions with very few to no talks. Therefore, we used the average number of citations per paper from the Psychology, Economics and Business, and Social Science departments at the institution of the focal author as determined by the Web of Science. This average was downloaded on August 29, 2019.

A fifth deviation was to use Pearson correlations as a measure of reliability instead of a rank order correlation. As the ratings were standardized within each reviewer, using a rank order or Pearson correlation made little difference. We chose the latter as it was mathematically and statistically more feasible to work with.

A sixth deviation was to use Bayesian regression models with weakly informative priors. As our primary focus was on quantifying effect sizes this Bayesian approach facilitates this better providing moderate regularization to the estimates and helping reduce errant estimates. We report credible intervals around mean posterior estimates instead of maximum likelihood estimate effects with confidence intervals. Posterior distributions do not change based on the number of planned or unplanned tests. Thus, there is no correction for the number of tests as the statistics and inferences from them are based on the posterior distribution Gelman et al. (2012a); Kruschke & Liddell (2018a). Another reason we adopted a Bayesian approach as it permitted a more rigorous model comparison approach based on leave-one-out cross-validation (Vehtari et al., 2017). Note the posted primary regression analyses on [OSF](#) have a frequentist version of the multi-level model posted for the key analyses in terms of bias. The conclusions remain the same with conventional p-values based on our planned comparisons, but if we include corrections for multiple comparisons the following interactions are not significant: the interaction between gender and review condition, the interaction between senior coauthorship and review condition, and the interaction between first author race and review condition. Again our focus is not on null hypothesis significance testing, but on quantifying the effects; thus we focus on the posterior distribution over the effects.

A seventh deviation is that when writing the pre-registration we failed to account for the fact that single-author submissions by definition, do not have characteristics for co-authors. Thus, to include single-author submissions in the regression analyses, for all the submissions, we replaced the coauthor variables with variables for all authors (e.g., the proportion of male authors instead of male coauthors). However, this had the effect of losing a clean measure of coauthor characteristics (e.g., the proportion of co-author males). Thus, we also ran a second set of all of our regression analyses with multiauthor submissions.

2 Methods

2.1 Power Analyses

There are several issues to consider for a power analysis for our study. All of our primary analyses and conclusions are based on Bayesian multi-level regressions. A power analysis with multi-level regression is not straightforward, entailing many assumptions. Moreover, our goal in the field study was a quantitative approach focusing on the effect sizes for all the collected variables. This quantitative approach combined with our Bayesian methods implies we are not relying on null hypothesis significance testing but focusing on the posterior distribution. To highlight particular results, we ask whether the 95% CI excludes 0 and used the term *credible* to describe the effect. These credible effects are the ones we focus on in the paper. Finally, as our study was a field study, we had little control over the sample size with our only goal being to maximize the number of observations that we could use in our analyses given the number of submissions that were entered and the number of talks at the conference.

Given these qualifiers, we conducted a Bayesian power analysis focusing on our lowest powered test: the Pearson r correlation. We designed the power analysis in line with Kruschke & Liddell (2018a) evaluating the probability of obtaining a particular goal. In this case the goal was if a 95% credible interval for a given posterior distribution to exclude 0.

One set of analyses focused on the posterior distribution over the Pearson correlation r and the probability that a 95% credible interval would exclude 0. To estimate this probability, simulated data were created by generating bivariate normally distributed random values each with a mean of 0 and a standard distribution of 1. We varied the correlation between the two variables between $\rho = 0.10$ and $\rho = 0.40$. In calculating the posterior distribution of the Pearson r for a given dataset, we used the R package **correlationBF** (Morey & Rouder, 2018). This is the function we used in all of our analyses to calculate correlations. Noninformative priors are assumed for the population means and variances of the two populations. A beta distribution (shifted) is assumed for the population correlation ρ with the 2 parameters of the beta distribution were set to the default values of $\frac{1}{3}$ creating a “medium” spread in the beta distribution centered over 0 and extending between -1 and 1. The code for the power analysis is on the OSF site for the paper.

Table 1 lists the proportion of times a 95% credible interval excluded 0 for a range of correlations. We calculated these proportions for a range of different sample sizes that reflect the different sample sizes used in the paper. The sample sizes included the number of submissions that were accepted as talks ($n = 108$), the number of multiauthor submissions with a complete set of variables used in the regressions ($n = 370$), the number of multiauthor submissions for which the first author was the corresponding author ($n = 430$), the number of submissions for which the first author was the corresponding author ($n = 470$), and the number of submissions ($n = 530$). The table shows that for analyses other than those that focused only on the submissions that were given as talks, data generated with a correlation of $\rho = .15$ had greater than an 80% chance of being identified as credible. For analyses that focused only on the submissions that were given as talks (i.e., when $n = 108$) data generated with a correlation of $r = .27$ had an 80% chance of being identified as credible. Convention in psychology has been to treat a correlation of .10 as a weak to small association and values of .30 as moderate associations. Thus, field study was adequately powered to detect small to moderate associations with our lowest powered test.

The focus of the study was the difference between single- and double-blind. These comparisons were primarily done via interaction terms via multilevel regressions. But, the lowest powered comparison is via a difference in correlations. We also ran these power analyses. To do so, we generated two sets of bivariate normal variables each with the same given sample size. One pair of variables was generated with a correlation of $\rho = .15$. The other pair of variables was generated with a range of different correlations starting with $\rho = .25$ and ranging up to $\rho = .55$.¹ Then for each pair of variables we calculated the posterior distribution over the Pearson correlation and took the difference in the two distributions. Table 2 lists the proportion of times (out of 10,000 repetitions) that the 95% credible interval of the posterior distribution of differences in correlation contained 0 for a given difference in the correlation. The table shows that for the analyses on the larger set of submissions (i.e., other than the ones focused on the submissions given as talks), a sample of data with differences in correlations of $\Delta\rho = 0.20$ or larger was identified as a credible difference in the Pearson correlation greater than 80% of the time. For the analyses based on the smaller sample of submissions given only as a talk, the corresponding difference in correlations was $\Delta\rho = 0.35$. Again we conclude that our field study was adequately powered to detect small to moderate differences in associations with our lowest-powered test.

These are baseline estimates from correlations and differences in correlations and represent the lowest powered tests. Most of our analyses and conclusions are based on more powerful regressions that increase the power of detecting if a variable was credibly associated with a dependent variable of interest. There were several aspects by-which our regressions increased the power. First, we used simultaneous regressions controlling for 11 author and submission characteristics and 5 reviewer characteristics. Second, each submission received at least 3 single-blind and 3 double-blind reviews. This property created a within-submission comparison of single- and double-blind review. In our statistical models, we accounted for this via multi-level regressions making submission a random effect (a random slope). Because each reviewer reviewed approximately 30 submissions, we also could treat the reviewer as a random effect. Similarly in the outcome data the models accounted for random effects in the session as well as other variables. Third, we used vague but informed priors to regularize our regressions and results to help guard against overly sensitive analyses of extreme data. Interaction effects in multiple regressions are essentially differences in correlations, so these same statements are true for the interaction effects.

Altogether, we believe our field study is informative and well-powered for the debate on single- vs. double-

¹These values were relatively arbitrary and adjusting these values within the range of intermediate correlations ($-.5 \geq \rho \geq .5$) did not impact the estimates as the power analysis was based on the difference in correlations.

Table 1: Proportion of times a 95% credible interval for the posterior distribution over the Pearson-r correlation ρ excluded 0 for a sample of multivariate-normal data generated with a correlation.

Correlation	Prop. Excludes 0				
	n = 108	n = 370	n = 430	n = 456	n = 530
0.10	0.1676	0.4878	0.5425	0.5667	0.6321
0.11	0.1829	0.5530	0.6237	0.6512	0.7169
0.12	0.2213	0.6322	0.6996	0.7279	0.7858
0.13	0.2479	0.7072	0.7671	0.7893	0.8570
0.14	0.2885	0.7670	0.8310	0.8463	0.9012
0.15	0.3185	0.8159	0.8814	0.8943	0.9390
0.16	0.3737	0.8720	0.9123	0.9258	0.9544
0.17	0.3997	0.9091	0.9441	0.9536	0.9749
0.18	0.4531	0.9410	0.9634	0.9693	0.9877
0.19	0.4849	0.9558	0.9800	0.9853	0.9933
0.20	0.5402	0.9708	0.9872	0.9897	0.9962
0.21	0.5727	0.9849	0.9923	0.9948	0.9985
0.22	0.6152	0.9886	0.9970	0.9972	0.9991
0.23	0.6575	0.9937	0.9982	0.9984	0.9970
0.24	0.6996	0.9970	0.9989	0.9988	0.9970
0.25	0.7375	0.9975	0.9996	0.9996	1.0000
0.26	0.7704	0.9993	0.9999	0.9997	1.0000
0.27	0.8019	0.9997	1.0000	1.0000	1.0000
0.28	0.8306	1.0000	1.0000	1.0000	1.0000
0.29	0.8596	1.0000	1.0000	1.0000	1.0000
0.30	0.8786	1.0000	1.0000	1.0000	1.0000
0.31	0.8990	1.0000	1.0000	1.0000	1.0000
0.32	0.9216	1.0000	1.0000	1.0000	1.0000
0.33	0.9317	1.0000	1.0000	1.0000	1.0000
0.34	0.9489	1.0000	1.0000	1.0000	1.0000
0.35	0.9598	1.0000	1.0000	1.0000	1.0000
0.36	0.9693	1.0000	1.0000	1.0000	1.0000
0.37	0.9749	1.0000	1.0000	1.0000	1.0000
0.38	0.9831	1.0000	1.0000	1.0000	1.0000
0.39	0.9875	1.0000	1.0000	1.0000	1.0000
0.40	0.9898	1.0000	1.0000	1.0000	1.0000

The proportion of times the 95% credible interval excluded 0 was calculated for multiple sample sizes including the number of submissions that were accepted as talks ($n = 108$), the number of multiauthor submissions with a complete set of variables used in the regressions ($n = 370$), the number of multiauthor submissions for which the first author was the corresponding author ($n = 430$), the number of submissions for which the first author was the corresponding author ($n = 470$), and the number of submissions ($n = 530$). The proportions are based on 10,000 repetitions of generating a sample of multivariate normal random variables with a given sample size and a given correlation.

blind review. We explicitly chose to study this question in the field as opposed to a lab study as the field seems the best test of these different systems of peer-review. We also worked to report our results in the most informative manner possible, providing measures of effect sizes with credible intervals for all the results.

Table 2: Proportion of times a 95% credible interval for the posterior distribution over the difference in two Pearson-r correlations ρ excluded 0 for a sample of multivariate-normal data generated with a correlation.

Difference in Correlations	Prop. 95% Credible Interval Excludes 0				
	n = 108	n = 370	n = 430	n = 456	n = 530
0.10	0.1095	0.2944	0.3291	0.3440	0.3892
0.11	0.1251	0.3448	0.3861	0.4136	0.4612
0.12	0.1491	0.3925	0.4403	0.4717	0.5249
0.13	0.1591	0.4529	0.5139	0.5422	0.5883
0.14	0.1828	0.5138	0.5735	0.6036	0.6685
0.15	0.1937	0.5682	0.6472	0.6642	0.7345
0.16	0.2186	0.6172	0.6963	0.7289	0.7891
0.17	0.2445	0.6838	0.7562	0.7740	0.8313
0.18	0.2720	0.7364	0.8038	0.8192	0.8708
0.19	0.2988	0.7854	0.8393	0.8661	0.9074
0.20	0.3263	0.8264	0.8830	0.8985	0.9363
0.21	0.3699	0.8675	0.9051	0.9267	0.9545
0.22	0.3895	0.8935	0.9335	0.9482	0.9726
0.23	0.4230	0.9230	0.9498	0.9637	0.9794
0.24	0.4539	0.9426	0.9708	0.9752	0.9891
0.25	0.4901	0.9577	0.9764	0.9828	0.9915
0.26	0.5238	0.9730	0.9865	0.9885	0.9972
0.27	0.5689	0.9785	0.9899	0.9948	0.9979
0.28	0.5981	0.9840	0.9954	0.9959	0.9995
0.29	0.6317	0.9916	0.9973	0.9983	0.9992
0.30	0.6690	0.9946	0.9984	0.9990	0.9999
0.31	0.6982	0.9965	0.9992	0.9994	0.9999
0.32	0.7269	0.9968	0.9994	0.9999	0.9999
0.33	0.7582	0.9989	0.9998	1.0000	1.0000
0.34	0.7857	0.9992	0.9999	1.0000	1.0000
0.35	0.8139	0.9997	1.0000	1.0000	1.0000
0.36	0.8433	1.0000	1.0000	1.0000	1.0000
0.37	0.8590	1.0000	1.0000	1.0000	1.0000
0.38	0.8875	1.0000	1.0000	1.0000	1.0000
0.39	0.9031	0.9999	1.0000	1.0000	1.0000
0.40	0.9209	1.0000	1.0000	1.0000	1.0000

The proportion of times the 95% credible interval excluded 0 was calculated for multiple sample sizes including the number of submissions that were accepted as talks ($n = 108$), the number of multiauthor submissions with a complete set of variables used in the regressions ($n = 370$), the number of multiauthor submissions for which the first author was the corresponding author ($n = 430$), the number of submissions for which the first author was the corresponding author ($n = 470$), and the number of submissions ($n = 530$). The proportions are based on 10,000 repetitions. To estimate the power two samples of a given sample size of multivariate normal random variables were generated. One sample was generated with a correlation of $\rho = .15$ and the other with a range of correlations starting at $\rho = .25$ up to $\rho = .55$. Thus, the difference ranged from $\Delta = 0.10$ to $\Delta = 0.40$. The specific values for ρ were relatively arbitrary and changes in the values (so long as they stayed roughly between $\rho \approx -0.50$ and $\rho \approx 0.50$) had little impact on the power of identifying a credible difference.

2.2 Conference Materials

The call for conference submission is available at this link <http://www.sjdm.org/programs/2018-cfp.html>. The program for the conference is available at this link <https://sjdm.org/programs/2018-program.pdf>.

Table 3: Characteristics of authors and reviewers.

