Supplementary Material

Contents

EC.1. Preregistration

The study was pregistered. We filed it on July 18, 2018. This was done after the submissions were collected, but before sending the submissions out for review. It can be found at the OSF at this link: OSF Peer Review Link. There are eight deviations from the pre-registration that we should note here.

Deviation 1

The first deviation was in analyzing the reviewer ratings. We planned 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 the disciplinary area was not asked, and it was not feasible to classify authors ex-post. The number of studies was not a useful statistic due to the heterogeneity in the submissions: some studies had zero studies and some 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 use 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 EC.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.

Deviation 2

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 abstracts. As we finalized the talk rating instrument, we modified the dimensions to include significance, methods, results, conclusion, innovation, and uniqueness. We found these dimensions to be more straightforward to rate. Moreover, in the final 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 in the ratings. However, all ratings were standardized among raters, which should help minimize differences.

Deviation 3

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. For instance, we created one group we referred to as Asian or of Asian descent that included people who 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."

Deviation 4

The fourth deviation was our measure of author prestige. Similar to Tomkins et al., we planned to use the count of the number of talks in the previous 5 years of the SJDM conference. Due to the number of early career researchers, we extended this measure to include the count of the number of talks and posters at SJDM. Furthermore, we were able to extend the measure to include the

counts over the previous 10 years. We also stated we would explore using the citation counts from Web of Science and Google Scholar as a measure of prestige. However, it was impractical to obtain these measures across all the authors (approximately 1,100). Therefore, we used the author counts as our measure of prestige for the author.

Deviation 5

The fifth 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. We expanded this count to cover the previous 10 years and include talks and posters. We did this because this variable was highly skewed with a few institutions having most of the talks and many other institutions with very few to no talks. To further address this issue, we obtained two other measures of prestige. One measure was the average citation rate for psychology, economics, business, and social science departments at the institutions. This number was extracted from the Web of Science on on August 29, 2019. The other measure we collected was reviewers' subjective ratings of the prestige of the institutions. To do this, in the Spring of 2023 we contacted the reviewers and asked them to rate the prestige of a random subset of the institutions. We aggregated across reviewers to obtain a single measure of prestige per institution. These three statistics were correlated ($M = 0.55$, $SD = 0.17$). Therefore, after standardizing the three different prestige statics we averaged across them to form an overall measure of prestige and used it as our measure of prestige.

Deviation 6

A sixth deviation was the use of 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.

Deviation 7

A seventh 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. (2012b), Kruschke and Liddell (2018). 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 of the effects.

Deviation 8

An eight 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.

EC.2. Methods

EC.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. An *a priori* 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 asked 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 and Liddell (2018) evaluating the probability of obtaining a particular goal. In this case, the goal was if a 95% credible interval for a given posterior distribution excluded 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 and 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. 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.

EC.1 lists the proportion of times (out of 10,000 repetitions) 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 including (a) the number of submissions that were accepted as talks ($n = 108$); (b) the number of multiauthor submissions with a complete set of variables used in the regressions ($n = 370$); (c) the number of multiauthor submissions for which the first author was the corresponding author ($n = 430$); (d) the number of submissions for which the first author was the corresponding author ($n = 470$); and (e) 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, we conclude the 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 $\rho = .55$.[11] Then, for each pair of variables, we calculated the posterior distribution over the Pearson correlation and took the difference in the two distributions. Table EC.2 lists the proportion of times (out of 10,000 repetitions) that the 95% credible interval of the posterior distribution of differences in correlation

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

contained 0 for a given difference in the correlation. The table shows that for the analyses on the more extensive 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 greater than 80% of the time. For the analyses based on the smaller sample of submissions given only as a talk, the corresponding difference 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 13 author and submission characteristics and 5 reviewer characteristics. Second, each submission received at least three single-blind and three 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 modeled as 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 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 hold for the interaction effects.

Altogether, we believe our field study is informative and well-powered for the debate on single- vs. double-blind review. We explicitly chose to study this question in the field instead of 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 EC.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.

	Prop. Excludes 0				
Correlation	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

Table EC.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.

	Prop. 95% Credible Interval Excludes 0				
Difference in Correlations	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

Table EC.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. 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.

EC.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.

EC.2.3. Submission Characteristics

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

EC.2.4. Materials

EC.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, a 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 EC.4 Distribution of the number of authors.

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

EC.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, to make a decision, one set of three double-blind and one set of three single-blind reviews were randomly selected to aggregate across.

EC.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. Upon 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

EC.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

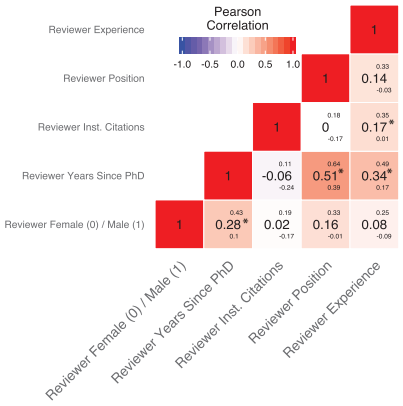

Figure EC.1 Intercorrelations between reviewer characteristics. Values in the center of each cell are posterior estimates of the Pearson correlations, and values in the upper and lower right-hand corner are the corresponding credible intervals. Starred values indicate credible interval excludes 0.

EC.2.4.5. Intercorrelations Between Reviewer Characteristics.

EC.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 who were representative of the society in terms of gender, nationality, and area of expertise (Table EC.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:

Table EC.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%)

EC.2.5.1. Characteristics of Raters.

EC.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?

EC.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 the student and faculty ratings together, giving them equal weight.

EC.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 a count of raised hands indicated potential questions 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.

EC.2.6. Prestige Rating Survey.

In May 2024, we sought additional measures of institutional prestige. This additional survey was approved the Carnegie Mellon University Institutional Review Board. We targeted the original reviewers of the submissions and asked them to complete a survey. Of the $N = 113$ reviewers, we had contact information for $N = 110$. In creating the survey, we identified 328 unique institutions that were associated with at least one author in our 2018 dataset.[12] We randomly assigned a subset of these 328 institutions to each reviewer such that each reviewer was asked to rate the prestige of approximately one quarter of the 328 institutions. They assigned each institution to 1 of 4 prestige levels.

1. World renowned for JDM research.
2. Known for JDM research.
3. A little known for JDM research.

[12] When we include reviewers, there are 334 institutions. As we include the reviewer's institutional prestige in the regressions, it was an error not to include these six institutions. We addressed this issue by averaging the three different prestige measures to form a prestige score so that every institution for authors and reviewers was assigned a prestige score. See EC.2.8.

Table EC.6 Reviewer characteristics who rated the prestige of institutions.

Gender	
Men	35
Women	27
Prefer not to answer	2
Mean Age	44.6 (SD = 8.3, Range 31 to 67; N = 56)
Race	
Asian or Asian American	13
Black, African, or African American	2
White or European	48
prefer not to answer	4
Department	
Cognitive psychology	10
Social psychology	6
Other psychology	1
Marketing	27
Organizational behavior	4
Economics	1
Policy	1
Decision Sciences	14
prefer not to answer	1
Position Rank	
Assistant Professor	12
Associate Professor	22
Full Professor	29
Clinical Professor / Professor of Practice	1
prefer to self describe	1
Continent	
Asia	7
Europe	8
North America	50

4. Not known for JDM research.

We set the *a priori* stop rule to continue data collection until we received at least 20 ratings for each institution. In the end, $N = 65$ reviewers responded to our request and thus we had

approximately 16 ratings for each institution. We used the average rating for each institution and one component of our prestige measure. Table EC.6 lists the characteristics of the respondents. The characteristics indicate our sample of respondents was representative of the original reviewers (Table EC.3.