		First Author	coauthors	Reviewers
Gender	Female	237 (45%)	355 (38%)	50 (44%)
	Male	288 (54%)	571 (61%)	63 (56%)
	Not reported	5 (1%)	13 (1%)	-
Race/Ethnicity	Arab	-	-	1 (0.9%)
	Arab, Caucasian, White or European	1 (0.1%)	-	-
	Asian	109 (20.6%)	-	12 (10.6%)
	Asian, Caucasian, White or European	3 (0.6%)	-	1 (0.9%)
	Asian, Caucasian, White or European, Other	3 (0.6%)	-	-
	Asian, Other	1 (0.2%)	-	-
	Black, African, or African American	4 (0.8%)	-	-
	Caucasian, White or European	325 (61.3%)	-	57 (50.4%)
	Caucasian, White or European, Other	1 (0.2%)	-	-
	Hispanic or Latino origin	6 (1.1%)	-	-
	Hispanic or Latino origin, Caucasian, White or European	7 (1.3%)	-	3 (2.7%)
	Native American or Alaska Native, Caucasian, White or European	1 (0.2%)	-	-
	Other	12 (2.3%)	-	1 (0.9%)
	Not reported	57 (11%)	-	38 (45.5%)
Citizenship	U.S.	241 (45%)	-	56 (50.0%)
	Non-U.S.	289 (55%)	-	27 (23.9%)
	Not reported	-	-	30 (26.5%)
Ave. Year of Ph.D. (SD)		2013 (8)	-	2008 (7)
Area of Ph.D.	Economics	39 (7.4%)	-	14 (12.4%)
	Management	57 (10.8%)	-	13 (11.5%)
	Marketing	121 (22.8%)	-	25 (22.1%)
	Psychology	253 (47.7%)	-	52 (46.0%)
	Other	60 (11.3%)	-	8 (7.1%)
	Not reported	-	-	1 (0.9%)
Position	Student	0	2 (0.2%)	-
	Undergraduate	1 (0.2%)	17 (1.8%)	-
	Masters Student	8 (1.5%)	20 (2.2%)	-
	Doctoral Student	196 (37.0%)	96 (10.3%)	-
	Postdoctoral Researcher	70 (13.2%)	36 (3.9%)	-
	Research Scientist	17 (3.2%)	48 (5.2%)	5 (4.4%)
	Practitioner	12 (2.3%)	27 (2.9%)	2 (1.8%)
	Assistant Professor	133 (25.1%)	217 (23.4%)	48 (42.4%)
	Associate Professor	48 (9.1%)	144 (15.5%)	31 (27.4%)
	Full Professor	41 (7.7%)	321 (34.6%)	27 (23.9%)
	Not reported	4 (0.8%)	-	-

Many people selected multiple race and ethnicity categories and instead of marginalizing across categories we report the joint selections. Except for gender and position, the characteristics of the first author are based on the subset of submissions when the corresponding author was the first author.

2.3 Submission Characteristics

A summary of the characteristics of the submitting authors is listed in Table 3.

2.4 Materials

2.4.1 Information Collected at Submission.

The call for submissions for the 2018 Annual Meeting for the Society for Judgment and Decision Making invited members and non-members to submit a 600-word long-abstract to be considered for a spoken presentation. Submissions also included a title, a 100-word short abstract (to be used in the conference program), a listing of the authors, up to 3 keywords, and an indication as to whether the submission should be considered for a poster if it was not selected for a talk. Authors could also request to present a poster in which case they were only asked to submit a title, 100 word abstract for the program, and a list of authors. Here we focus only on the submissions for a spoken presentation.

At the time of submission, we also collected the following information from all authors:

Position Undergraduate student; Masters student; PhD student; Post doctoral researcher; Research scientist; Lecturer; Assistant Professor; Associate Professor; Full Professor; Practitioner; Other; Prefer not to answer

Department Text answer

Institution Text Answer

Gender Female; Male; Other (please specify); Prefer not to answer

The submitting or corresponding author was also asked to report the following information

Highest degree obtained or in progress Bachelors; Masters; Ph.D.; MD; MPH; Other; Prefer not to answer

Primary area of study for highest degree completed or in progress Economics; Marketing; Management; Psychology (Cognitive); Psychology (Social); Psychology (Other); Statistics; Other (Please specify); Prefer not to answer

Department and institution where the degree was obtained Text Answer

Graduation year for this degree Four digit year

Race/ethnic group Arab; Asian; Black, African, or African American; Hispanic or Latino origin; Native American or Alaska Native; Native Hawaiian or Other Pacific Islander; Caucasian, White, or European; Other (please specify); Prefer not to answer

Nationality List of countries

Table 4: Distribution of the number of authors.

Number of Authors	Frequency
1	26
2	218
3	176
4	78
5	26
6	6

2.4.2 Conflicts of Interest and Selection Process.

Across all assignments, we worked to minimize conflicts of interest by ensuring no submission was assigned to a reviewer who shared a common institution with one of the institutions of the submitting authors. In addition, upon sending reviewers their assigned submissions, they were asked to inspect the submissions and identify any conflict so that it could be reassigned. If a conflict became apparent during the review, then reviewers were instructed to leave the abstract unrated if a conflict became apparent during the review. However, no submission was left unrated. The ratings were then standardized within each reviewer and aggregated across all six ratings forming an average reviewer rating for each submission. The Conference Chair (XXX) used the average ratings to decide on accepting a submission for presentation. In the case of the 53 submissions that had two sets of reviews from each process, for the purpose of making a decision one set of three double-blind and one set of three single-blind reviews were randomly selected to aggregate across.

2.4.3 Collected Variables on Reviewers at Agreement.

Besides rating each of their assigned submissions on a 1 *poor* to 9 *excellent* point scale and entering any comments they might have, reviewers were also asked a series of questions. At agreeing to review they were asked:

Institution Current academic institution

Department Current academic department

Position Current position

Area of study of highest degree Primary area of study for highest degree completed or in progress: (If you ave more than one PhD in a JDM-related field, select the area for your first PhD)

Year of highest degree Graduation year for this degree (enter all 4 digits, i.e, 2008)

Department and school where they obtained their highest degree

Reviewing experience How many times they had reviewed for the annual meeting in the past

Gender Female; Male; Other (please specify); Prefer not to answer

Nationality Dropdown list of countries

Race Arab; Asian; Black, African, or African American; Hispanic or Latino origin; Native American or Alaska Native; Native Hawaiian or Other Pacific Islander; Caucasian, White or European; Other (please specify); Prefer not to answer

2.4.4 Collected Variables from Reviewers During and After Completing Reviews.

During the review, reviewers were asked to rate each abstract on a nine-point scale ranging from 1 (Poor) to 9 (Excellent). There was also a field to enter any submission comments.

After their reviews, these questions asked them the following items.

Overall confidence How confident are you that your evaluations of the abstracts are an accurate rating of their actual quality? 0 = 0% confident; 10 = 100% confident

How did you complete the reviews? To what extent did you evaluate the abstracts? 1 = Fully on paper; 6 = Fully on screen

When did they complete the reviews? To what extent did you evaluate the abstracts: 1 = Within 1-2 days; 6 = Over several weeks

Did you revise while you reviewed? To what extent did you evaluate the abstracts: 1 = Without any revisions to ratings; 6 = With significant revisions to ratings

Did you consult with colleagues? Did you consult with any colleagues while you were reviewing because you had questions about the process? Yes; No

Comments Text

2.4.5 Intercorrelations Between Reviewer Characteristics.

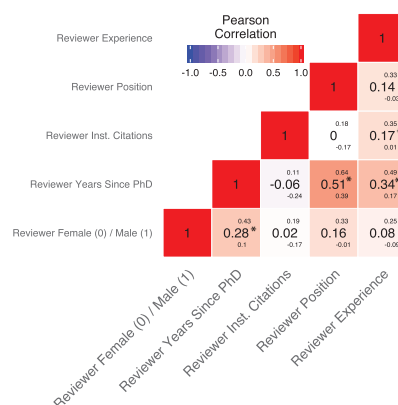


Figure 1: Intercorrelations between reviewer characteristics. Values in the center of each cell are posterior estimates of the Pearson correlations and values in the upper and lower right-hand corner are the corresponding credible intervals. Starred values indicate credible interval excludes 0.

2.5 Talk Evaluation and Statistics.

A total of $N = 18$ faculty and $N = 12$ pre- and post-doctoral students (henceforth students) were recruited to rate the talks at the conference. We sought to recruit faculty and students that were representative of the society in terms of gender, nationality, and area of expertise (Table 5). Travel expenses were covered for the students to attend the conference while faculty volunteered their time. The conference had nine sessions each with three tracks of four talks. Based on availability, faculty were randomly assigned to one of the three tracks in three of the nine sessions during the conference. Students were randomly assigned to rate one track of talks in either the first half or second half of the conference. In the end, we aimed to have two faculty raters and two student raters at each of the 108 talks.

Each faculty member and graduate student rated the talks along the following dimensions:

2.5.1 Characteristics of Raters.

Table 5: Distribution of rater characteristics.

		Pre- and Post-doctoral Students	Faculty	Total
Gender	Female	7 (58%)	7 (39%)	14 (47%)
	Male	5 (42%)	11 (61%)	16 (53%)
Ethnicity	White	9 (75%)	15 (83%)	24 (80%)
	Not White	2 (17%)	2 (11%)	4 (13%)
	No Response	1 (8%)	1 (6%)	2 (7%)
Country of Home Institution (Nationality)	US	7 (58%)	14 (78%)	21 (70%)
	International	4 (33%)	3 (17%)	7 (23%)
	No Response	1 (8%)	1 (5%)	2 (7%)
Area of Specialty	Psychology	2 (17%)	12 (67%)	14 (47%)
	Marketing	5 (42%)	1 (5.5%)	6 (20%)
	Economics	1 (8%)	1 (5.5%)	2 (7%)
	Other	4 (33%)	4 (22%)	8 (26%)

2.5.2 Rated Dimensions.

Each faculty member and graduate student rated the talks along the following dimensions:

Significance Is the topic of the talk significant? Does it concern a scientifically important subject or is it relevant for policy or other applications?

Methods Are the methods scientifically sound?

Results Are results presented in enough detail and in an understandable way?

Conclusion Do the conclusions follow from the results? Are they justified? Are the results generalizable?

Innovation Is there something innovative about the presented material?

Uniqueness Is this talk different from other talks typically at SJDM?

2.5.3 Judged Overall Rating.

For each rater, we standardized the ratings across all dimensions and all talks they rated. Then for each rater, and each talk we averaged across the ratings to form an overall rating. Then for each talk we averaged across the student faculty ratings giving them equal weight.

2.5.4 Collected Talk Statistics.

Each research assistant tracked the following statistics:

Attendance Estimated the attendance at the talk 5 minutes into the talk.

Potential questions Estimated the number of potential questions for each talk, where potential questions were indicated by a count of raised hands throughout the talk.

Number of questions answered The number of questions answered for each talk.

Duration of talk The length of the talk.

Duration of Q & A The length of the question and answer period.

2.6 Poster Ratings.

There were a total of $N = 96$ posters that were submitted for the student-poster competition during the conference and judged on a similar set of dimensions as those used to judge the quality of the talks. A subset of these posters ($N = 57$) had been originally submitted for talks to the conference had single- and double-blind abstract evaluation scores. Thus, we took the poster ratings, standardized them across each judge, averaged them across the dimensions to arrive at an overall rating for each poster, and integrated these ratings with the ratings for the submissions given as a talk, creating a new talk/poster rating.

2.7 Preference for Review process Survey Items.

Shortly before the conference, we surveyed members of the Society for Judgment and Decision Making listserv, asking them to report their preference between single- vs. double-blind review processes and their judged fairness of each review process. We sent the request for participation out on October 25, 2019. We sent a reminder on November 1, 2019, and we stopped data collection on November 7, 2019. There were $N = 173$ responses to the survey.

The survey asked the following questions.

Author characteristics predictions Indicate the extent to which they believed that specific author characteristics (e.g. gender, proportion of male co-authors, institution, etc.) would increase versus decrease the chances a talk would be accepted as a talk at the conference if using single- vs. double-blind review on a 7-point scale ranging from significantly decreases to significant increases.

Unfair vs. fair to use author characteristics Rate [7-point scale] the extent to which they believed it was very unfair to very fair for reviewers to take into account each of the above author characteristics when judging abstracts.

Preference for review process Rate preferences for single- versus double-blind review as an author v. author v. conference attendee v. overall (1 = definitely prefer single-blind, 7 = definitely prefer double-blind).

Fairness of review process Rate each review process according to its fairness (1 = very unfair, 7 = very fair).

Agreement Rate the degree of agreement between the two review processes.

Validity Rate which review process would better predict the aspects of judged talk quality (1 = much better predicted by single-blind review, 7 = much better predicted by double-blind review).

Demographics Demographics were collected.

Knowledge of study Yes/No you had heard about the current blind-review study being conducted within the society.

2.8 Journal Publications.

Two years after the conference, we sought to identify submissions that were published. To do so, we searched Google Scholar using as a search term the first author’s last name and the title of the submission. In total, $N = 130$ submissions were identified as published in a peer-reviewed journal. Then we contacted via email all of the corresponding authors of the remaining submissions and asked if their submission had been published in a peer-reviewed journal and what the journal was. The initial email was sent in October 2020 and we collected emails up until December 18, 2020. We also separately identified the impact factor of the journal via Web of Science and Google Scholar. Responses to the email identified an additional $N = 54$ submissions that had been published. We used the (logged) impact factor from Web of Science as a metric of impact.

2.9 Analyses.

We used Bayesian estimation methods for data analysis. Statistics were conducted using R 3.6.2 (R Core Team, 2019), the rstanarm package (v2.19.3; Goodrich et al., 2020), and the BayesFactor package (0.9.12-4.2; Morey & Rouder, 2018). We report the posterior predicted mean of the parameter of interest and in brackets its 95% Credible Interval (CI). The CI summarizes the posterior distribution excluding the 2.5% of the distribution at each tail. We use the term *credible* when the 95% CI excludes 0. For model comparisons, we used leave-one-out cross validation (LOO) to compare models based on the expected log pointwise predictive density for a new dataset (elpd_{loo}). The difference between models on elpd_{loo} quantifies the difference in predictive accuracy between them. The standard error (SE) captures the uncertainty around each elpd_{loo} . In general, we consider $|\Delta \text{elpd}_{\text{loo}}| < 1 \text{ SE}$ to be weak evidence for a particular model. In all cases, we used the default priors, which were set to be weakly informative providing moderate regularization to the estimates (i.e., priors that are skeptical of extreme parameter values). We inspected MCMC chains for representation and accuracy, and we sought to have all reported parameters based on an effective sample size of 10,000. Posterior distributions do not change based on the number of planned or unplanned tests. Thus, there is no correction for the number of tests as statistics and inferences from them are based on the posterior distribution (Gelman et al., 2012b; Kruschke & Liddell, 2018b).

3 Modeling the Text of the Long Abstracts

3.1 Text Corpus

The text corpus was the *Wiley Blackwell Handbook of Judgment and Decision Making* edited by Keren and Wu published in 2015. This two volume handbook has 35 chapters that provides a comprehensive, examination of the field of judgment and decision making. The handbook has authors from across the discipline including psychology, economics, marketing, finance, public policy, sociology, and philosophy. The review covered traditional topics, controversies, new topics, as well as applications. Therefore, the text provides a good representation of the field. We obtained the text from all 35 chapters and prepared it to model by removing

- Front and back matter
- Titles
- Author names, and cited authors
- Page numbers
- Equations, numbers, and mathematical notation
- Figure and figure caption
- Table and table captions

In the end, we created a .txt file where each line was a chapter. Then we applied the following pre-processing steps in Matlab. See [OSF](#) for further details.

- Converted all the text to lower-case.
- Removed hyphens
- Removed numbers
- Made a second pass to remove authors using the author index as a reference.
- Removed punctuation
- Removed stop words like “a”, “and”, ”to”, and ”the”.
- Removed words two letters or shorter.
- Removed words 15 letters or longer.
- Tokenized the text (via Matlab’s ‘tokenizedDocument’ function.
- Removed words that occurred 3 or fewer times across the text.

The end text corpus had 35 chapters with 5020 words. Figure 2 shows a word cloud of the frequent words and it suggests that the handbook, indeed, contains the relevant words one would expect in a handbook about how people make judgments and decisions.

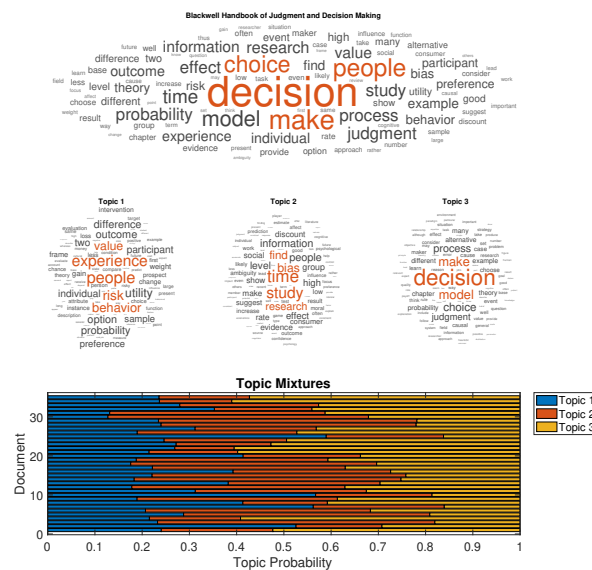


Figure 2: Topic modeling of the 2015 *Wiley Blackwell Handbook of Judgment and Decision Making*. The top plot is a word cloud of the most frequent words in the handbook. The middle plots show the most frequent words associated with the three topics from the LDA topic model. The bottom plot shows the distribution of topics across the 35 chapters.

3.2 Topics Modeling

We fit a latent Dirichlet allocation (LDA) topic model to the *Wiley Handbook of Judgment and Decision Making* text corpus using Gibbs sampling (Blei et al., 2003; Griffiths & Steyvers, 2004). The LDA model discovers underlying topics in a collection of documents and infers word probabilities in topics. It was fit using a bag-of-n-grams model treating the n-grams as individual words.

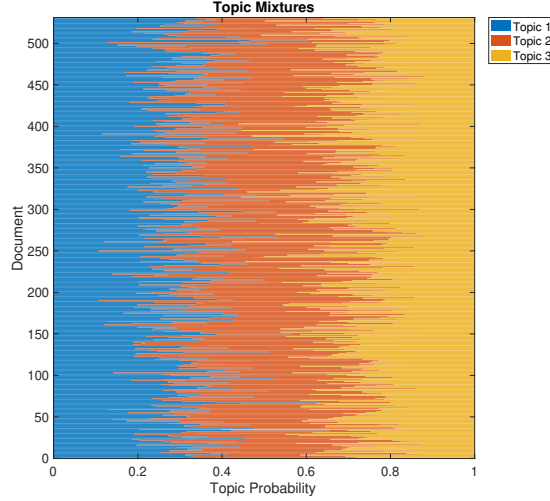


Figure 3: Topic distribution across the 530 long abstracts. The topic model came from the best fitting LDA topic model of the 2015 *Wiley Blackwell Handbook of Judgment and Decision Making* (see Figure 2).

To identify the number of topics that best fit the handbook, we carried out a cross validation holding out 7 of the chapters and fitting a topics model with 1 to 15 topics to the remaining 28 chapters. Then we calculated a perplexity score—the geometric mean of the inverse marginal probability of each word in the held-out set of documents—as a measure of goodness of fit. We repeated this process 10 times using a different partition of chapters each time. Across iterations, the median number of topics that returned a minimum perplexity score was 2.5. Therefore we selected 3 topics for the best fitting model. Figure 2 shows the word clouds of the most frequent words associated with each topic and the topic mixtures of the different chapters. Roughly the three topics were

1. Theories of judgment and decision making
2. Empirical phenomena of judgment and decision making
3. Psychological processes of judgment and decision making

These labels correspond well to the chapters. For instance, for Chapter 2 Topic 1 had the greatest probability. It focused on the theory of how people make decisions under risk, a core topic of judgment and decision making. For Chapter 6 Topic 2 had the greatest probability and it focused on the empirical phenomena of overconfidence in judgment. And for Chapter 22 Topic 3 had the greatest probability and it focused on learning processes during decision making.

We then used the model to calculate the posterior probability that each long abstract belonged to each of these three topics and assigned each submission the topic with the greatest posterior probability. Figure 3 shows the distribution of topics across the 530 long abstracts. The code and output for the topic modeling is available on [OSF](#).

3.3 Sentiment Analysis

The sentiment analysis assigned a score to each long abstract according to whether it had on average a positive or negative sentiment. The score was based on a sentiment analysis conducted by training a support vector machine classifier on a subset of 6,800 positive and negative words from Hu & Liu (2004). Then after validating the classifier on the hold-out set, we predicted the sentiment of the words in the long abstract with the classifier. Note both the positive and negative words and the abstract words were represented by 300 dimensional numerical vectors via a pre-trained word embedding of 1 million English words available in Matlab. Figure 4 shows the word clouds of the words with the strongest negative and positive sentiment

4.1 Characteristics of Respondents

Table 6: Characteristics of the respondents to the survey of the Society for Judgment and Decision Making on single- vs. double-blind review.

	Group	Count
Gender	Female	73 (42%)
	Male	85 (49%)
	Other	1 (1%)
	No Response	14 (8%)
Area of Ph.D.	Cognitive Psychology	36 (21%)
	Social Psychology	26 (15%)
	Other Psychology	10 (6%)
	Marketing	31 (18%)
	Organizational Behavior	9 (5%)
	Management	6 (3%)
	Economics	1 (1%)
	Accounting	0 (0%)
	Finance	1 (1%)
	Public Policy	2 (1%)
	Law	0 (0%)
	Medicine	0 (0%)
	Decision Science	24 (14%)
	Other	10 (6%)
	No Response	17 (10%)
Department Area	Cognitive Psychology	21 (12%)
	Social Psychology	11 (6%)
	Other Psychology	8 (5%)
	Marketing	48 (28%)
	Organizational Behavior	11 (6%)
	Management	13 (8%)
	Economics	3 (2%)
	Accounting	0 (0%)
	Finance	0 (0%)
	Public policy	4 (2%)
	Law	0 (0%)
	Medicine	2 (1%)
	Decision Science	15 (9%)
	Other	19 (11%)
	No Response	18 (10%)
Position	Undergraduate	1 (1%)
	Masters student / graduate	0 (0%)
	PhD student / graduate	35 (20%)
	Postdoctoral Researcher	18 (10%)
	Research Scientist	4 (2%)
	Assistant Professor	49 (28%)
	Associate Professor	24 (14%)
	Full Professor	25 (14%)
	Other	3 (2%)
	No Response	14 (8%)
Member of SJDM	Yes	154 (89%)
	No	8 (5%)
	No Response	11 (6%)

4.2 Preference Ratings

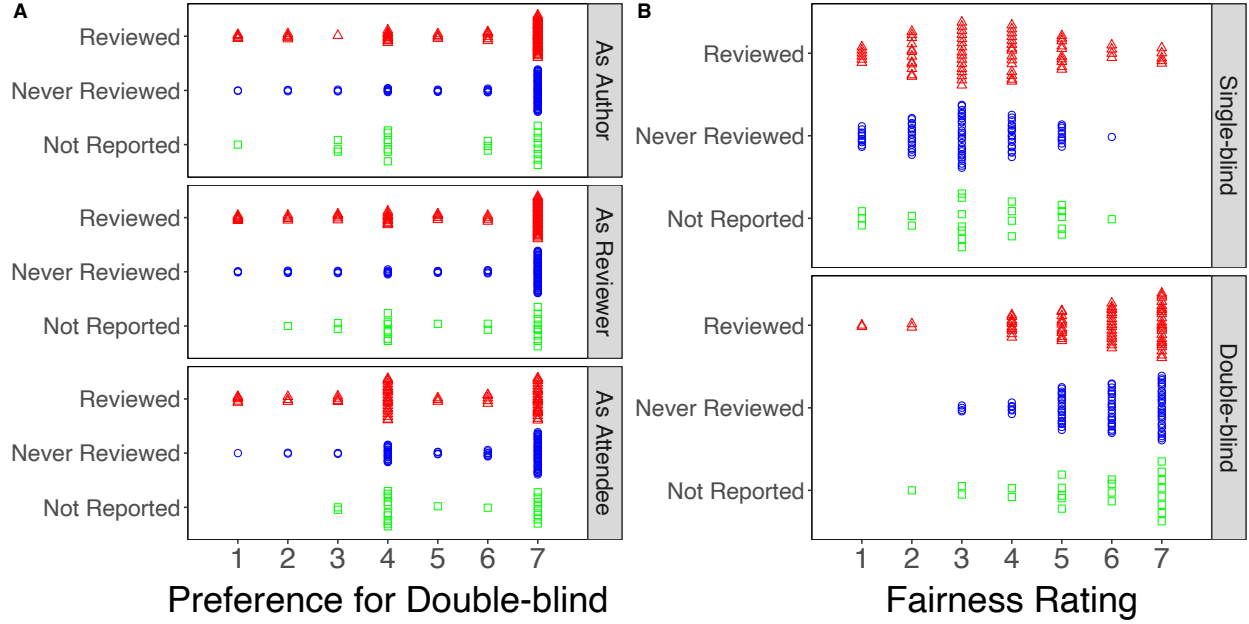


Figure 5: **A:** Society for Judgment and Decision Making members’ surveyed preference for single-blind versus double-blind review for selecting submissions for the annual conference from the perspective of different roles (1 = Strong preference for single-blind, 7 = Strong preference for double-blind). **B:** Judged fairness of the two review processes (1 = Very unfair, 7 = Very fair). Reviewing experience was analyzed as a continuous variable. For the figure, we split the responses into whether respondents had never reviewed (0 years of reviewing), reviewed (> 0 years of reviewing), or did not report their reviewing experience.

4.2.1 Regression Coefficients for Predictors of Double- vs. Single-blind Preference.

Table 7: Coefficients from regressing perspective, reviewing experience, and fairness on double-blind vs. single-blind preference.

	w/o Fairness		w Fairness	
	M	95% CI	M	95% CI
Intercept	0.139	[-0.025, 0.298]	0.125	[-0.011, 0.261]
Reviewer vs. Author	-0.112	[-0.24, 0.012]	-0.113	[-0.242, 0.017]
Attendee vs. Author	-0.225	[-0.353, -0.1]	-0.229	[-0.36, -0.098]
Reviewing Experience (z)	-0.206	[-0.364, -0.049]	-0.091	[-0.231, 0.047]
Reviewer vs. Author x Reviewing Experience	-0.042	[-0.17, 0.082]	-0.043	[-0.165, 0.084]
Attendee vs. Author x Reviewing Experience	-0.008	[-0.135, 0.119]	-0.008	[-0.135, 0.117]
Fairness of Double- vs. Single-Blind (z)	-	-	0.538	[0.422, 0.652]

Bolded values indicated credible interval excludes 0. Preference and reviewing experience were standardized in the model. Respondent id was a random intercept in the regression model.

4.3 Fairness Perceptions

Table 8: Coefficients from regressing perspective and reviewing experience onto fairness.

	M	95% CI
Double (1) or Single (0)	1.31	[1.14, 1.48]
Reviewing Experience (z)	0.20	[0.08, 0.32]
Double or Single x Reviewer Experience (z)	-0.24	[-0.41, -0.07]

Bolded values indicated credible interval excludes 0. Fairness ratings and reviewing experience were standardized in the model. Respondent id was a random intercept in the regression model.

Figure 5 also shows the distribution of fairness ratings. Double-blind review was rated as more fair (Table 8). Moreover, reviewing experience was also associated with the difference in fairness between single- and double-blind with more experienced reviewers rating seeing less of a difference in fairness (Tables 7). This pattern raises the question whether differences in fairness perceptions accounts for the relationship between review experience and reviewer preference for double-blind review? When fairness is included in the regression predicting the preference, the relationship between reviewing experience and review-system preference dissipates once fairness perceptions are accounted for ($\beta = -0.09$, $[-0.23, 0.05]$) (Table 8). This result suggests that part of the reason for the differences in preference for double-blind and single-blind across levels of reviewing experience is the different perceptions of fairness.

5 Reliability of Review Systems

5.1 Comparing Bottom vs. Top Half of Submission

As reported in the main paper, median split by the single-blind rating reveals more agreement between single- and double blind review for the bottom half of the submissions ($r = 0.45$, $[0.35, 0.54]$) compared to the top half ($r = 0.19$, $[0.07, 0.29]$). To establish if this degree of agreement was different between the bottom vs. top half, we regressed the double-blind rating onto the single-blind rating, a dummy variable for bottom- vs. top-half, and the interaction. The interaction was credible (Table 9) indicating a credible difference between bottom and top-half in terms of agreement.

Table 9: Regressing double-blind review rating onto single-blind review rating, whether the submission was in the bottom vs. top half of the median split, and the interaction.