EC.2.7. Author and Institution Count.

As an additional component of our institutional prestige measure, we examined how often scholars from each institution presented at the SJDM meetings. Across authors and reviewers, there were 334 unique institutions. For each institution, we tabulated the number of times it appeared in the SJDM Annual Meeting program over 10 years (2008 to 2017), affiliated with an author of either a talk or a poster.[13] We did the same thing for author names. That is, for each author who submitted an abstract in 2018, we tabulated how often they appeared in the SJDM programs from 2008 to 2017. To do so, we acquired tab-delimited text files of all the paper and poster submissions to the SJDM conferences between 2008 and 2017. We used the Python function `fuzz.partial_ratio` to identify a matching string of characters with a target string of characters. This function has a similarity value that we can adjust to determine the degree of match needed to count as a match. For institutions, we set this value at .95, and for authors, we set this value at .91. We chose these values after extensive tests with a range of values revealed these values ensured we minimized the chance of misses. But, since we minimized misses, we had an increased chance of false alarms. Thus, three researchers inspected each match and confirmed a match. For the institutions, per year, there were between 5 to 10 false alarms. For the author names, per year there were 3 to 5 false alarms. We then tabulated across all the identified matches, counting the number of matches per talk and poster per year.

EC.2.8. Prestige Score

Figure EC.2 displays the relationship between institutional citation rates, the number of institutional appearances at the annual SJDM meetings from 2008 to 2017, and the prestige rating. The three different measures were correlated. Thus, we standardized each measure and averaged across the three measures to create a prestige score. For some institutions, we did not have all three measures. We addressed this issue by averaging and ignored missing values.

[13] When we solicited prestige ratings, we identified 328 institutions. This difference is because we included reviewers' institutions in the count.

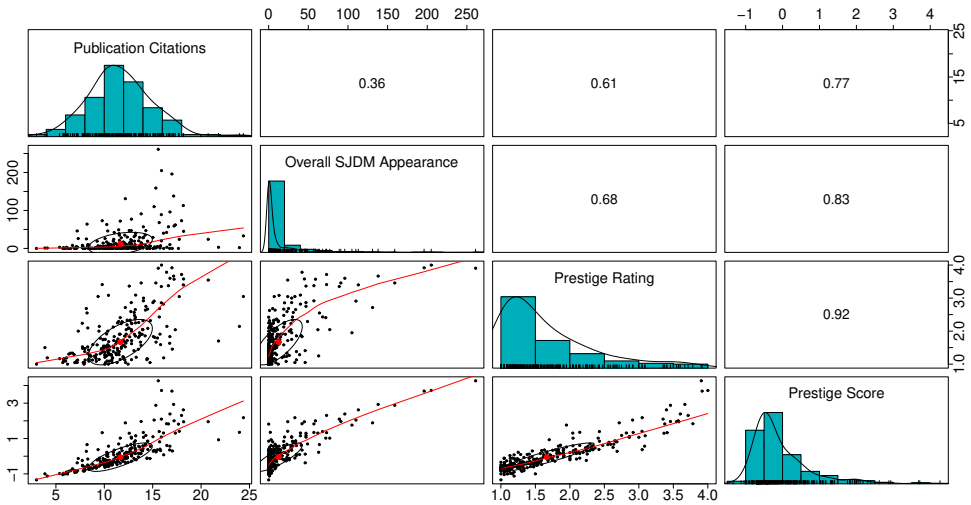

Figure EC.2 Scatterplot matrix of the measures of institutional prestige. Prestige score is the omnibus measure calculated as the mean of the three other standardized measures.

EC.2.9. Poster Ratings.

There were a total of $N = 96$ posters that were submitted for the student-poster competition during the 2018 conference and judged on a similar set of dimensions as those used to judge the quality of the talks. During the conference, student posters were judged for a student poster competition award. To do so, faculty members are recruited to rate the posters. Initially, 94 members volunteered to complete the ratings. Each rater was assigned 5 to 6 posters. The assignment was random with the constraint that the raters did not share an institution or other apparent conflicts of interest (e.g., faculty mentor) with the poster author. Raters were also instructed to withhold ratings if there was a conflict of interest.

A few days before the conference, raters received links to the electronic version of each poster. They were instructed to rate the posters either via the electronic file or in person at the conference. They rated the posters on a 7-point scale using the following 5 dimensions.

Visual presentation and organization Clearly labeled abstract/introduction, methods, results, discussion/conclusions. Visual appeal, appropriate use of figures and tables, clarity of presentation, decent font sizes.

Methodological quality Are the methods appropriate for the topics, in terms of research design and proper use of statistical analyses?

Appropriateness of interpretation Abstract appropriately written, results don't over-interpret or misinterpret, conclusions and discussion don't generalize inappropriately. Conclusion addresses purpose of study as described in introduction.

Significance / Theoretical importance of contribution Does the poster address an important issue? Does the study advance knowledge?

Originality Is the topic treated in a substantially new way? Does the design elaborate or improve on the standard paradigm? Do the results allow for new interpretations or provide a novel source of evidence?

In the end, $N = 69$ raters finished the ratings, resulting in each poster receiving 4 to 5 ratings. A subset of the posters ($N = 57$) had been originally submitted for talks at the conference and 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 used them to investigate how well the review rating systems predicted these ratings. Due to the similar rating systems in how the posters and talks were judged, we also integrated these ratings with the ratings for the submissions given as a talk, creating a new talk/poster rating. Note one poster was also accepted and given as a talk; thus, it has a rating as both a talk and a poster.

EC.2.10. Preference for Review Process Survey.

Shortly before the 2019 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.

EC.2.11. Journal Publications.

Two and six years after the conference, we sought to identify submissions that were published. In total, $N = 276$ submissions were identified as published in a peer-reviewed journal. Our process for identifying publication status was a follows. Two years after the conference, we searched Google Scholar using as a search term the first author's last name and the title of the submission. That yielded 130 published submissions. 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. Responses to the email identified an additional $N = 54$ submissions that had been published. Then in March 2025, we repeated the Google Scholar search and also searched author C.V.s to establish publication status. In total, we identified 276 submissions that were published as of March 13, 2025. We also separately identified the impact factor of each journal via Web of Science and Google Scholar. We used the (logged) impact factor from Web of Science as a metric of impact.

EC.2.12. 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 and 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 $|\Delta elpd_{loo}| < 1\ SE$ 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. 2012a, Kruschke and Liddell 2018).

EC.3. Modeling the Text of the Long Abstracts

EC.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 EC.3 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.

EC.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 and Steyvers 2004).

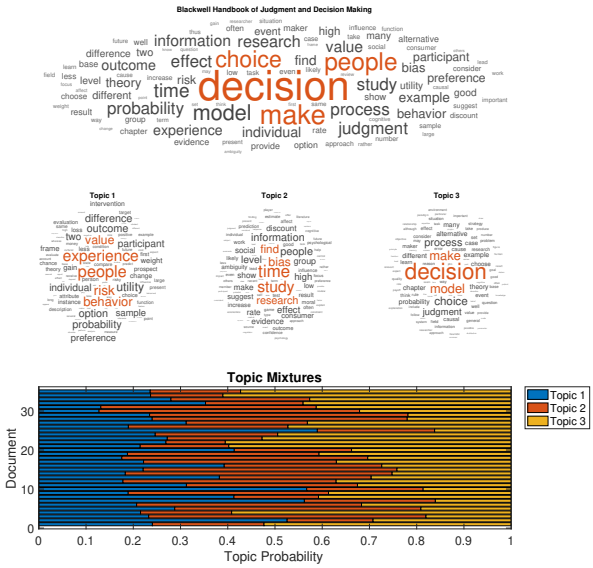

Figure EC.3 Topic modeling of the 2015 Wiley Blackwell Handbook of Judgment and Decision Making. The top plot is a word cloud of the most frequent words in the handbook. The middle plots show the most frequent words associated with the three topics from the LDA topic model. The bottom plot shows the distribution of topics across the 35 chapters.