	M	95% CI
Intercept	0.098	[-0.013, 0.209]
Single-blind Review	0.686	[0.534, 0.84]
Bottom (0) vs. Top (1) Half	0.035	[-0.136, 0.206]
Single-blind x Bottom vs. Top	-0.400	[-0.645, -0.162]
sigma	0.578	[0.545, 0.614]

Bolded values indicated credible intervals exclude 0.

6 Differential Use Between Single- and Double-blind Review

6.1 Model Comparison

Our first step was to assess the overall evidence for submission characteristics like the author’s identity being used differently between single- and double-blind using a model comparison. We compared a model that had interaction terms between each of the eleven characteristics and the condition (*Differential Use (DU) Model*

) to a model that did not have these interaction terms (*No Differential Use (NoDU) Model*) using leave-one-out cross validation. The difference between the DUU ($elpd_{loo} = -3325.206$; $SE(elpd_{loo}) = 32.918$) and the simpler NoDU model ($elpd_{loo} = -3318.547$; $SE(elpd_{loo}) = 32.532$) in terms of the expected log-pointwise predictive density was $\Delta elpd_{loo} = 6.7$ favoring the simpler NoDU model without the interaction terms, but this difference was within the margin of error of $SE = 6.9$ indicating very weak evidence against an omnibus differential use of submission characteristics during single-blind review.

Our main analysis entered all thirteen characteristics (Table 1) simultaneously in a multilevel regression model with the ratings for the single- and double-blind conditions as the outcome variables. Though reviewers were randomly assigned to conditions and submission, we also included the six reviewer characteristics listed in Table 1 as predictors. In the analyses reported in the main paper, the reviewer characteristics were entered alone. We also explored whether there were potential interactions between the review condition and reviewer characteristics. None of the reviewer characteristics showed credible interactions with the review condition (Table 13). Moreover, in a model comparison between the Differential Use Model (DUM) and the model including an interaction between reviewer characteristics and condition as well as potential three-way interactions with author specific author characteristics led to a substantially worse fit of the regression model (Table 10).

Table 10: Model comparison between models with interactions between reviewer characteristics and review condition.

	$\Delta elpd$	$se(\Delta elpd)$	$elpd_{loo}$	$se(elpd_{loo})$
Differential Use Model	0	0	-3325.206	32.918
Differential Use Model + Reviewer x Condition	-7.452	2.299	-3332.658	32.969
Differential Use Model + Reviewer x Condition + Reviewer x Submission x Condition	-35.993	11.050	-3361.199	33.384

The comparisons were made with the expected log pointwise predictive density for a new dataset estimated from the leave-one-out cross validation ($elpd_{loo}$) for each model. The difference between two models with this statistic ($\Delta elpd_{loo}$) quantifies the difference in predictive accuracy between the two and offers a measure of the relative support for either model. There is uncertainty around each $elpd_{loo}$ and this is captured by the standard error (SE) around each estimate.

6.2 Regression Coefficients With All Characteristics Entered Simultaneously

6.2.1 All Submissions.

Table 11: Coefficients from regressing single- and double-blind ratings onto submission characteristics.

Category	Variable	M [95% CI]		ELPD
		Predictor	Predictor x Condition	Interaction - DUM
	Intercept	-0.025, [-0.501, 0.451]		
	Condition; Single (0) or Double (1)	0.221, [-0.254, 0.692]		
First Author	<i>Male (1) or Female (0)</i>	0.134, [-0.054, 0.324]	0.255, [0.066, 0.445]	-0.110 (1.738)
	Years Since Ph.D.	-0.134, [-0.269, 0.001]	-0.023, [-0.157, 0.113]	3.056 (2.945)
	Domestic (0) or International (1)	0.039, [-0.109, 0.185]	-0.048, [-0.198, 0.101]	-0.701 (1.152)
	<i>Asian vs White</i>	-0.054, [-0.225, 0.115]	0.18, [0.01, 0.351]	-5.223 (1.823)
	Black vs White	-0.232, [-1.184, 0.721]	-0.299, [-1.278, 0.687]	"
	Hispanic or Latino vs White	0.159, [-0.246, 0.567]	-0.096, [-0.508, 0.321]	"
	Native American vs White	0.106, [-1.233, 1.439]	-0.042, [-1.412, 1.322]	"
	Other vs White	0.424, [-0.02, 0.869]	-0.181, [-0.637, 0.273]	"
	No Answer vs White	0.068, [-0.171, 0.311]	-0.07, [-0.315, 0.176]	"
	Ave. Citations of 1st Author's Institute	0.025, [-0.125, 0.174]	0.007, [-0.147, 0.162]	0.801 (2.230)
	Economics Ph.D. vs. Psychology	0.087, [-0.206, 0.38]	0.038, [-0.258, 0.333]	-2.155 (2.195)
	Management Ph.D. vs Psychology	0.078, [-0.167, 0.323]	0.079, [-0.17, 0.334]	"
	Marketing Ph.D. vs. Psychology	0.033, [-0.156, 0.224]	-0.033, [-0.224, 0.16]	"
	Other Ph.D. vs. Psychology	0.09, [-0.14, 0.319]	-0.097, [-0.329, 0.137]	"
	Undergraduate vs. Full	0.115, [-1.323, 1.557]	-1.28, [-2.765, 0.227]	-0.819 (4.038)
	Masters vs. Full	-0.145, [-0.848, 0.552]	-0.214, [-0.937, 0.504]	"
	<i>Ph.D. vs. Full</i>	-0.016, [-0.529, 0.489]	-0.552, [-1.074, -0.026]	"
	Post Doc vs. Full	-0.077, [-0.552, 0.4]	-0.386, [-0.878, 0.107]	"
	Research Scientist vs. Full	0.035, [-0.592, 0.658]	-0.548, [-1.182, 0.093]	"
	Practitioner vs. Full	-0.459, [-1.218, 0.302]	-0.297, [-1.036, 0.446]	"
	Assistant vs. Full	0.073, [-0.366, 0.508]	-0.358, [-0.809, 0.096]	"
	Associate vs. Full	-0.098, [-0.504, 0.31]	-0.208, [-0.629, 0.214]	"
Co-Author	Proportion Male Authors	0.016, [-0.078, 0.112]	-0.103, [-0.198, -0.008]	-0.580 (0.979)
	Ave. Citations of Authors' Institutes	0.137, [-0.013, 0.286]	-0.054, [-0.207, 0.099]	1.939 (2.672)
	Ave. Seniority of Authors	0.132, [0.053, 0.213]	-0.07, [-0.151, 0.011]	-0.861 (0.934)
	Empirical Studies vs. Traditional Topic	-0.019, [-0.179, 0.142]	0.073, [-0.088, 0.234]	-1.861 (1.350)
	Psychology of JDM vs. Traditional Topic Sentiment	-0.049, [-0.259, 0.161]	0.069, [-0.145, 0.284]	"
Reviewer	Reviewer Female (0) or Male (1)	-0.026, [-0.105, 0.053]		
	Reviewer Years Since Ph.D.	-0.022, [-0.081, 0.037]		
	Reviewer Ave. Citations of 1st Author's Institute	0.027, [-0.011, 0.065]		
	Economics Ph.D. vs. Psychology	0.029, [-0.096, 0.155]		
	Management Ph.D. vs Psychology	-0.002, [-0.119, 0.114]		
	Marketing Ph.D. vs. Psychology	-0.035, [-0.135, 0.066]		
	Other Ph.D. vs. Psychology	0.025, [-0.131, 0.179]		
	Research Scientist vs. Full	-0.141, [-0.334, 0.052]		
	Practitioner vs. Full	-0.165, [-0.571, 0.246]		
	Assistant vs. Full	-0.035, [-0.179, 0.108]		
	Associate vs. Full	-0.01, [-0.129, 0.108]		
	Reviewer # Times Review Past 7 Years	-0.013, [-0.053, 0.027]		

Bolded values indicate credible intervals exclude 0. All coauthor characteristic characteristics were represented by the statistics based on all the authors (e.g., proportion of male authors instead of male coauthors) so that single-author submissions were included in the analysis. Thus, the regression coefficients for the first author variables represent incremental change for the variables in which there are corresponding coauthor characteristics. Review ratings and numerical predictors were standardized as indicated by the z. The values in the column under Δ ELPD₁₀₀ are the differences in the expected log pointwise predictive density for a new dataset between the NoDU model including the interaction between the predictor and condition and the NoDU model.

6.2.2 Multiauthor Submissions.

Table 12: Coefficients from regressing single- and double-blind ratings onto submission characteristics for multi-authored submissions.

		M [95% CI]	ELPD	
	Characteristic	Predictor	Predictor x Condition	Interaction - DUM
First Author	Intercept	0.08 [-0.394, 0.549]		
	Condition; Single (0) or Double (1)	0.15 [-0.319, 0.627]		
	Female (0) or Male (1)	0.172 [0.031, 0.315]	0.143 [-0.003, 0.289]	0.042 (1.719)
	Years Since Ph.D.	-0.134 [-0.281, 0.015]	-0.033 [-0.18, 0.113]	2.200 (2.749)
	Domestic (0) or International (1)	0.072 [-0.08, 0.22]	-0.054 [-0.208, 0.103]	-0.693 (1.390)
	<i>Asian vs White</i>	-0.062 [-0.234, 0.11]	<i>0.202 [0.022, 0.382]</i>	<i>-4.778 (2.078)</i>
	Black vs White	-0.222 [-1.152, 0.707]	-0.318 [-1.296, 0.667]	"
	Hispanic or Latino vs White	0.149 [-0.25, 0.551]	-0.128 [-0.545, 0.285]	"
	Native American vs White	0.088 [-1.204, 1.378]	-0.039 [-1.409, 1.317]	"
	Other vs White	0.492 [0.033, 0.948]	-0.164 [-0.645, 0.317]	"
	No Answer vs White	0.102 [-0.144, 0.345]	-0.093 [-0.347, 0.16]	"
	Ave. Citations of 1st Author's Institute	0.063 [-0.036, 0.163]	-0.011 [-0.113, 0.091]	1.304 (2.396)
	Economics Ph.D. vs. Psychology	0.175 [-0.127, 0.474]	0.012 [-0.296, 0.322]	-0.574 (2.823)
	Management Ph.D. vs. Psychology	0.153 [-0.1, 0.402]	0.045 [-0.217, 0.306]	"
	Marketing Ph.D. vs. Psychology	0.049 [-0.143, 0.242]	-0.013 [-0.213, 0.187]	"
	Other Ph.D. vs. Psychology	0.155 [-0.079, 0.386]	-0.129 [-0.369, 0.111]	"
	Undergraduate vs. Full	-0.355 [-1.752, 1.049]	-1.073 [-2.542, 0.401]	0.107 (3.827)
	Masters vs. Full	-0.425 [-1.106, 0.263]	-0.069 [-0.790, 0.652]	"
	Ph.D. vs. Full	-0.26 [-0.765, 0.25]	-0.399 [-0.925, 0.121]	"
	Post Doc vs. Full	-0.308 [-0.791, 0.176]	-0.252 [-0.757, 0.248]	"
	Research Scientist vs. Full	-0.185 [-0.804, 0.439]	-0.408 [-1.062, 0.239]	"
	Practitioner vs. Full	-0.623 [-1.382, 0.13]	-0.22 [-0.975, 0.523]	"
Assistant vs. Full	-0.052 [-0.507, 0.408]	-0.294 [-0.768, 0.176]	"	
Associate vs. Full	-0.137 [-0.557, 0.282]	-0.216 [-0.655, 0.218]	"	
Co-Author	Proportion Male Couthors	0.032 [-0.039, 0.103]	-0.073 [-0.147, 0.000]	2.018 (2.619)
	Ave. Citations of Coauthors' Institutes	0.098 [0.001, 0.195]	-0.043 [-0.144, 0.057]	3.046 (2.955)
	<i>Ave. Seniority of Coauthors</i>	0.103 [0.026, 0.18]	<i>-0.09 [-0.168, -0.011]</i>	<i>7.820 (4.270)</i>
	Empirical Studies vs. Traditional Topic	-0.024 [-0.187, 0.137]	0.114 [-0.051, 0.276]	-1.482 (1.732)
	Psychology of JDM vs. Traditional Topic	-0.043 [-0.257, 0.172]	0.123 [-0.100, 0.342]	"
	Sentiment	-0.07 [-0.14, 0]	0.046 [-0.027, 0.119]	-1.201 (0.629)
Reviewer	Reviewer Female (0) or Male (1)	-0.03 [-0.113, 0.051]		
	Reviewer Years Since Ph.D.	-0.018 [-0.079, 0.043]		
	Reviewer Ave. Citations of 1st Author's Institute	0.034 [-0.005, 0.073]		
	Economics Ph.D. vs. Psychology	0.029 [-0.101, 0.157]		
	Management Ph.D. vs. Psychology	-0.019 [-0.139, 0.1]		
	Marketing Ph.D. vs. Psychology	-0.033 [-0.138, 0.071]		
	Other Ph.D. vs. Psychology	0.05 [-0.111, 0.21]		
	Research Scientist vs. Full	-0.109 [-0.315, 0.094]		
	Practitioner vs. Full	-0.151 [-0.579, 0.279]		
	Assistant vs. Full	0 [-0.148, 0.147]		
	Associate vs. Full	0.007 [-0.115, 0.131]		
	Reviewer # Times Review Past 7 Years	-0.011 [-0.051, 0.031]		

Bolded values indicated credible intervals exclude 0. Review ratings and numerical predictors were standardized. The data were subsetted to only include multi-author submissions. ELPD stands for the expected log pointwise predictive density for a new dataset calculated with leave one out cross validation (loo) indicating the predictive accuracy of the respective model. The values in the column under $\Delta \text{ELPD}_{\text{loo}}$ indicate the difference between the NoDU model for multiauthor submissions including the interaction between the predictor and condition and the NoDU model for multiauthor submissions.

6.3 Regression Coefficients With Each Variable Entered Alone

Table 13: Coefficients from regression with each variable entered alone.

Category	Variable	Single Blind Coef (Alone)		Double Blind Coef (Alone)	
		M	95% CI	M	95% CI
First Author	Female (0) or Male (1)	0.094	[-0.046, 0.230]	0.163	[0.035, 0.286]
	Years Since Ph.D.*	-0.102	[-0.172, -0.032]	-0.03	[-0.099, 0.037]
	International (0) or Domestic (1)	0.172	[0.034, 0.307]	0.151	[0.023, 0.276]
	Asian vs. White	-0.035	[-0.202, 0.128]	0.086	[-0.073, 0.246]
	Black vs. White	0.161	[-0.850, 1.198]	-0.144	[-1.115, 0.865]
	Hispanic or Latino vs. White	0.332	[-0.098, 0.769]	0.240	[-0.167, 0.645]
	Native American vs. White	0.100	[-1.373, 1.630]	0.023	[-1.422, 1.447]
	Other vs. White	0.245	[-0.221, 0.725]	0.156	[-0.298, 0.617]
	No Answer vs. White	-0.016	[-0.254, 0.212]	-0.057	[-0.274, 0.164]
	Ave. Citations of 1st Author's Institute	0.156	[0.090, 0.224]	0.095	[0.032, 0.160]
	Economics Ph.D. vs. Psychology	0.092	[-0.174, 0.363]	0.157	[-0.113, 0.416]
	Management Ph.D. vs. Psychology	0.245	[0.022, 0.474]	0.300	[0.083, 0.513]
	Marketing Ph.D. vs. Psychology	0.222	[0.050, 0.389]	0.191	[0.031, 0.353]
	Other Ph.D. vs. Psychology	0.360	[0.137, 0.580]	0.220	[0.015, 0.429]
	Undergraduate vs. Full	0.232	[-1.056, 1.543]	-0.401	[-1.606, 0.904]
	Masters vs. Full*	-0.206	[-0.722, 0.370]	-0.196	[-0.694, 0.378]
	Ph.D. vs. Full*	0.342	[0.072, 0.592]	-0.007	[-0.264, 0.254]
	Post Doc vs. Full	0.267	[-0.046, 0.552]	0.073	[-0.221, 0.375]
	Research Scientist vs. Full	0.340	[-0.085, 0.757]	-0.089	[-0.509, 0.321]
	Practitioner vs. Full	-0.031	[-0.683, 0.520]	-0.172	[-0.788, 0.404]
	Assistant vs. Full	0.372	[0.131, 0.645]	0.168	[-0.097, 0.426]
	Associate vs. Full	0.053	[-0.274, 0.382]	-0.038	[-0.356, 0.280]
Authors & Co-Authors	Proportion Male Coauthors*	0.041	[-0.027, 0.110]	-0.034	[-0.100, 0.031]
	Proportion Male Authors	0.051	[-0.018, 0.118]	0.018	[-0.047, 0.84]
	Ave. Citations of Co-Authors' Institutes*	0.196	[0.129, 0.262]	0.118	[0.053, 0.183]
	Ave. Citations of Authors' Institutes*	0.192	[0.128, 0.254]	0.125	[0.060, 0.188]
	Ave. Seniority of Coauthors*	0.157	[0.087, 0.227]	0.005	[-0.061, 0.077]
	Ave. Seniority of Authors	0.112	[0.047, 0.179]	0.063	[0.001, 0.126]
Text	Empirical Studies vs. Traditional Topic	0.013	[-0.146, 0.168]	0.045	[-0.105, 0.199]
	Psychology of JDM vs. Traditional Topic	-0.123	[-0.322, 0.079]	-0.092	[-0.286, 0.108]
	Sentiment	-0.012	[-0.079, 0.055]	0.017	[-0.044, 0.081]
Reviewer	Female (0) or Male (1)	-0.009	[-0.105, 0.086]	-0.010	[-0.109, 0.091]
	Years Since Ph.D.	-0.001	[-0.008, 0.006]	-0.003	[-0.011, 0.004]
	Ave. Citation of Institution	0.007	[-0.036, 0.051]	0.017	[-0.036, 0.069]
	Economics Ph.D. vs. Psychology	0.041	[-0.113, 0.201]	0.028	[-0.131, 0.183]
	Management Ph.D. vs. Psychology	-0.022	[-0.187, 0.135]	0.040	[-0.117, 0.197]
	Marketing Ph.D. vs. Psychology	0.002	[-0.125, 0.130]	-0.008	[-0.134, 0.119]
	Other Ph.D. vs. Psychology	-	-	0.022	[-0.133, 0.172]
	Research Scientist vs. Full	-0.077	[-0.302, 0.146]	-0.078	[-0.337, 0.176]
	Practitioner vs. Full	-	-	-0.037	[-0.3000, 0.221]
	Assistant vs. Full	0.051	[-0.078, 0.179]	-0.031	[-0.146, 0.089]
	Associate vs. Full	0.028	[-0.097, 0.160]	-0.042	[-0.188, 0.099]
	Experience reviewing past 7 years	-0.012	[-0.044, 0.018]	0.003	[-0.024, 0.030]

Bolded values indicated credible interval excludes 0. Starred variables (*) indicate a credible interaction between the variable and the condition in Table 15. Review ratings and numerical predictors were standardized. The coefficients in the single- and double-blind were entered alone.

Table 14: Coefficients from regression with each variable entered alone along with the review condition.

Category	Variable	M [95% CI]				Δ ELPD _{loo} (SE)
		Intercept	Condition (Coef)	Predictor Coef	Single x Double Coef	
First Author	Female (0) or Male (1)	-0.065 [-0.160, 0.031]	-0.036 [-0.128, 0.053]	0.092 [-0.039, 0.223]	0.081 [-0.041, 0.204]	0.208 (1.395)
	Years Since Ph.D.	-0.029 [-0.097, 0.038]	0.009 [-0.052, 0.071]	-0.103 [-0.172, -0.035]	0.073 [0.012, 0.134]	2.014 (2.498)
	International (0) or Domestic (1)	-0.097 [-0.185, -0.009]	0.016 [-0.067, 0.099]	0.173 [0.043, 0.303]	-0.023 [-0.146, 0.101]	-1.240 (0.489)
	Asian vs White	-0.026 [-0.108, 0.057]	-0.010 [-0.087, 0.069]	-0.032 [-0.193, 0.125]	0.115 [-0.032, 0.257]	-4.140 (2.229)
	Black vs White	"	"	0.163 [-0.838, 1.184]	-0.317 [-1.278, 0.668]	"
	Hispanic or Latino vs White	"	"	0.319 [-0.094, 0.739]	-0.089 [-0.481, 0.304]	"
	Native American vs White	"	"	0.104 [-1.332, 1.504]	-0.084 [-1.403, 1.291]	"
	Other vs White	"	"	0.252 [-0.206, 0.716]	-0.102 [-0.537, 0.331]	"
	No Answer vs White	"	"	-0.009 [-0.230, 0.212]	-0.063 [-0.268, 0.141]	"
	Ave. Citations of 1st Author's Institute	0.044 [-0.021, 0.109]	-0.003 [-0.067, 0.060]	0.154 [0.090, 0.220]	-0.059 [-0.124, 0.007]	0.553 (1.935)
	Economics Ph.D. vs. Psychology	-0.14 [-0.266, -0.051]	0.020 [-0.065, 0.106]	0.108 [-0.155, 0.370]	0.034 [-0.208, 0.276]	-2.186 (1.770)
	Management Ph.D. vs Psychology	"	"	0.243 [0.023, 0.464]	0.06 [-0.150, 0.268]	"
	Marketing Ph.D. vs. Psychology	"	"	0.219 [0.055, 0.385]	-0.033 [-0.187, 0.122]	"
	Other Ph.D. vs. Psychology	"	"	0.352 [0.141, 0.561]	-0.138 [-0.336, 0.064]	"
	Undergraduate vs. Full	-0.287 [-0.523, -0.052]	0.238 [0.029, 0.449]	0.19 [-1.159, 1.569]	-0.628 [-1.953, 0.689]	0.628 (3.723)
	Masters vs. Full	"	"	-0.177 [-0.714, 0.359]	-0.017 [-0.508, -0.472]	"
	Ph.D. vs. Full*	"	"	0.332 [0.078, 0.585]	-0.333 [-0.566, -0.104]	"
	Post Doc vs. Full	"	"	0.267 [-0.033, 0.565]	-0.205 [-0.479, 0.062]	"
	Research Scientist vs. Full	"	"	0.346 [-0.073, 0.754]	-0.442 [-0.820, 0.063]	"
	Practitioner vs. Full	"	"	-0.032 [-0.631, 0.555]	-0.148 [-0.669, 0.374]	"
	Assistant vs. Full	"	"	0.369 [0.100, 0.642]	-0.205 [-0.449, 0.033]	"
	Associate vs. Full	"	"	0.057 [-0.266, 0.388]	-0.097 [-0.390, 0.193]	"
Authors & Co-Authors	Proportion Male Coauthors	0.005 [-0.061, 0.072]	0.002 [-0.061, 0.063]	0.042 [-0.026, 0.109]	-0.075 [-0.139, -0.012]	1.809 (2.458)
	Proportion Male Authors	-0.021 [-0.054, 0.066]	0.006 [-0.54, 0.066]	0.051 [-0.014, 0.166]	-0.03 [-0.090, 0.031]	-0.939 (1.048)
	Ave. Citations of Co-Authors' Institutes	0.044 [-0.022, 0.108]	-0.013 [-0.078, 0.051]	0.195 [0.129, 0.259]	-0.076 [-0.141, -0.011]	2.007 (2.550)
	Ave. Citations of Authors' Institutes	0.023 [-0.041, 0.088]	-0.008 [-0.069, 0.053]	0.191 [0.128, 0.255]	-0.067 [-0.127, -0.007]	0.658 (2.315)
	Ave. Seniority of Coauthors	-0.007 [-0.072, 0.060]	0.018 [-0.046, 0.081]	0.151 [0.0.084, 0.219]	-0.141 [-0.204, -0.078]	9.632 (4.555)
	Ave. Seniority of Authors	-0.020 [-0.084, 0.045]	0.006 [-0.054, 0.066]	0.11 [0.047, 0.173]	-0.047 [-0.105, 0.012]	0.263 (1.662)
Text	Empirical Studies vs. Traditional Topic	-0.010 [-0.141, 0.119]	-0.017 [-0.135, 0.100]	0.021 [-0.136, 0.179]	0.030 [-0.114, 0.175]	-1.939 (0.543)
	Psychology of JDM vs. Traditional Topic	"	"	-0.114 [-0.316, 0.087]	0.032 [-0.152, 0.220]	"
	Sentiment	-0.020 [-0.085, 0.045]	-0.006 [-0.054, 0.067]	0.011 [-0.077, 0.054]	0.026 [-0.033, 0.086]	0.051 (0.957)
Reviewer	Female (0) or Male (1)	-0.008 [-0.088, 0.070]	-0.012 [-0.111, 0.086]	-0.024 [-0.114, 0.068]	0.034 [-0.100, 0.167]	-0.577 (0.747)
	Years Since Ph.D.	-0.014 [-0.109, 0.084]	0.033 [-0.084, 0.146]	-0.001 [-0.008, 0.007]	-0.003 [-0.013, 0.008]	-0.809 (0.776)
	Ave. Citatiton of Institution	-0.020 [-0.086, 0.044]	0.004 [-0.055, 0.063]	0.015 [-0.028, 0.056]	0.008 [-0.057, 0.073]	0.314 (0.608)
	Economics Ph.D. vs. Psychology	-0.027 [-0.105, 0.050]	0.013 [-0.081, 0.110]	0.074 [-0.74, 0.220]	-0.063 [-0.279, 0.150]	-3.549 (1.026)
	Management Ph.D. vs Psychology	"	"	-0.016 [-0.167, 0.134]	0.055 [-0.161, 0.271]	"
	Marketing Ph.D. vs. Psychology	"	"	-0.001 [-0.123, 0.119]	-0.018 [-0.190, 0.157]	"
	Other Ph.D. vs. Psychology	"	"	-0.004 [-0.144, 0.137]	-	"
	Research Scientist vs. Full	-0.043 [-0.146, 0.062]	0.070 [-0.065, 0.200]	-0.052 [-0.260, 0.162]	-0.082 [-0.404, 0.237]	-0.745 (1.384)
	Practitioner vs. Full	"	"	-0.073 [-0.332, 0.179]	-	"
	Assistant vs. Full	"	"	0.035 [-0.086, 0.158]	-0.073 [-0.237, 0.092]	"
	Associate vs. Full	"	"	0.038 [-0.086, 0.162]	-0.114 [-0.297, 0.079]	"
	Experience reviewing past 7 years	-0.006 [-0.083, 0.067]	-0.012 [-0.094, 0.071]	-0.009 [-0.037, 0.019]	0.013 [-0.025, 0.050]	-0.595 (0.861)

Bolded values indicated credible interval excludes 0. Review ratings and numerical predictors were standardized. The interaction term was run entering the variable predicting both single- and double-blind ratings together with condition dummy coded (0: single-blind; 1: double-blind). ELPD stands for the expected log pointwise predictive density for a new dataset calculated with leave one out cross validation (loo) indicating the predictive accuracy of the respective model (Vehtari et al., 2017). The values in the column under Δ ELPD_{loo} indicate the difference between model where the interaction between the predictor and condition was in and the model without the interaction and only the predictor. Positive values of Δ ELPD_{loo} indicate more predictive accuracy for the model with the interaction indicating differential use of that predictor between conditions. Negative values indicate more predictive accuracy for the model without the interaction term. The standard error is the variability associated the difference and indexes the uncertainty around the difference in predictive accuracy of the model.

6.4 Regression Coefficients with All Variables Simultaneously

Table 15: Coefficients from regressing submission characteristics onto single- or double-blind review ratings separately.