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.

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 EC.3 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

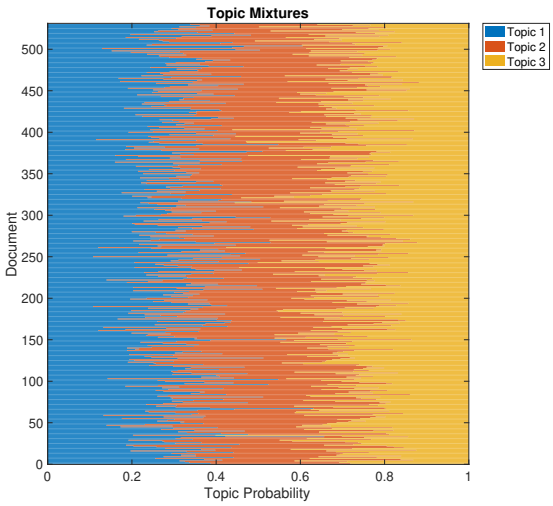

Figure EC.4 Topic distribution across the 530 long abstracts. The topic model came from the best fitting LDA topic model of the 2015 Wiley Blackwell Handbook of Judgment and Decision Making (see Figure EC.3).

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 EC.4 shows the distribution of topics across the 530 long abstracts. The code and output for the topic modeling is available on OSF.

EC.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 and 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 EC.5 shows the word clouds of the words with the strongest negative and positive sentiment in the abstracts. The sentiment of the abstract was determined by averaging across the sentiment scores of each word in the abstract. The average sentiment is shown in the bottom panel of Figure EC.5. The code and output for the sentiment analysis is available on OSF.

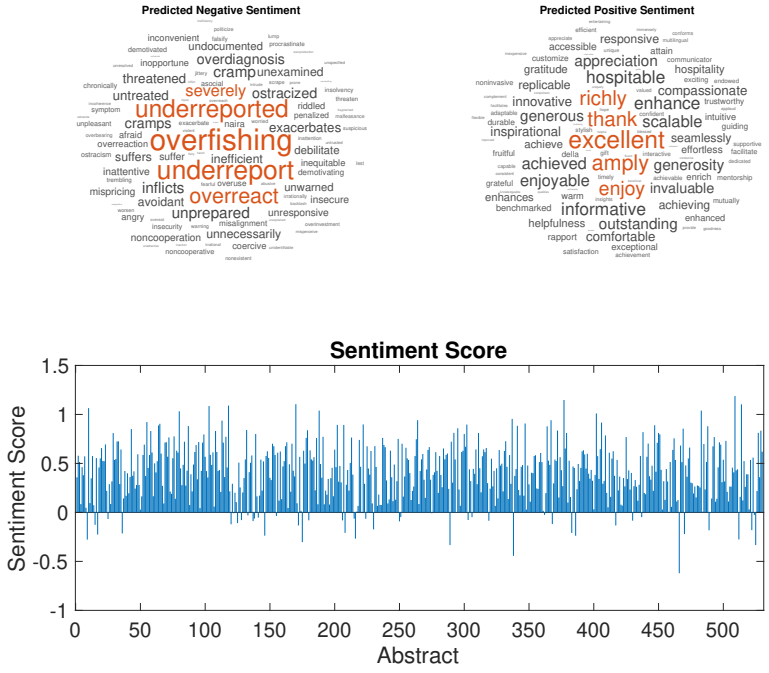

Figure EC.5 Sentiment analysis of the long abstracts. The word clouds show the words with the strongest negative and positive sentiments in the abstracts. The bottom plot is the average sentiment of each of the 530 abstracts.

EC.4. Survey of SJDM Members' Preferences

We survey SJDM member's preferences regarding single- and double-blind review. We asked them to report their preferences from the view point of an author, reviewer, and as an attendee. Details on the methods are in Section EC.2.10. Here we report the analyses.

EC.4.1. Respondent Characteristics

Table EC.7 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%)

EC.4.2. Preference Ratings

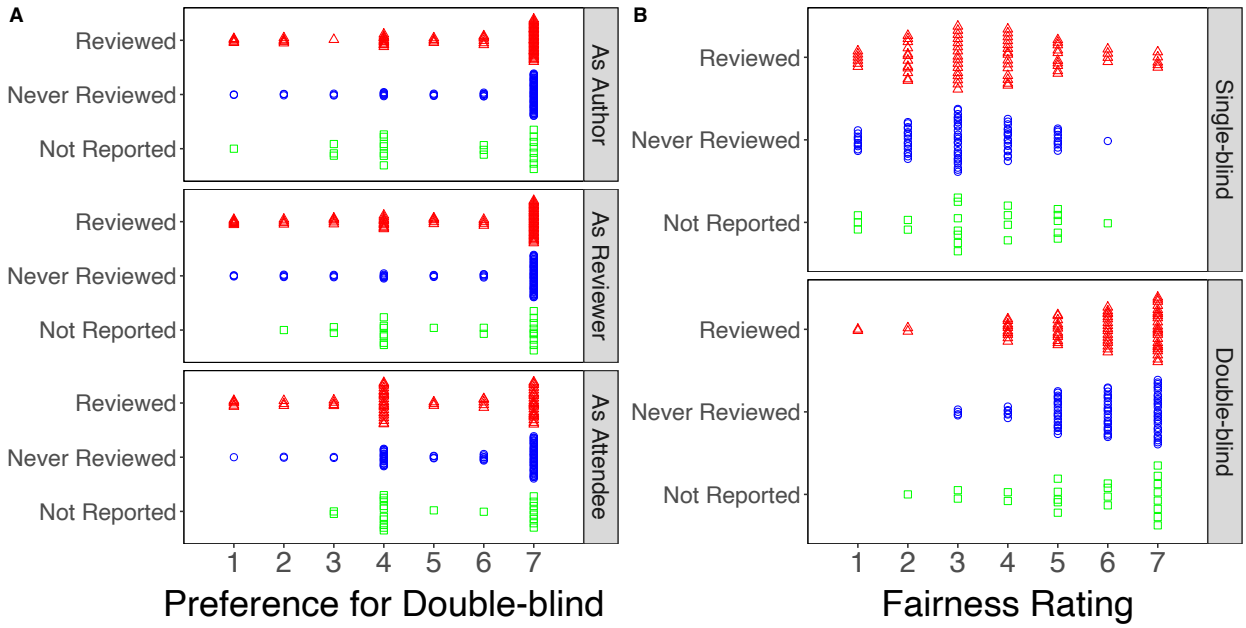

Figure EC.6 **A: Society for Judgment and Decision Making members' surveyed preference for single-blind versus double-blind review for selecting submissions for the annual conference from the perspective of different roles (1 = Strong preference for single-blind, 7 = Strong preference for double-blind). 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.**

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

EC.4.3. Fairness Perceptions

Table EC.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 EC.6 also shows the distribution of fairness ratings. Double-blind review was rated as more fair (Table EC.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 EC.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 EC.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.

EC.5. Differential Use Between Single- and Double-blind Review

EC.5.1. Zero-order Correlations

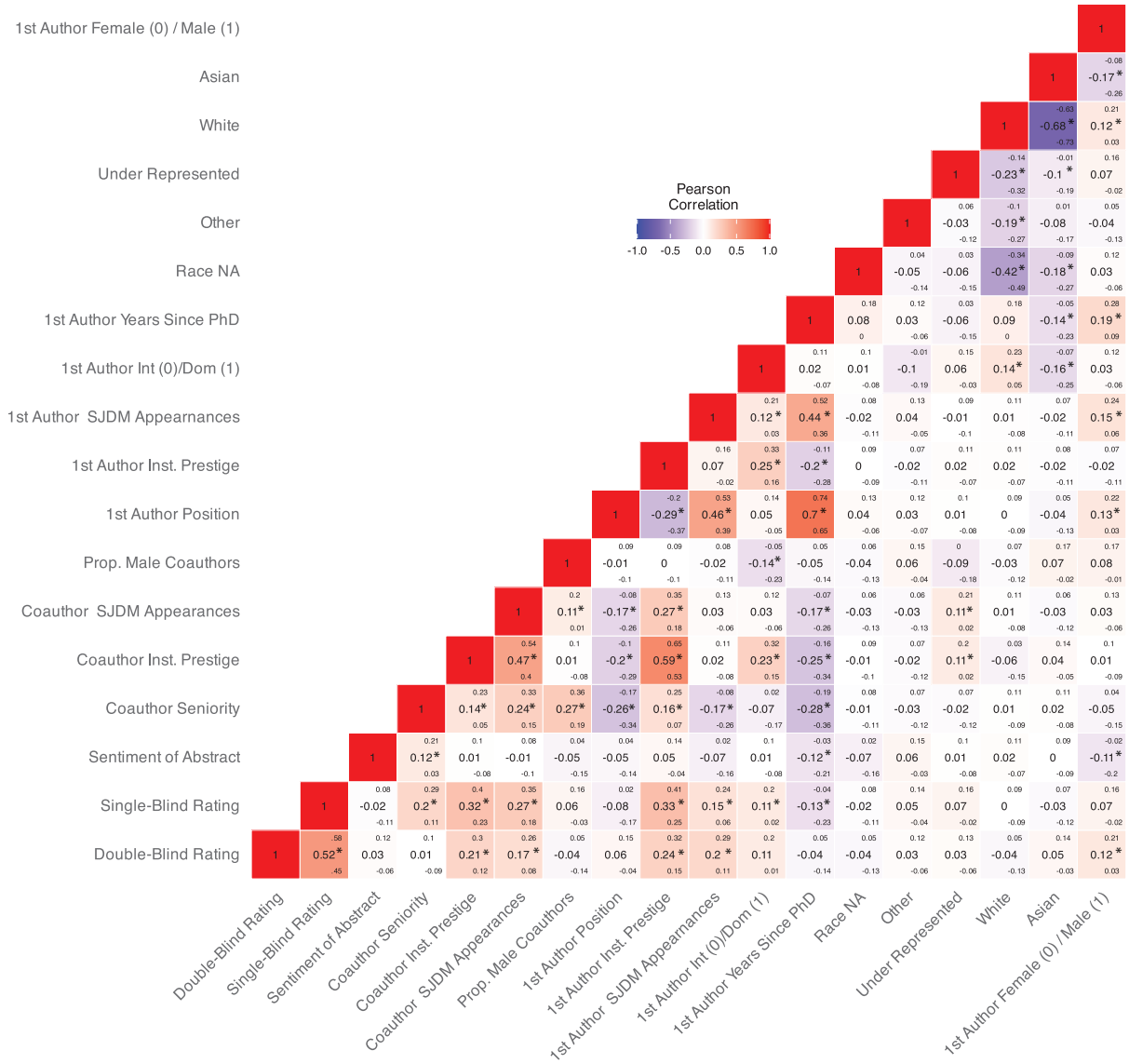

Figure EC.7 Correlations between author and text characteristics and single- vs. double-blind review ratings. The central value is the mean posterior estimate. The values in the upper and lower right-hand corners are the estimates of the upper and lower bounds of the 95% credible interval. Starred values 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 Asian, White, Other, Under Represented, and No Answer were calculated by comparing each group to all others. Under Represented corresponds to Black, Hispanic, and Native American authors who had very low numbers of authors (Table EC.3). 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 EC.10-EC.12.

EC.5.2. Model Comparison

Table EC.9 Model comparison between models with interactions between reviewer characteristics and review condition.

	Δelpd	se(Δelpd)	elpd_{loo}	se(elpd_{loo})
SvB DIF Model	0	0	-3700.311	35.397
SvB DIF Model + Reviewer x Condition	-7.017	2.617	-3707.329	35.441
SvB DIF Model + Reviewer x Condition + Reviewer x Submission x Condition	-38.638	11.186	-3738.950	35.840

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 (Δ 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.

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 fourteen characteristics and the condition (*Single vs Double Review Differential Item Functioning (SvB DIF) Model*) to a model that did not have these interaction terms (*Differential Item Functioning (NoDIF) Model)*) using leave-one-out cross validation. The difference between the SvB DIF ($elpd_{loo} = -3700.31; SE(elpd_{loo}) = 35.40$) and the simpler DIF model ($elpd_{loo} = -3693.62; SE(elpd_{loo}) = 35.05$) in terms of the expected log-pointwise predictive density was $\Delta elpd_{loo} = 6.7$ favoring the simpler DIF model without the interaction terms, but this difference was within the margin of error of $SE = 7.35$ indicating very weak evidence against an omnibus differential item or merit evaluation functioning between single and double-blind review.

Our main analysis entered all fourteen 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 EC.13). Moreover, a

model comparison between the SvB DIF model 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 EC.9).

EC.5.3. Regression Coefficients With All Characteristics Entered Simultaneously

As discussed in the text, to assess the degree to which review systems may advantage an individual or group, we entered review type (single vs. double), all fourteen author and submission characteristics, and the 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 main regression replaced the coauthor variables with variables for all authors (e.g., the proportion of male authors instead of male coauthors) for all the submissions so as to retain the single author submissions in the analysis. The regression coefficients for this regression are reported in Table EC.10. Because our approach to handling single author submissions introduced multicollinearity, we also reran the regression using multi-author submissions. Those coefficients are reported in Table EC.11.

Table EC.10 Coefficients when regressing single- and double-blind ratings onto author and submission characteristics simultaneously.