Category	Predictor	Single	Double
	Intercept	-0.134 [-0.64, 0.366]	0.249 [-0.256, 0.757]
First Author	Female (0) or Male (1)	0.146 [-0.044, 0.336]	0.372 [0.18, 0.562]
	Proportion Male Authors	0.014 [-0.081, 0.109]	-0.082 [-0.177, 0.015]
	Years Since Ph.D.	-0.124 [-0.259, 0.012]	-0.156 [-0.295, -0.017]
	Domestic (0) or International (1)	0.033 [-0.113, 0.177]	-0.008 [-0.155, 0.142]
	Asian vs White	-0.059 [-0.229, 0.114]	0.125 [-0.048, 0.297]
	Black vs White	-0.23 [-1.164, 0.704]	-0.55 [-1.512, 0.406]
	Hispanic or Latino vs White	0.156 [-0.25, 0.561]	0.073 [-0.325, 0.48]
	Native American vs White	0.107 [-1.219, 1.434]	0.076 [-1.257, 1.394]
	Other vs White	0.409 [-0.037, 0.857]	0.237 [-0.204, 0.679]
	No Answer vs White	0.055 [-0.19, 0.299]	0.009 [-0.236, 0.254]
	Ave. Citations of 1st Author's Institute	0.029 [-0.121, 0.178]	0.032 [-0.121, 0.185]
	Ave. Citations of Authors' Institutes	0.135 [-0.014, 0.286]	0.079 [-0.073, 0.232]
	Economics Ph.D. vs. Psychology	0.089 [-0.203, 0.382]	0.133 [-0.166, 0.433]
	Management Ph.D. vs Psychology	0.078 [-0.169, 0.326]	0.155 [-0.091, 0.4]
	Marketing Ph.D. vs. Psychology	0.039 [-0.149, 0.23]	0.008 [-0.185, 0.2]
	Other Ph.D. vs. Psychology	0.102 [-0.122, 0.331]	-0.005 [-0.235, 0.221]
	Undergraduate vs. Full	0.13 [-1.317, 1.566]	-1.177 [-2.64, 0.269]
	Masters vs. Full	-0.1 [-0.8, 0.599]	-0.356 [-1.067, 0.362]
	Ph.D. vs. Full	0.023 [-0.494, 0.546]	-0.574 [-1.095, -0.052]
	Post Doc vs. Full	-0.056 [-0.54, 0.437]	-0.458 [-0.95, 0.04]
	Research Scientist vs. Full	0.041 [-0.588, 0.675]	-0.506 [-1.155, 0.142]
	Practitioner vs. Full	-0.44 [-1.189, 0.32]	-0.757 [-1.53, 0.021]
	Assistant vs. Full	0.102 [-0.345, 0.552]	-0.285 [-0.734, 0.169]
	Associate vs. Full	-0.094 [-0.5, 0.32]	-0.294 [-0.705, 0.118]
Co-Author	Ave. Seniority of Authors	0.138 [0.059, 0.217]	0.061 [-0.019, 0.141]
	Empirical Studies vs. Traditional Topic	-0.022 [-0.184, 0.138]	0.046 [-0.117, 0.208]
	Psychology of JDM vs. Traditional Topic	-0.049 [-0.263, 0.165]	0.005 [-0.207, 0.22]
	Sentiment	-0.07 [-0.14, 0]	-0.013 [-0.083, 0.058]
Reviewer	Reviewer Female (0) or Male (1)	-0.01 [-0.129, 0.107]	-0.008 [-0.142, 0.126]
	Reviewer Years Since Ph.D.	0.039 [-0.054, 0.132]	-0.078 [-0.166, 0.011]
	Reviewer Ave. Citations of 1st Author's Institute	0.021 [-0.032, 0.074]	0.022 [-0.043, 0.086]
	Economics Ph.D. vs. Psychology	0.002 [-0.181, 0.185]	0.027 [-0.169, 0.223]
	Management Ph.D. vs Psychology	-0.002 [-0.18, 0.175]	0.036 [-0.143, 0.215]
	Marketing Ph.D. vs. Psychology	-0.011 [-0.155, 0.133]	-0.034 [-0.202, 0.132]
	Other Ph.D. vs. Psychology		0.047 [-0.145, 0.24]
	Research Scientist vs. Full	-0.068 [-0.352, 0.325]	-0.163 [-0.48, 0.151]
	Practitioner vs. Full		-0.203 [-0.654, 0.244]
	Assistant vs. Full	0.095 [-0.135, 0.325]	-0.13 [-0.341, 0.082]
	Associate vs. Full	0.058 [-0.126, 0.271]	-0.069 [-0.284, 0.146]
	Reviewer # Times Review Past 7 Years	-0.032 [-0.098, 0.034]	0.01 [-0.054, 0.073]

Bolded values indicated credible interval excludes 0. All co-author characteristic characteristics were represented by the statistics based on all the authors (e.g., proportion of male authors instead of male coauthors) so that single-author submissions were included in the analysis. Thus, the regression coefficients for the first author variables represent incremental change for the variables in which there are corresponding coauthor characteristics. Review ratings and numerical predictors were standardized. The z indicates when the variables were standardized. Review ratings and numerical predictors were standardized. The characteristics were entered simultaneously for both regression models.

Table 16: Coef. from regressing submission characteristics onto single- or double-blind review ratings separately for multi-author submissions.

Category	Predictor	Single	Double
	Intercept	0.002 [-0.495, 0.497]	0.257 [-0.245, 0.755]
First Author	Female (0) or Male (1)	0.182 [0.04, 0.325]	0.302 [0.158, 0.445]
	Proportion Male Authors	0.033 [-0.038, 0.103]	-0.039 [-0.112, 0.033]
	Years Since Ph.D.	-0.121 [-0.267, 0.028]	-0.17 [-0.319, -0.019]
	Domestic (0) or International (1)	0.067 [-0.084, 0.216]	0.021 [-0.132, 0.175]
	Asian vs White	-0.067 [-0.239, 0.106]	0.135 [-0.043, 0.311]
	Black vs White	-0.218 [-1.146, 0.707]	-0.566 [-1.514, 0.386]
	Hispanic or Latino vs White	0.137 [-0.258, 0.53]	0.028 [-0.378, 0.43]
	Native American vs White	0.097 [-1.182, 1.37]	0.059 [-1.242, 1.368]
	Other vs White	0.476 [0.02, 0.922]	0.32 [-0.142, 0.776]
	No Answer vs White	0.092 [-0.153, 0.34]	0.012 [-0.241, 0.265]
	Ave. Citations of 1st Author's Institute	0.065 [-0.031, 0.163]	0.048 [-0.053, 0.15]
	Ave. Citations of Authors' Institutes	0.097 [0.001, 0.193]	0.051 [-0.048, 0.15]
	Economics Ph.D. vs. Psychology	0.171 [-0.126, 0.468]	0.203 [-0.106, 0.512]
	Management Ph.D. vs. Psychology	0.155 [-0.093, 0.405]	0.194 [-0.061, 0.45]
	Marketing Ph.D. vs. Psychology	0.057 [-0.136, 0.249]	0.046 [-0.151, 0.246]
	Other Ph.D. vs. Psychology	0.166 [-0.065, 0.396]	0.03 [-0.204, 0.264]
	Undergraduate vs. Full	-0.336 [-1.72, 1.05]	-1.429 [-2.843, -0.015]
	Masters vs. Full	-0.395 [-1.087, 0.291]	-0.487 [-1.188, 0.215]
	Ph.D. vs. Full	-0.231 [-0.735, 0.276]	-0.671 [-1.191, -0.144]
	Post Doc vs. Full	-0.296 [-0.778, 0.185]	-0.559 [-1.054, -0.06]
	Research Scientist vs. Full	-0.181 [-0.801, 0.438]	-0.593 [-1.235, 0.051]
	Practitioner vs. Full	-0.613 [-1.354, 0.14]	-0.852 [-1.616, -0.077]
	Assistant vs. Full	-0.027 [-0.48, 0.424]	-0.357 [-0.821, 0.105]
	Associate vs. Full	-0.138 [-0.56, 0.275]	-0.347 [-0.775, 0.081]
Co-Author	Ave. Seniority of Authors	0.107 [0.031, 0.183]	0.012 [-0.067, 0.09]
	Empirical Studies vs. Traditional Topic	-0.033 [-0.195, 0.128]	0.079 [-0.086, 0.245]
	Psychology of JDM vs. Traditional Topic	-0.05 [-0.263, 0.165]	0.065 [-0.153, 0.284]
	Sentiment	-0.07 [-0.14, 0]	-0.027 [-0.099, 0.045]
Reviewer	Reviewer Female (0) or Male (1)	-0.022 [-0.144, 0.099]	-0.021 [-0.159, 0.118]
	Reviewer Years Since Ph.D.	0.048 [-0.048, 0.142]	-0.066 [-0.157, 0.026]
	Reviewer Ave. Citations of 1st Author's Institute	0.036 [-0.018, 0.091]	0.031 [-0.037, 0.099]
	Economics Ph.D. vs. Psychology	-0.035 [-0.222, 0.153]	0.07 [-0.132, 0.271]
	Management Ph.D. vs. Psychology	-0.041 [-0.225, 0.145]	0.05 [-0.134, 0.234]
	Marketing Ph.D. vs. Psychology	-0.042 [-0.19, 0.107]	0.005 [-0.168, 0.181]
	Other Ph.D. vs. Psychology	-0.084 [-0.38, 0.209]	0.094 [-0.107, 0.293]
	Research Scientist vs. Full		-0.067 [-0.396, 0.262]
	Practitioner vs. Full		-0.206 [-0.674, 0.261]
	Assistant vs. Full	0.123 [-0.112, 0.357]	-0.069 [-0.285, 0.149]
	Associate vs. Full	0.086 [-0.101, 0.271]	-0.061 [-0.28, 0.161]
	Reviewer # Times Review Past 7 Years	-0.034 [-0.102, 0.035]	0.014 [-0.052, 0.081]

Bolded values indicated credible interval excludes 0. Review ratings and numerical predictors were standardized. The characteristics in the single- and double-blind were simultaneously for both regressions. The data were subsetted to only include multi-author submissions.

7 Comparing Single- vs. Double-blind in the Top 108

The comparison between single- and double-blind review revealed there were three author characteristics that had a differential impact on single- and double-blind review: first author gender, whether the first author identified as Asian or not, and the seniority of the coauthors. We were curious how these effects would be realized if the top 108 submissions were identified by single-blind or double-blind reviews. In terms of the gender difference, if we look at the top 108 as specified by the single-blind review ratings and do the same for double-blind review, then double-blind review has less of a gender difference than single-blind. In the top 108 submission according to single-blind review there were 35 submissions (32%) with female first authors while according to double-blind review there were 39 submissions (36%). Note across all 530 submissions there were 246 female first authors (46%). Thus, for both review systems there would be a gender difference with less representation of women in the hypothetical conference, but if anything the gender difference would be slightly stronger for single-blind review. This difference reinforces the fact that the difference between single- and double-blind in terms of gender is subtle and should be interpreted with caution.

In terms of race, if the top 108 submissions were identified by the single-blind review ratings the number of Asian first authors would be $N = 19$ (18%). In contrast, if the top 108 were identified by double-blind review ratings then the number of Asian first authors would be $N = 29$ (26%). This difference illustrates an important aspect of double-blind review that it can help address potential inequities that arise with single-blind review.

We can do the same analysis and ask how the representation of senior coauthorship would change if the selection was done with single-blind review or double-blind review. If the top 108 submissions were identified by the single-blind review ratings the mean coauthor rank would be 8.3 ($SD = 1.7$) (i.e., between an Associate and a Full Professor). In contrast, if the top 108 were identified by double-blind review then

the mean coauthor rank would be 8.0 ($SD = 1.9$) (i.e., an Associate Professor). Across all 530 submissions the mean coauthor rank excluding the first author was 7.9 ($SD = 1.9$) (i.e., just below Associate Professor). Thus, from the perspective of the conference threshold, selection of talks by single-blind review would have advanced submissions with more senior coauthors, but not in double-blind.

8 Predictive Validity of Review Ratings and Submission Characteristics

8.1 Summary of Talk Outcomes Across sessions

Table 17: Mean (M) and standard deviation (SD) of talk ratings, attendance, and total potential questions

Session	Overall Rating (Std.)		Attendance		Potential Questions	
	M	SD	M	SD	M	SD
1	-0.13	0.45	99.2	23.8	3.3	1.5
2	-0.04	0.31	117.3	21.4	3.3	2.0
3	0.09	0.40	102.8	30.3	2.5	1.6
4	0.07	0.39	89.5	20.9	2.0	1.1
5	0.20	0.40	100.6	37.6	3.4	1.3
6	-0.08	0.29	106.7	43.0	2.5	1.9
7	0.14	0.43	102.4	44.7	3.1	1.8
8	0.05	0.39	62.6	28.8	2.8	1.3
9	0.19	0.27	65.1	17.1	1.9	1.8

Talk ratings were standardized within each talk evaluator.

8.2 Correlations Between Outcomes



Figure 6: Correlations between outcome data. *Note:* The central value is the mean posterior estimate. The values in the upper and lower right-hand corners are the estimates of the upper and lower bounds of the 95% credible interval. Starred values indicate the credible interval excludes 0. Inter correlations between talk rating and talk/poster rating and published and journal impact factor are not informative and are thus excluded. Journal impact factor was log-transformed before computing the correlation.

8.3 Regression of Outcomes onto Single- and Double-blind Review Ratings

Table 18: Coefficients from regressing conference outcomes onto single- and double-blind review ratings simultaneously.

	Single		Double	
	M	95% CI	M	95% CI
Talk Rating (z)	0.048	[-0.092, 0.190]	-0.018	[-0.139, 0.106]
Attendance (z)	-0.119	[-0.438, 0.207]	-0.117	[-0.397, 0.166]
Total Potential Questions (z)	0.017	[-0.326, 0.368]	0.164	[-0.158, 0.495]
Talk & Poster Rating (z)	0.326	[0.112, 0.536]	0.191	[-0.022, 0.400]
Published	0.184	[-0.049, 0.410]	0.368	[0.132, 0.604]
Journal Impact Factor (z)	0.204	[0.001, 0.409]	0.131	[-0.075, 0.342]

Bolded values indicated credible interval excludes 0. All regressions included a dummy variable indicating whether the submission was presented at the conference or not. The z indicates that the variables were standardized.

8.4 Regression of Outcomes Onto Author and Submission Characteristics

8.4.1 Regression with Standardized Outcome Values.

Table 19: Coefficients from regressing conference outcomes onto submission characteristics.