Category	Variable	Coefficient	Interaction Coefficient
	Intercept	-0.604 [-1.037, -0.165]	
	Condition: Single (0) or Double (1)	0.216 [-0.2, 0.627]	
First Author	Male vs Female	0.172 [-0.008, 0.352]	**0.225 [0.043, 0.408]**
	No Answer vs Female	-0.413 [-1.124, 0.3]	-0.24 [-0.963, 0.489]
	Years Since Ph.D.	-0.04 [-0.161, 0.081]	-0.029 [-0.147, 0.089]
	Domestic (0) or International (1)	0.066 [-0.072, 0.203]	-0.023 [-0.162, 0.117]
	Asian vs. White	-0.067 [-0.228, 0.095]	**0.166 [0.005, 0.328]**
	Black vs White	-0.099 [-1.051, 0.859]	-0.34 [-1.308, 0.634]
	Hispanic or Latino vs. White	0.072 [-0.337, 0.48]	-0.076 [-0.49, 0.337]
	Native American vs. White	0.345 [-0.985, 1.672]	-0.131 [-1.487, 1.227]
	Other vs. White	0.375 [-0.064, 0.812]	-0.141 [-0.59, 0.311]
	No Answer vs. White	0.081 [-0.148, 0.313]	-0.046 [-0.279, 0.186]
	SJDM Appearances	**0.104 [0.009, 0.198]**	0.012 [-0.083, 0.106]
	Prestige Score	**0.169 [0.044, 0.295]**	-0.055 [-0.182, 0.073]
	Economics Ph.D. vs Psychology	**0.169 [0.044, 0.295]**	-0.055 [-0.182, 0.073]
	Management Ph.D. vs Psychology	0.042 [-0.089, 0.173]	-0.032 [-0.164, 0.1]
	Marketing Ph.D. vs Psychology	0.137 [-0.149, 0.422]	0.083 [-0.208, 0.372]
	Other Ph.D. vs Psychology	0.081 [-0.146, 0.308]	0.078 [-0.149, 0.308]
	Undergraduate vs. Full	0.888 [-0.533, 2.319]	-1.291 [-2.769, 0.167]
	Masters vs. Full	0.336 [-0.31, 0.98]	-0.298 [-0.939, 0.34]
	Ph.D. vs. Full	**0.539 [0.07, 1.008]**	**-0.493 [-0.962, -0.025]**
	Post Doc vs. Full	**0.522 [0.072, 0.969]**	-0.34 [-0.788, 0.11]
	Research Scientist vs. Full	**0.631 [0.126, 1.139]**	**-0.578 [-1.094, -0.06]**
	Practitioner vs. Full	0.209 [-0.427, 0.847]	-0.212 [-0.842, 0.409]
	Assistant vs. Full	**0.519 [0.127, 0.908]**	-0.316 [-0.703, 0.078]
	Associate vs. Full	0.336 [-0.034, 0.707]	-0.196 [-0.571, 0.176]
Coauthors	Proportion Male Authors	-0.04 [-0.131, 0.052]	**-0.107 [-0.2, -0.016]**
	Ave. SJDM Appearances	**0.11 [0.028, 0.192]**	-0.01 [-0.094, 0.074]
	Ave. Prestige Score	0.042 [-0.089, 0.173]	-0.032 [-0.164, 0.1]
	Ave. Seniority	**0.146 [0.067, 0.226]**	**-0.088 [-0.168, -0.007]**
Submission	Empirical Studies vs Traditional	-0.034 [-0.186, 0.118]	0.019 [-0.131, 0.17]
	Psychology of JDM vs Traditional	0.01 [-0.191, 0.207]	-0.009 [-0.208, 0.193]
	Sentiment	-0.05 [-0.117, 0.016]	0.046 [-0.021, 0.112]
Reviewer	Male vs Female	-0.028 [-0.102, 0.045]	
	Years Since Ph.D.	-0.013 [-0.07, 0.043]	
	Prestige Score	0.023 [-0.015, 0.061]	
	Economics Ph.D. vs Psychology	0.035 [-0.084, 0.154]	
	Management Ph.D. vs Psychology	0.018 [-0.095, 0.128]	
	Marketing Ph.D. vs Psychology	-0.002 [-0.097, 0.092]	
	Other Ph.D. vs Psychology	-0.011 [-0.157, 0.134]	
	Research Scientist vs. Full	-0.084 [-0.266, 0.102]	
	Practitioner vs. Full	-0.079 [-0.452, 0.295]	
	Assistant vs. Full	-0.013 [-0.151, 0.125]	
	Associate vs. Full	0.006 [-0.109, 0.121]	
	Reviewer No. Times Review Past 7 Years	-0.018 [-0.057, 0.019]	

Regression based on $N = 440$ submissions. Bolded values indicate credible intervals exclude 0. All coauthor 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.

Table EC.11 Coefficients when regressing single- or double-blind review ratings onto author and submission characteristics simultaneously for multiauthor submissions.

Category	Variable	Coefficient	Interaction Coefficient
	Intercept	-0.459 [-0.882, -0.04]	
	Condition: Single (0) or Double (1)	0.146 [-0.262, 0.552]	
First Author	Male vs Female	**0.141 [0.006, 0.275]**	0.102 [-0.035, 0.238]
	No Answer vs Female	-0.075 [-1.05, 0.901]	0.461 [-0.538, 1.457]
	Years Since Ph.D.	-0.024 [-0.157, 0.11]	-0.046 [-0.176, 0.083]
	Domestic (0) or International (1)	0.092 [-0.05, 0.232]	-0.045 [-0.191, 0.1]
	Asian vs. White	-0.077 [-0.243, 0.086]	0.181 [0.013, 0.349]
	Black vs White	-0.09 [-1.03, 0.848]	-0.331 [-1.317, 0.654]
	Hispanic or Latino vs. White	0.069 [-0.334, 0.47]	-0.108 [-0.521, 0.308]
	Native American vs. White	0.317 [-1.003, 1.637]	-0.119 [-1.482, 1.238]
	Other vs. White	0.453 [-0.006, 0.906]	-0.13 [-0.597, 0.338]
	No Answer vs. White	0.081 [-0.151, 0.312]	-0.038 [-0.269, 0.196]
	SJDM Apperances	**0.137 [0.049, 0.224]**	0.02 [-0.069, 0.11]
	Prestige Score	**0.175 [0.081, 0.27]**	-0.06 [-0.157, 0.036]
	Economics Ph.D. vs Psychology	0.186 [-0.114, 0.49]	0.087 [-0.219, 0.394]
	Management Ph.D. vs Psychology	0.137 [-0.095, 0.367]	0.056 [-0.183, 0.299]
	Marketing Ph.D. vs Psychology	0.122 [-0.064, 0.307]	0.002 [-0.187, 0.189]
	Other Ph.D. vs Psychology	0.134 [-0.104, 0.372]	-0.073 [-0.314, 0.168]
	Undergraduate vs. Full	0.413 [-0.987, 1.813]	-1.04 [-2.488, 0.395]
	Masters vs. Full	0.049 [-0.579, 0.676]	-0.117 [-0.747, 0.51]
	Ph.D. vs. Full	0.312 [-0.136, 0.766]	-0.317 [-0.768, 0.137]
	Post Doc vs. Full	0.316 [-0.121, 0.753]	-0.22 [-0.659, 0.219]
	Research Scientist vs. Full	0.455 [-0.044, 0.959]	-0.462 [-0.973, 0.046]
	Practitioner vs. Full	0.179 [-0.506, 0.866]	-0.141 [-0.81, 0.526]
	Assistant vs. Full	**0.437 [0.044, 0.829]**	-0.264 [-0.657, 0.132]
	Associate vs. Full	0.339 [-0.039, 0.713]	-0.217 [-0.595, 0.163]
Coauthors	Proportion Male Coauthors	-0.009 [-0.078, 0.061]	**-0.072 [-0.142, -0.002]**
	Ave. SJDM Coauthor Appearances	**0.109 [0.033, 0.186]**	-0.016 [-0.093, 0.062]
	Ave. Coauthor Prestige Score	0.027 [-0.067, 0.123]	-0.027 [-0.122, 0.069]
	Ave. Coauthor Seniority	**0.117 [0.042, 0.193]**	**-0.104 [-0.179, -0.028]**
Submission	Empirical Studies vs Traditional	-0.031 [-0.189, 0.126]	0.07 [-0.087, 0.228]
	Psychology of JDM vs Traditional	0.007 [-0.199, 0.213]	0.064 [-0.144, 0.273]
	Sentiment	-0.044 [-0.112, 0.024]	0.028 [-0.041, 0.098]
Reviewer	Male vs Female	-0.028 [-0.105, 0.049]	
	Years Since Ph.D.	-0.019 [-0.076, 0.04]	
	Prestige Score	0.025 [-0.014, 0.064]	
	Economics Ph.D. vs Psychology	0.032 [-0.09, 0.154]	
	Management Ph.D. vs Psychology	0.016 [-0.099, 0.13]	
	Marketing Ph.D. vs Psychology	-0.004 [-0.102, 0.094]	
	Other Ph.D. vs Psychology	0.001 [-0.151, 0.153]	
	Research Scientist vs. Full	-0.087 [-0.281, 0.105]	
	Practitioner vs. Full	-0.072 [-0.461, 0.321]	
	Assistant vs. Full	-0.012 [-0.152, 0.13]	
	Associate vs. Full	0.001 [-0.117, 0.12]	
	Reviewer No. Times Review Past 7 Years	-0.018 [-0.057, 0.021]	