		M [95% CI]					
	Characteristic	Overall (z)	Attendance (z)	Total Potential Questions (z)	Talk/Poster Rating (z)	Published Dec 2020	Impact Factor of Journal Publi
First Authors	(Intercept)	0.114 [-0.677, 0.897]	0.209 [-1.327, 1.748]	0.557 [-1.354, 2.458]	-0.382 [-2.177, 1.408]	0.874 [-0.733, 2.479]	-0.266 [-1.418, 0.884]
	Female (0) or Male (1)	0.032 [-0.282, 0.346]	-0.604 [-1.206, 0.001]	-0.355 [-1.102, 0.387]	0.361 [-0.225, 0.955]	-0.059 [-0.774, 0.637]	-0.081 [-0.598, 0.432]
	Years Since Ph.D. (z)	-0.176 [-0.486, 0.143]	0.444 [-0.169, 1.07]	-0.216 [-0.966, 0.552]	-0.342 [-1.059, 0.388]	-0.457 [-0.949, 0.011]	-0.045 [-0.41, 0.322]
	Domestic (0) or International (1)	-0.005 [-0.274, 0.268]	0.113 [-0.412, 0.641]	0.264 [-0.405, 0.937]	0.063 [-0.425, 0.552]	-0.105 [-0.663, 0.462]	0.049 [-0.382, 0.478]
	Asian vs White	0.358 [0.035, 0.68]	-0.122 [-0.752, 0.508]	-0.22 [-1.007, 0.574]	0.699 [0.144, 1.264]	-0.007 [-0.672, 0.638]	0.173 [-0.311, 0.653]
	Black vs White					-25.527 [-71.874, -0.422]	
	Hispanic or Latino vs White	0.311 [-0.22, 0.848]	0.804 [-0.225, 1.851]	0.76 [-0.549, 2.066]	0.417 [-0.928, 1.729]	0.138 [-1.384, 1.606]	0.775 [-0.233, 1.786]
	Native American vs White	-0.053 [-0.98, 0.868]	0.404 [-1.359, 2.181]	-0.626 [-2.831, 1.597]	-0.229 [-2.527, 2.046]	-35.845 [-104.303, -0.209]	
	Other vs White	-0.109 [-0.66, 0.442]	-0.214 [-1.286, 0.853]	-0.328 [-1.655, 1.027]	-0.159 [-1.519, 1.176]	-0.638 [-2.795, 1.346]	0.682 [-0.751, 2.119]
	No Answer vs White	-0.14 [-0.585, 0.308]	-0.195 [-1.055, 0.662]	-0.614 [-1.674, 0.461]	0.33 [-0.551, 1.228]	-0.174 [-1.087, 0.728]	0.231 [-0.489, 0.955]
	Economics Ph.D. vs. Psychology	0.124 [-0.384, 0.629]	-0.504 [-1.487, 0.47]	0.45 [-0.759, 1.65]	0.022 [-1.059, 1.148]	-1.416 [-2.661, -0.285]	0.751 [-0.258, 1.76]
	Marketing Ph.D. vs. Psychology	0.05 [-0.245, 0.352]	-0.063 [-0.636, 0.515]	0.332 [-0.393, 1.045]	0.103 [-0.469, 0.672]	-1.494 [-2.259, -0.747]	0.177 [-0.45, 0.8]
	Other Ph.D. vs. Psychology	0.124 [-0.201, 0.456]	0.543 [-0.081, 1.176]	0.553 [-0.246, 1.346]	0.302 [-0.391, 0.992]	-0.286 [-1.076, 0.493]	-0.127 [-0.711, 0.465]
	Ave. Citations of 1st Author's Institute (z)	0.281 [0.05, 0.513]	0.748 [0.309, 1.188]	0.015 [-0.551, 0.584]	0.394 [-0.175, 0.955]	-0.371 [-0.926, 0.166]	0.143 [-0.277, 0.56]
	Undergraduate vs. Full				-4.373 [-7.67, -1.167]	-36.434 [-103.785, -1.079]	
	Masters vs. Full	-0.492 [-1.529, 0.563]	0.685 [-1.315, 2.751]	-0.38 [-2.868, 2.155]	-1.874 [-4.1, 0.348]	-2.072 [-4.747, 0.473]	0.18 [-1.581, 1.967]
	Ph.D. vs. Full	-0.674 [-1.592, 0.273]	-0.562 [-2.381, 1.321]	-0.865 [-3.094, 1.434]	-1.483 [-3.548, 0.588]	-1.479 [-3.244, 0.262]	0.084 [-1.149, 1.327]
	Post Doc vs. Full	-0.581 [-1.402, 0.265]	-0.244 [-1.858, 1.422]	-0.488 [-2.473, 1.569]	-1.178 [-3.049, 0.735]	-1.048 [-2.706, 0.59]	0.281 [-0.838, 1.408]
	Research Scientist vs. Full	0.018 [-1.134, 1.201]	-0.489 [-2.749, 1.755]	-1.287 [-4.115, 1.531]	0.118 [-2.675, 2.837]	-16.051 [-40.117, -3.427]	
	Practitioner vs. Full	-0.205 [-1.315, 0.916]	0.461 [-1.678, 2.621]	0.921 [-1.736, 3.618]	-0.615 [-3.276, 2.105]	-20.059 [-52.881, -2.717]	
All Authors	Assistant vs. Full	-0.536 [-1.272, 0.212]	-0.047 [-1.498, 1.454]	-0.444 [-2.238, 1.391]	-0.937 [-2.603, 0.705]	-1.014 [-2.547, 0.516]	0.162 [-0.837, 1.152]
	Associate vs. Full	-0.497 [-1.153, 0.161]	-0.816 [-2.08, 0.464]	-1.128 [-2.704, 0.444]	-0.782 [-2.347, 0.765]	-1.075 [-2.527, 0.323]	0.237 [-0.8, 1.266]
	Proportion Male Authors (z)	0.003 [-0.151, 0.157]	0.19 [-0.1, 0.489]	0.31 [-0.082, 0.698]	-0.203 [-0.499, 0.096]	0.525 [0.16, 0.898]	0.174 [-0.093, 0.45]
Submission	Ave. Citations of Authors' Institutes (z)	-0.183 [-0.426, 0.069]	-0.353 [-0.819, 0.117]	0.135 [-0.454, 0.729]	-0.163 [-0.706, 0.382]	0.321 [-0.223, 0.871]	0.198 [-0.246, 0.64]
	Ave. Seniority of Authors (z)	0.034 [-0.134, 0.206]	-0.288 [-0.611, 0.036]	-0.213 [-0.612, 0.188]	0.011 [-0.304, 0.33]	0.203 [-0.121, 0.532]	0.041 [-0.199, 0.28]
Submission	Empirical Studies vs. Traditional Topic	0.198 [-0.089, 0.483]	0.34 [-0.224, 0.895]	0.361 [-0.331, 1.052]	0.393 [-0.12, 0.9]	0.077 [-0.527, 0.69]	-0.127 [-0.601, 0.351]
	Psychology of JDM vs. Traditional Topic	0.105 [-0.299, 0.507]	0.48 [-0.297, 1.245]	0.001 [-0.962, 0.977]	0.304 [-0.422, 1.029]	0.292 [-0.509, 1.084]	-0.056 [-0.625, 0.528]
Session	Sentiment (z)	-0.047 [-0.171, 0.08]	-0.04 [-0.287, 0.208]	0.013 [-0.288, 0.314]	-0.164 [-0.419, 0.091]	-0.192 [-0.458, 0.072]	0.043 [-0.161, 0.243]
	Session 2 vs. 1	0.23 [-0.232, 0.698]	1.24 [0.359, 2.124]	-0.111 [-1.248, 1.04]	0.614 [-0.517, 1.75]		
	Session 3 vs. 1	0.074 [-0.413, 0.564]	0.42 [-0.525, 1.345]	-0.646 [-1.803, 0.543]	0.308 [-0.868, 1.464]		
	Session 4 vs. 1	0.184 [-0.262, 0.621]	0.392 [-0.463, 1.255]	-0.312 [-1.384, 0.756]	0.376 [-0.754, 1.471]		
	Session 5 vs. 1	0.063 [-0.378, 0.509]	-0.568 [-1.431, 0.299]	0.275 [-0.872, 1.464]	0.483 [-0.612, 1.556]		
	Session 6 vs. 1	-0.041 [-0.504, 0.421]	0.101 [-0.807, 0.989]	-0.192 [-1.336, 0.958]	0.393 [-0.737, 1.511]		
	Session 7 vs. 1	0.061 [-0.389, 0.517]	-0.186 [-1.045, 0.681]	-0.393 [-1.489, 0.722]	0.438 [-0.739, 1.582]		
	Session 8 vs. 1	0.242 [-0.182, 0.671]	-1.134 [-1.946, -0.325]	-0.239 [-1.247, 0.798]	0.852 [-0.193, 1.88]		
	Session 9 vs. 1	0.073 [-0.359, 0.509]	-0.804 [-1.643, 0.052]	-1.162 [-2.217, -0.087]	0.513 [-0.535, 1.565]		
	Poster Session 10 vs. 1				1.126 [0.225, 1.999]		
	Presentation					0.215 [-0.428, 0.865]	0.104 [-0.377, 0.592]
	Noise	sigma	0.386 [0.309, 0.486]	0.742 [0.596, 0.932]	0.921 [0.734, 1.17]	0.991 [0.844, 1.172]	0.961 [0.83, 1.114]

Bolded values indicated credible intervals exclude 0. As with the regressions comparing single- vs. double-blind reviews, all coauthor characteristic characteristics were represented by the statistics based on all the authors (e.g., the proportion of male authors instead of male coauthors) so that single-author submissions were included in the analysis. Thus, the regression coefficients for the first author variables represent incremental change for the variables with corresponding coauthor characteristics. The z indicates when the variables were standardized. When regressing conference outcomes (talk rating, attendance, questions, and talk/poster rating) conference session was entered as a covariate in the model. When regressing publication outcomes, whether the submission was presented at the conference or not was entered as a covariate. Table 21 provides the regression coefficients when subsetting the data on multiple author submissions.

8.4.2 Regression with Raw Outcome Values.

Table 20: Coefficients from regressing conference outcomes onto single- and double-blind review ratings simultaneously.

		M [95% CI]	
	Characteristic	Attendance	Total Potential Questions
	(Intercept)	101.6 [48.8, 154.1]	3.8 [0.7, 6.9]
First Authors	Female (0) or Male (1)	-21.1 [-42.5, -0.3]	-0.6 [-1.8, 0.7]
	Years Since Ph.D. (z)	15.2 [-5.6, 36.7]	-0.4 [-1.6, 0.8]
	Domestic (0) or International (1)	3.9 [-14.6, 22.6]	0.4 [-0.7, 1.5]
	Asian vs White	-4.3 [-26.2, 17.7]	-0.4 [-1.7, 0.9]
	Black vs White	-	-
	Hispanic or Latino vs White	27.8 [-9, 63.8]	1.2 [-0.9, 3.3]
	Native American vs White	13.9 [-47.8, 74.7]	-1 [-4.6, 2.6]
	Other vs White	-7.5 [-44, 28.6]	-0.5 [-2.8, 1.6]
	No Answer vs White	-6.8 [-36.8, 23.5]	-1 [-2.8, 0.7]
	Economics Ph.D. vs. Psychology	-17.5 [-51.9, 16.2]	0.7 [-1.2, 2.7]
	Marketing Ph.D. vs. Psychology	-2.1 [-22.2, 17.4]	0.6 [-0.6, 1.7]
	Other Ph.D. vs. Psychology	19.1 [-2.8, 40.8]	0.9 [-0.4, 2.2]
	Ave. Citations of 1st Author's Institute (z)	26.1 [10.1, 41.8]	0 [-0.9, 1]
	Undergraduate vs. Full	-	-
	Masters vs. Full	22.8 [-48, 92.6]	-0.7 [-4.7, 3.3]
	Ph.D. vs. Full	-20.1 [-84.5, 43.4]	-1.5 [-5.1, 2.1]
	Post Doc vs. Full	-9 [-65.8, 47.7]	-0.8 [-4.1, 2.3]
	Research Scientist vs. Full	-17.5 [-96.6, 61.9]	-2.1 [-6.7, 2.4]
	Practitioner vs. Full	15.3 [-58.9, 89.8]	1.5 [-2.8, 5.7]
	Assistant vs. Full	-2.1 [-52.6, 48.6]	-0.8 [-3.6, 2.2]
	Associate vs. Full	-28.6 [-72.9, 16.3]	-1.9 [-4.5, 0.7]
All Authors	Proportion Male Authors (z)	6.6 [-3.5, 16.9]	0.5 [-0.1, 1.2]
	Ave. Citations of Authors' Institutes (z)	-12.2 [-28.7, 4.2]	0.2 [-0.7, 1.2]
	Ave. Seniority of Authors (z)	-10.1 [-21.5, 1.3]	-0.4 [-1, 0.3]
	Empirical Studies vs. Traditional Topic	12 [-7.4, 31.1]	0.6 [-0.6, 1.7]
	Psychology of JDM vs. Traditional Topic	16.8 [-10.5, 44]	0 [-1.6, 1.6]
	Sentiment (z)	-1.4 [-9.7, 7]	0 [-0.5, 0.5]
Session	Session 2 vs. 1	43.2 [11.5, 74.9]	-0.2 [-2.1, 1.7]
	Session 3 vs. 1	14.6 [-18.5, 47.5]	-1.1 [-3, 0.9]
	Session 4 vs. 1	13.5 [-16.5, 43.5]	-0.5 [-2.3, 1.3]
	Session 5 vs. 1	-19.6 [-49.8, 10.3]	0.5 [-1.5, 2.4]
	Session 6 vs. 1	3.8 [-27.3, 35.1]	-0.3 [-2.2, 1.6]
	Session 7 vs. 1	-6.4 [-36.3, 23.2]	-0.7 [-2.5, 1.1]
	Session 8 vs. 1	-39.4 [-67.7, -10.8]	-0.4 [-2.1, 1.3]
	Session 9 vs. 1	-27.7 [-57.6, 2.4]	-1.9 [-3.7 -, 0.1]
Noise	sigma	25.8 [20.7, 32.6]	1.5 [1.2, 1.9]

Bolded values indicated credible interval excludes 0. All regressions included a dummy variable indicating whether the submission was presented at the conference or not. The z indicates when the variables were standardized.

8.4.3 Regression with Multi-author Submissions.

Table 21: Coefficients from regressing conference outcomes onto submission characteristics for multiple author papers.

		M [95% CI]					
Characteristic		Overall (z)	Attendance (z)	Total Potential Questions (z)	Talk/Poster Rating (z)	Published Dec 2020	Impact Factor of Journal Pub
(Intercept)		-0.079 [-0.874, 0.712]	-0.042 [-1.587, 1.513]	0.265 [-1.643, 2.176]	-0.691 [-2.445, 1.082]	1.133 [-0.502, 2.726]	-0.168 [-1.524, 1.184]
First Authors	Female (0) or Male (1)	0.076 [-0.204, 0.354]	-0.553 [-1.085, -0.011]	-0.048 [-0.727, 0.639]	0.21 [-0.284, 0.705]	0.599 [0.052, 1.145]	0.074 [-0.347, 0.489]
	Years Since Ph.D. (z)	-0.081 [-0.417, 0.252]	0.31 [-0.34, 0.969]	-0.215 [-1.015, 0.588]	-0.223 [-0.984, 0.551]	-0.764 [-1.395, -0.188]	-0.095 [-0.769, 0.582]
	Domestic (0) or International (1)	-0.028 [-0.304, 0.245]	0.173 [-0.36, 0.705]	0.256 [-0.445, 0.962]	0.128 [-0.352, 0.623]	-0.157 [-0.759, 0.434]	0.155 [-0.294, 0.604]
	Asian vs White	0.385 [0.059, 0.705]	-0.137 [-0.748, 0.48]	-0.293 [-1.077, 0.503]	0.721 [0.174, 1.27]	-0.021 [-0.719, 0.662]	0.167 [-0.313, 0.652]
	Black vs White	-	-	-	-	-25.487 [-71.859, -0.616]	-
	Hispanic or Latino vs White	0.324 [-0.235, 0.88]	0.706 [-0.369, 1.772]	0.818 [-0.535, 2.202]	0.33 [-0.973, 1.652]	0.005 [-1.514, 1.509]	0.669 [-0.342, 1.675]
	Native American vs White	0.019 [-0.878, 0.929]	0.301 [-1.436, 2.054]	-0.512 [-2.709, 1.701]	-0.085 [-2.285, 2.092]	-34.544 [-98.133, -0.34]	-
	Other vs White	-0.123 [-0.691, 0.431]	-0.248 [-1.326, 0.831]	-0.269 [-1.649, 1.1]	-0.19 [-1.506, 1.13]	-0.626 [-2.847, 1.361]	0.632 [-0.821, 2.086]
	No Answer vs White	-0.114 [-0.569, 0.334]	-0.14 [-0.997, 0.723]	-0.668 [-1.76, 0.407]	0.204 [-0.687, 1.12]	-0.151 [-1.11, 0.78]	0.202 [-0.569, 0.974]
	Economics Ph.D. vs. Psychology	0.227 [-0.3, 0.754]	-0.854 [-1.887, 0.171]	0.352 [-0.94, 1.648]	0.246 [-0.874, 1.362]	-1.397 [-2.688, -0.23]	0.703 [-0.304, 1.718]
	Marketing Ph.D. vs. Psychology	0.132 [-0.182, 0.442]	-0.185 [-0.805, 0.434]	0.207 [-0.58, 0.988]	0.251 [-0.323, 0.833]	-1.513 [-2.301, -0.776]	0.216 [-0.419, 0.847]
	Other Ph.D. vs. Psychology	0.131 [-0.207, 0.457]	0.574 [-0.07, 1.208]	0.544 [-0.282, 1.359]	0.366 [-0.313, 1.04]	-0.331 [-1.148, 0.467]	0.063 [-0.545, 0.684]
	Ave. Citations of 1st Author's Institute (z)	0.148 [-0.04, 0.335]	0.568 [0.212, 0.931]	0.097 [-0.35, 0.556]	0.159 [-0.241, 0.557]	-0.214 [-0.593, 0.155]	0.23 [-0.046, 0.513]
	Undergraduate vs. Full	-	-	-	-4.114 [-7.095, -1.147]	-35.918 [-101.084, -1.991]	-
	Masters vs. Full	-0.342 [-1.338, 0.664]	0.712 [-1.278, 2.727]	-0.054 [-2.483, 2.383]	-1.669 [-3.776, 0.433]	-2.902 [-5.608, -0.369]	0.009 [-2.03, 2.057]
	Ph.D. vs. Full	-0.468 [-1.335, 0.407]	-0.486 [-2.212, 1.26]	-0.491 [-2.613, 1.626]	-1.014 [-2.991, 0.98]	-2.248 [-4.076, -0.394]	-0.216 [-1.879, 1.436]
	Post Doc vs. Full	-0.335 [-1.137, 0.477]	-0.108 [-1.719, 1.486]	-0.259 [-2.249, 1.742]	-0.657 [-2.517, 1.258]	-1.746 [-3.478, -0.015]	0.045 [-1.397, 1.477]
	Research Scientist vs. Full	0.134 [-1.04, 1.327]	-0.475 [-2.793, 1.766]	-1.168 [-4.056, 1.671]	0.537 [-2.172, 3.289]	-16.001 [-38.412, -3.963]	-
	Practitioner vs. Full	0.044 [-1.095, 1.224]	0.365 [-1.869, 2.612]	0.877 [-1.979, 3.677]	-0.21 [-2.905, 2.563]	-19.75 [-50.954, -3.129]	-
	Assistant vs. Full	-0.324 [-1.093, 0.434]	-0.139 [-1.653, 1.385]	-0.37 [-2.229, 1.487]	-0.474 [-2.178, 1.238]	-1.236 [-2.855, 0.392]	-0.066 [-1.322, 1.182]
	Associate vs. Full	-0.586 [-1.23, 0.063]	-0.404 [-1.628, 0.834]	-0.799 [-2.379, 0.812]	-1.09 [-2.627, 0.476]	-1.219 [-2.695, 0.243]	0.198 [-0.936, 1.321]
Co-Authors	Proportion Male Authors (z)	-0.027 [-0.148, 0.095]	0.105 [-0.131, 0.344]	0.239 [-0.068, 0.536]	-0.162 [-0.389, 0.062]	0.392 [0.11, 0.686]	0.141 [-0.078, 0.36]
	Ave. Citations of Authors' Institutes (z)	-0.108 [-0.279, 0.062]	-0.176 [-0.493, 0.144]	0.097 [-0.318, 0.508]	-0.102 [-0.464, 0.258]	0.182 [-0.195, 0.564]	0.041 [-0.238, 0.324]
	Ave. Seniority of Authors (z)	0.028 [-0.105, 0.161]	-0.314 [-0.573, -0.059]	-0.08 [-0.41, 0.246]	-0.048 [-0.314, 0.216]	0.037 [-0.239, 0.322]	-0.013 [-0.222, 0.193]
Submission	Empirical Studies vs. Traditional Topic	0.149 [-0.147, 0.451]	0.383 [-0.194, 0.953]	0.305 [-0.424, 1.026]	0.217 [-0.299, 0.739]	-0.013 [-0.633, 0.612]	-0.111 [-0.592, 0.366]
	Psychology of JDM vs. Traditional Topic	0.101 [-0.291, 0.503]	0.421 [-0.342, 1.194]	-0.144 [-1.112, 0.825]	0.191 [-0.535, 0.915]	0.269 [-0.541, 1.095]	-0.082 [-0.675, 0.515]
	Sentiment (z)	-0.056 [-0.184, 0.073]	0.003 [-0.246, 0.257]	0.001 [-0.317, 0.321]	-0.213 [-0.462, 0.038]	-0.234 [-0.518, 0.044]	0.065 [-0.138, 0.273]
Session	Session 2 vs. 1	0.217 [-0.256, 0.688]	1.426 [0.528, 2.318]	-0.205 [-1.357, 0.922]	0.607 [-0.503, 1.735]	-	-
	Session 3 vs. 1	0.102 [-0.415, 0.614]	0.706 [-0.277, 1.71]	-0.702 [-1.952, 0.553]	0.379 [-0.802, 1.569]	-	-
	Session 4 vs. 1	0.081 [-0.397, 0.551]	0.622 [-0.29, 1.514]	-0.275 [-1.449, 0.871]	0.255 [-0.864, 1.354]	-	-
	Session 5 vs. 1	0.192 [-0.284, 0.664]	-0.285 [-1.205, 0.629]	0.227 [-0.966, 1.421]	0.852 [-0.284, 1.987]	-	-
	Session 6 vs. 1	0.013 [-0.473, 0.499]	0.447 [-0.483, 1.403]	-0.168 [-1.372, 1.037]	0.621 [-0.491, 1.732]	-	-
	Session 7 vs. 1	0.112 [-0.334, 0.564]	-0.094 [-0.97, 0.768]	-0.405 [-1.511, 0.699]	0.464 [-0.673, 1.591]	-	-
	Session 8 vs. 1	0.278 [-0.154, 0.705]	-0.893 [-1.723, -0.072]	-0.22 [-1.292, 0.833]	0.935 [-0.111, 1.971]	-	-
	Session 9 vs. 1	0.098 [-0.349, 0.549]	-0.614 [-1.477, 0.243]	-1.121 [-2.228, -0.032]	0.706 [-0.333, 1.736]	-	-
	Poster Session 10 vs. 1	-	-	-	1.234 [0.341, 2.142]	-	-
	Presentation	-	-	-	-	0.224 [-0.433, 0.876]	0.047 [-0.464, 0.549]
	Noise	sigma	0.386 [0.308, 0.489]	0.737 [0.588, 0.936]	0.944 [0.753, 1.2]	0.963 [0.816, 1.141]	0.963 [0.833, 1.12]

Bolded values indicated credible interval excludes 0. The z indicates when the variables were standardized. The data were subsetted only to include multi-author submissions. When regressing conference outcomes (talk rating, attendance, questions, and talk/poster rating) conference session was entered as a covariate in the model. When regressing publication outcomes whether the submission was presented at the conference or not was entered as a covariate.

8.4.4 Regression with Raw outcome Values for Muti-author Submissions.

Table 22: Coefficients from regressing conference outcomes onto single- and double-blind review ratings simultaneously.

Characteristic		Attendance	Total Potential Questions
(Intercept)		91.982 [38.445, 146.689]	3.273 [0.051, 6.397]
First Authors	Female (0) or Male (1)	-19.254 [-37.611, -0.576]	-0.082 [-1.211, 1.028]
	Years Since Ph.D. (z)	-	-
	Domestic (0) or International (1)	10.934 [-11.746, 33.845]	-0.349 [-1.703, 1.017]
	Asian vs White	5.919 [-12.811, 24.45]	0.424 [-0.725, 1.577]
	Black vs White	-	-
	Hispanic or Latino vs White	-4.734 [-26.095, 17.173]	-0.478 [-1.797, 0.831]
	Native American vs White	24.527 [-12.376, 61.372]	1.35 [-0.897, 3.594]
	Other vs White	10.072 [-49.449, 70.552]	-0.835 [-4.531, 2.84]
	No Answer vs White	-8.625 [-45.982, 28.777]	-0.434 [-2.756, 1.839]
	Economics Ph.D. vs. Psychology	-4.754 [-34.54, 25.015]	-1.098 [-2.901, 0.736]
	Marketing Ph.D. vs. Psychology	-30.197 [-65.61, 5.29]	0.587 [-1.59, 2.763]
	Other Ph.D. vs. Psychology	-6.407 [-27.507, 14.976]	0.345 [-0.956, 1.635]
	Ave. Citations of 1st Author's Institute (z)	19.87 [-2.192, 41.439]	0.9 [-0.473, 2.246]
	Undergraduate vs. Full	-6.036 [-17.336, 5.043]	0.158 [-0.512, 0.827]
	Masters vs. Full	25.525 [-42.498, 94.053]	-0.082 [-4.132, 4.03]
	Ph.D. vs. Full	-16.728 [-75.718, 42.992]	-0.807 [-4.341, 2.725]
	Post Doc vs. Full	-3.366 [-59.333, 52.577]	-0.432 [-3.72, 2.936]
	Research Scientist vs. Full	-16.412 [-95.7, 62.75]	-1.904 [-6.642, 2.852]
	Practitioner vs. Full	12.792 [-64.396, 92.065]	1.447 [-3.304, 6.164]
	Assistant vs. Full	-4.308 [-56.56, 47.582]	-0.615 [-3.7, 2.513]
	Associate vs. Full	-13.704 [-56.517, 29.653]	-1.32 [-3.944, 1.36]
Coauthors	Proportion Male Coauthors (z)	3.543 [-4.804, 11.796]	0.389 [-0.108, 0.893]
	Ave. Citations of Coauthors' Institutes (z)	19.834 [7.634, 32.155]	0.158 [-0.597, 0.907]
	Ave. Seniority of Coauthors (z)	-10.979 [-19.796, -2.108]	-0.132 [-0.69, 0.422]
Submission	Empirical Studies vs. Traditional Topic	13.538 [-6.199, 33.002]	0.499 [-0.693, 1.724]
	Psychology of JDM vs. Traditional Topic	14.711 [-12.097, 41.176]	-0.241 [-1.852, 1.38]
	Sentiment (z)	0.071 [-8.448, 8.694]	0.002 [-0.518, 0.536]
Session	Session 2 vs. 1	49.923 [18.87, 81.116]	-0.335 [-2.254, 1.613]
	Session 3 vs. 1	24.838 [-9.127, 59.299]	-1.14 [-3.256, 0.995]
	Session 4 vs. 1	21.694 [-9.732, 53.346]	-0.455 [-2.356, 1.434]
	Session 5 vs. 1	-9.455 [-41.386, 22.649]	0.387 [-1.549, 2.338]
	Session 6 vs. 1	15.743 [-17.035, 48.375]	-0.268 [-2.261, 1.732]
	Session 7 vs. 1	-3.168 [-33.616, 27.078]	-0.666 [-2.495, 1.209]
	Session 8 vs. 1	-30.662 [-59.874, -1.834]	-0.352 [-2.088, 1.422]
	Session 9 vs. 1	-21.142 [-51.825, 9.55]	-1.834 [-3.616, 0.011]
	Poster Session 10 vs. 1	-	-
	Presentation	-	-
Noise	sigma	25.62 [20.409, 32.566]	1.56 [1.242, 1.993]

Bolded values indicated credible interval excludes 0. All regressions included a dummy variable indicating whether the submission was presented at the conference or not. The z indicates when the variables were standardized.

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Gelman, A., Hill, J., & Yajima, M. (2012a). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189–211. doi: 10.1080/19345747.2011.618213
- Gelman, A., Hill, J., & Yajima, M. (2012b). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189–211. doi: 10.1080/19345747.2011.618213
- Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *rstanarm: Bayesian applied regression modeling via Stan*. Retrieved from <https://mc-stan.org/rstanarm> (R package version 2.21.1)
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1), 5228–5235. doi: 10.1073/pnas.0307752101
- Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In *Aaai* (Vol. 4, pp. 755–760).
- Keren, G., & Wu, G. (2015). *The Wiley-Blackwell handbook of judgment and decision making*. John Wiley & Sons, Ltd.
- Kruschke, J. K., & Liddell, T. M. (2018a). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. doi: 10.3758/s13423-016-1221-4
- Kruschke, J. K., & Liddell, T. M. (2018b). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. doi: 10.3758/s13423-016-1221-4
- Morey, R. D., & Rouder, J. N. (2018). Bayesfactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=BayesFactor> (R package version 0.9.12-4.2)
- R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5), 1413–1432. doi: 10.1007/s11222-016-9696-4