This regression is based on $N = 413$ submissions. As discussed in the text, the main analysis sought to include all submissions and therefore used all the authors when entering statistics for the co-authors. But, this introduces multicollinearity. Thus, we reran the models with multi-author submissions only. Bolded values indicated credible interval excludes 0. 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.

EC.5.4. Regression Coefficients with All Variables Entered Simultaneously Predicting Single or Double-Blind

To better understand our results, particularly the interaction terms with review conditions, we also regressed single-blind review ratings onto author and submission characteristic. We did the same for double-blind review ratings. These regressions are presented side-by-side in Table EC.12.

Table EC.12 Coefficients when regressing single- or double-blind review ratings onto author and submission characteristics simultaneously.

Category	Variable	Single ($N = 440$)	Double ($N = 440$)
	Intercept	-0.754 [-1.202, -0.303]	-0.289 [-0.745, 0.172]
First Author	Male vs Female	0.179 [-0.001, 0.361]	**0.383 [0.197, 0.567]**
	No Answer vs Female	-0.43 [-1.151, 0.277]	-0.637 [-1.359, 0.095]
	Years Since Ph.D.	-0.022 [-0.142, 0.098]	-0.064 [-0.186, 0.058]
	Domestic (0) or International (1)	0.063 [-0.075, 0.202]	0.045 [-0.097, 0.186]
	Asian vs. White	-0.07 [-0.234, 0.09]	0.101 [-0.068, 0.268]
	Black vs White	-0.113 [-1.067, 0.843]	-0.449 [-1.421, 0.521]
	Hispanic or Latino vs. White	0.075 [-0.329, 0.481]	0.016 [-0.397, 0.424]
	Native American vs. White	0.353 [-0.972, 1.685]	0.228 [-1.123, 1.573]
	Other vs. White	0.367 [-0.074, 0.808]	0.243 [-0.194, 0.683]
	No Answer vs. White	0.078 [-0.158, 0.308]	0.046 [-0.193, 0.283]
	SJDM Apperances	0.11 [0.016, 0.204]	**0.113 [0.016, 0.211]**
	Institutional Prestige	**0.166 [0.038, 0.291]**	0.116 [-0.012, 0.246]
	Economics Ph.D. vs Psychology	0.135 [-0.152, 0.418]	0.218 [-0.071, 0.509]
	Management Ph.D. vs Psychology	0.076 [-0.154, 0.304]	0.163 [-0.069, 0.394]
	Marketing Ph.D. vs Psychology	0.137 [-0.045, 0.32]	0.112 [-0.076, 0.298]
	Other Ph.D. vs Psychology	0.084 [-0.148, 0.315]	0.028 [-0.202, 0.262]
	Undergraduate vs. Full	0.947 [-0.483, 2.374]	-0.418 [-1.878, 1.028]
	Masters vs. Full	0.384 [-0.251, 1.025]	0.053 [-0.601, 0.718]
	Ph.D. vs. Full	**0.615 [0.149, 1.074]**	0.046 [-0.429, 0.525]
	Post Doc vs. Full	**0.581 [0.137, 1.024]**	0.19 [-0.267, 0.649]
	Research Scientist vs. Full	**0.669 [0.156, 1.177]**	0.062 [-0.465, 0.586]
	Practitioner vs. Full	0.257 [-0.377, 0.893]	-0.005 [-0.648, 0.638]
	Assistant vs. Full	**0.575 [0.187, 0.958]**	0.205 [-0.189, 0.6]
	Associate vs. Full	0.361 [-0.007, 0.734]	0.15 [-0.23, 0.532]
Authors	Proportion Male Authors	-0.045 [-0.137, 0.047]	**-0.145 [-0.237, -0.05]**
	Ave. SJDM Appearances	**0.109 [0.027, 0.193]**	**0.106 [0.022, 0.191]**
	Ave. Institutional Prestige	0.048 [-0.084, 0.181]	0.005 [-0.13, 0.141]
	Ave. Seniority	**0.151 [0.072, 0.231]**	0.056 [-0.026, 0.137]
	Empirical Studies vs Traditional	-0.04 [-0.195, 0.111]	-0.024 [-0.179, 0.132]
Submission	Psychology of JDM vs Traditional	0.007 [-0.192, 0.206]	-0.011 [-0.214, 0.192]
	Sentiment	-0.055 [-0.121, 0.012]	-0.006 [-0.073, 0.062]
	Male vs Female	-0.011 [-0.119, 0.099]	-0.03 [-0.152, 0.092]
Reviewer	Years Since Ph.D.	0.047 [-0.041, 0.134]	-0.076 [-0.158, 0.006]
	Prestige Score	0.021 [-0.033, 0.075]	0.016 [-0.05, 0.082]
	Economics Ph.D. vs Psychology	0.068 [-0.103, 0.239]	-0.036 [-0.224, 0.152]
	Management Ph.D. vs Psychology	0.02 [-0.146, 0.186]	0.039 [-0.127, 0.206]
	Marketing Ph.D. vs Psychology	0.039 [-0.094, 0.172]	-0.036 [-0.191, 0.119]
	Other Ph.D. vs Psychology	-	0.001 [-0.18, 0.181]
	Research Scientist vs. Full	0.013 [-0.25, 0.273]	-0.156 [-0.455, 0.147]
	Practitioner vs. Full	-	-0.08 [-0.507, 0.349]
	Assistant vs. Full	0.122 [-0.093, 0.336]	-0.137 [-0.35, 0.075]
	Associate vs. Full	0.068 [-0.104, 0.241]	-0.067 [-0.288, 0.153]
	Reviewer No. Times Review Past 7 Years	-0.034 [-0.096, 0.028]	0.001 [-0.064, 0.065]

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. The N indicates the number of submissions used in each regression.

EC.5.5. Regression Coefficients With Each Variable Entered Alone

The zero-order correlations shown in Figures EC.8 and EC.7 do not control for submission and reviewer variability. Thus, for each author and submission characteristic, we conducted a regression, entering each characteristic alone and predicting either a single- or double-blind review rating. The coefficients for these regressions are reported in Table EC.13. We also sought to understand if there was a difference between single- and double-blind review ratings for each characteristic alone (without accounting for each characteristic). Thus, we ran separate regressions for each characteristic, regressing the review rating onto the characteristic, review condition, and interaction. The coefficients for those regressions are reported in Table EC.14

Table EC.13 Coefficients when single- or double-blind review ratings separately onto each author and submission characteristic individually.

		Single Blind Coef (Alone)			Double Blind Coef (Alone)		
Category	Variable	M	95% CI	N	M	95% CI	N
First Author	Male vs. Female*	0.095	[-0.044, 0.232]	456	**0.163**	**[0.036, 0.291]**	456
	No Answer vs. Female	-0.357	[-0.992, 0.274]	”	-0.475	[-1.080, 0.120]	”
	Years Since Ph.D.	**-0.102**	**[-0.172, -0.032]**	442	-0.030	[-0.100, 0.037]	442
	International (0) or Domestic (1)	**0.172**	**[0.035, 0.308]**	456	**0.151**	**[0.020, 0.279]**	456
	Asian vs. White	-0.036	[-0.201, 0.130]	456	0.084	[-0.078, 0.240]	456
	Black vs. White	0.161	[-0.857, 1.197]	”	-0.152	[-1.171, 0.852]	456
	Hispanic or Latino vs. White	0.325	[-0.121, 0.750]	”	0.245	[-0.155, 0.667]	”
	Native American vs. White	0.095	[-1.373, 1.623]	”	0.014	[-1.374, 1.363]	”
	Other vs. White	0.242	[-0.253, 0.711]	”	0.152	[-0.287, 0.599]	”
	No Answer vs. White	-0.017	[-0.244, 0.212]	”	-0.058	[-0.273, 0.161]	”
	SJDM Appearances	**0.121**	**[0.045, 0.195]**	456	**0.158**	**[0.090, 0.229]**	456
	Institutional Prestige	**0.267**	**[0.199,0.336]**	456	**0.184**	**[116, 0.252]**	456
	Economics Ph.D. vs. Psychology	0.094	[-0.174, 0.367]	456	0.159	[-0.108, 0.422]	456
	Management Ph.D. vs Psychology	**0.244**	**[0.024, 0.474]**	”	**0.302**	**[0.083, 0.525]**	”
	Marketing Ph.D. vs. Psychology	**0.224**	**[0.050, 0.398]**	”	**0.192**	**[0.029, 0.355]**	”
	Other Ph.D. vs. Psychology	**0.363**	**[0.141, 0.580]**	”	**0.220**	**[0.006, 0.430]**	”
	Undergraduate vs. Full	0.218	[-1.265, 1.680]	454	-0.437	[-1.857, 0.978]	454
	Masters vs. Full*	-0.206	[-0.741, 0.351]	”	-0.196	[-0.737, 0.346]	”
	Ph.D. vs. Full*	**0.340**	**[0.070, 0.605]**	”	-0.007	[-0.269, 0.243]	”
	Post Doc vs. Full	0.262	[-0.055, 0.571]	”	0.070	[-0.236, 0.360]	”
	Research Scientist vs. Full	0.340	[-0.069, 0.766]	”	-0.083	[-0.499, 0.330]	”
	Practitioner vs. Full	-0.041	[-0.687, 0.573]	”	-0.175	[-0.767, 0.437]	”
	Assistant vs. Full	**0.366**	**[0.083, 0.639]**	”	0.162	[-0.107, 0.415]	”
	Associate vs. Full	0.049	[-0.301, 0.385]	”	-0.039	[-0.362, 0.286]	”
Authors & Co-Authors	Proportion Male Co-Authors*	0.041	[-0.030 0.109]	430	-0.035	[-0.102, 0.031]	430
	Proportion Male Authors*	0.063	[0.004 0.126]	456	0.026	[-0.034 0.087]	456
	Ave. SJDM Appearances of Co-Authors*	**0.200**	**[0.129, 0.266]**	430	**0.122**	**[0.055, 0.188]**	430
	Ave. SJDM Appearances of Authors*	**0.235**	**[0.169, 0.301]**	456	**0.172**	**[0.106, 0.236]**	456
	Ave. Institutional Prestige of Co-Authors*	**0.237**	**[0.170, 0.304]**	430	**0.156**	**[0.091, 0.221]**	430
	Ave. Institutional Prestige of Authors*	**0.273**	**[0.208, 0.337]**	456	**0.183**	**[0.116, 0.250]**	456
	Ave. Seniority of Co-Authors	**0.156**	**[0.086, 0.226]**	428	0.008	[-0.059. 0.076]	428
	Ave. Seniority of Authors	**0.111**	**[0.045, 0.178]**	456	0.063	[-0.003 0.129]	456
Text	Empirical Studies vs. Traditional Topic	0.014	[-0.146, 0.183]	456	0.048	[-0.097, 0.197]	456
	Psychology of JDM vs. Traditional Topic	-0.123	[-0.322, 0.082]	456	-0.091	[-0.285, 0.101]	456
	Sentiment	-0.012	[-0.079, 0.055]	456	0.019	[-0.046, 0.083]	456
Reviewer	Female (0) or Male (1)	-0.010	[-0.110, 0.091]	456	-0.009	[-0.110, 0.091]	456
	Years Since Ph.D.	-0.001	[-0.008, 0.006]	456	-0.003	[-0.011, 0.004]	456
	Institutional Prestige	0.009	[-0.040, 0.059]	456	0.024	[-0.023, 0.073]	456
	Economics Ph.D. vs. Psychology	0.042	[-0.112, 0.194]	456	0.029	[-0.125, 0.186]	456
	Management Ph.D. vs Psychology	-0.023	[-0.182, 0.134]	”	0.040	[-0.121, 0.201]	”
	Marketing Ph.D. vs. Psychology	0.002	[-0.127, 0.127]	”	-0.008	[-0.132, 0.120]	”
	Other Ph.D. vs. Psychology	-	-	”	0.021	[-0.129, 0.170]	”
	Research Scientist vs. Full	-0.077	[-0.297, 0.144]	456	-0.082	[-0.352, 0.193]	456
	Practitioner vs. Full	-	-	”	-0.034	[-0.290, 0.229]	”
	Assistant vs. Full	0.050	[-0.082, 0.178]	”	-0.031	[-0.148, 0.088]	”
	Associate vs. Full	0.028	[-0.082, 0.178]	”	-0.045	[-0.190, 0.102]	”
	Experience reviewing past 7 years	-0.012	[-0.041, 0.019]	456	0.002	[-0.024, 0.029]	456

Bolded values indicated credible interval excludes 0. Starred variables () indicate a credible interaction between the variable and the condition in Table EC.12. Review ratings and numerical predictors were standardized. The coefficients in the single- and double-blind were entered alone. The N indicates the number of submissions used in each regression.

Table EC.14 Coefficients when single and double-blind review ratings together onto submission characteristic, the review condition, and the interaction as an individual set without any other characteristic.

Category	Variable	Coefficient	Interaction Coefficient	N
First Author	Male vs Female	0.094 [-0.036, 0.225]	0.081 [-0.040 0.204]	456
	No Answer vs Female	-0.361 [-0.966, 0.261]	-0.131 [-0.677 0.429]	456
	Years Since Ph.D.	**-0.103 [-0.172, -0.034]**	**0.073 [0.012, 0.136]**	442
	Domestic (0) or International (1)	**0.173 [0.042, 0.305]**	-0.024 [-0.145, 0.096]	456
	Asian vs. White	-0.031 [-0.192, 0.125]	0.116 [-0.031, 0.264]	456
	Black vs White	0.172 [-0.828, 1.187]	-0.318 [-1.268, 0.623]	”
	Hispanic or Latino vs. White	0.320 [-0.104, 0.742]	-0.089 [-0.487, 0.316]	”
	Native American vs. White	0.101 [-1.332, 1.525]	-0.079 [-1.388, 1.260]	”
	Other vs. White	0.244 [-0.223, 0.704]	-0.100[-0.535, 0.341]	”
	No Answer vs. White	-0.010 [-0.234, 0.211]	-0.062 [-0.272, 0.143]	”
	SJDM Appearances	0.119 [0.047, 0.190]	0.043 [-0.025, 0.111]	456
	Institutional Prestige	**0.266 [0.198, 0.336]**	**-0.083 [-0.148, -0.018]**	456
	Economics Ph.D. vs Psychology	0.111 [-0.157, 0.375]	0.033 [-0.215, 0.279]	456
	Management Ph.D. vs Psychology	**0.243 [0.023, 0.462]**	-0.032 [-0.149, 0.267]	”
	Marketing Ph.D. vs Psychology	**0.221 [0.060, 0.384]**	0.083 [-0.188, 0.121]	”
	Other Ph.D. vs Psychology	**0.353 [0.140, 0.566]**	0.078 [-0.341, 0.060]	”
	Undergraduate vs. Full	0.247 [-1.160, 1.680]	-0.688 [-2.103, 0.680]	454
	Masters vs. Full	-0.184 [-0.728, 0.349]	-0.017 [-0.508, 0.479]	”
	Ph.D. vs. Full	**0.337[0.081, 0.597]**	**-0.338 [-0.573, -0.105]**	”
	Post Doc vs. Full	0.271 [-0.022, 0.576]	-0.211 [-0.484, 0.058]	”
	Research Scientist vs. Full	0.354 [-0.065, 0.777]	**-0.451[-0.835, -0.071]**	”
	Practitioner vs. Full	-0.036 [-0.625, 0.575]	-0.151 [-0.684, 0.375]	”
	Assistant vs. Full	0.373 [0.106, 0.650]	-0.209 [-0.454, 0.034]	”
	Associate vs. Full	0.059[-0.263, 0.389]	-0.102 [-0.403, 0.191]	”
Authors	Proportion Male Coauthors	**0.042 [-0.025, 0.108]**	**-0.075 [-0.139, -0.012]**	430
	Proportion Male Authors	-0.04 [-0.131, 0.052]	**-0.107 [-0.2, -0.016]**	456
	Ave. Co-author SJDM Appearances	**0.202 [0.136, 0.267]**	-0.084 [-0.146, -0.022]	430
	Ave. SJDM Appearances	**0.236 [0.172, 0.300]**	-0.065 [-0.127, -0.004]	456
	Ave. Co-author Institutional Prestige	**0.237 [0.174, 0.302]**	**-0.082 [-0.145, -0.020]**	430
	Ave. Institutional Prestige	**0.272 [0.210, 0.337]**	-0.089 [-0.149, -0.028]	456
	Ave. Co-author Seniority	**0.152 [0.084, 0.220]**	**-0.141 [-0.205, -0.078]**	428
	Ave. Author Seniority	**0.110 [0.047, 0.174]**	-0.047 [-0.106, 0.013]	456
Submission	Empirical Studies vs Traditional	0.020 [-0.137 0.173]	0.031 [-0.110, 0.174]	456
	Psychology of JDM vs Traditional	-0.116 [-0.313, 0.084]	0.033 [-0.152, 0.217]	456
	Sentiment	-0.010 [-0.074, 0.055]	0.026 [-0.035, 0.087]	456
Reviewer	Male vs Female	-0.024 [-0.116, 0.070]	0.035 [-0.095, 0.164]	456
	Years Since Ph.D.	-0.001 [-0.008, 0.007]	0.003 [-0.013, 0.007]	456
	Institutional Prestige	0.015 [-0.033, 0.062]	0.009 [-0.056, 0.076]	456
	Economics Ph.D. vs Psychology	0.074 [-0.084, 0.154]	-0.061 [-0.275, 0.159]	456
	Management Ph.D. vs Psychology	0.015 [-0.095, 0.128]	0.054 [-0.159, 0.159]	”
	Marketing Ph.D. vs Psychology	0.001 [-0.097, 0.092]	-0.019 [-0.193, 0.145]	”
	Other Ph.D. vs Psychology	-0.003 [-0.157, 0.134]	-	”
	Research Scientist vs. Full	-0.084 [-0.266, 0.102]	-0.078 [-0.400, 0.247]	456
	Practitioner vs. Full	-0.079 [-0.452, 0.295]	-	”
	Assistant vs. Full	-0.013 [-0.151, 0.125]	-0.069 [-0.238, 0.093]	”
	Associate vs. Full	0.006 [-0.109, 0.121]	-0.111 [-0.302, 0.079]	”
	Reviewer No. Times Review Past 7 Years	-0.009 [-0.039, 0.020]	-0.013 [-0.025, 0.050]	456

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). The N indicates the number of submissions used in each regression.

EC.6. 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.

EC.7. Predictive Validity of Review Ratings and Author and Submission Characteristics

This section reports supplemental results for the outcome variables we collected as well as supporting analyses for evaluating the predictive validity of review ratings and author and submission characteristics.

EC.7.1. Summary of Talk Outcomes Across sessions

Table EC.15 Mean (M) and standard deviation (SD) of talk ratings, attendance, and total potential questions

	Overall Rating (Std.)		Attendance		Potential Questions	
Session	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.

EC.7.2. Correlations Between Outcomes

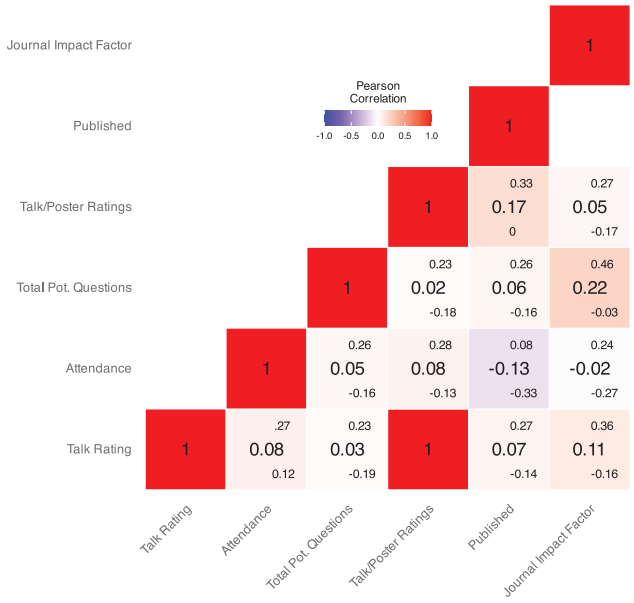

Figure EC.8 Correlations between outcome data. Note: The central value is the mean posterior estimate. The values in the upper and lower right-hand corners are the estimates of the upper and lower bounds of the 95% credible interval. Starred values indicate the credible interval excludes 0. Intercorrelations 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.

EC.7.3. Regression of Outcomes onto Single- and Double-blind Review Ratings

Table EC.16 Coefficients from regressing conference outcomes onto single- and double-blind review ratings simultaneously.

		Single		Double	
	N	M	95% CI	M	95% CI
Talk Rating (z)	107	0.048	[-0.092, 0.190]	-0.018	[-0.139, 0.106]
Attendance (z)	107	-0.119	[-0.438, 0.207]	-0.117	[-0.397, 0.166]
Total Potential Questions (z)	105	0.017	[-0.326, 0.368]	0.164	[-0.158, 0.495]
Talk & Poster Rating (z)	163	**0.326**	**[0.112, 0.536]**	0.191	[-0.022, 0.400]
Poster Rating (z)	57	**0.177**	**[0.058, 0.292]**	0.163	**[0.033 0.291]**
Published	530	0.184	[-0.049, 0.410]	**0.368**	**[0.132, 0.604]**
Journal Impact Factor (z)	173	**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.

EC.7.4. Regression of Outcomes Onto Author and Submission Characteristics

We collected several relevant outcomes including an overall rating of those submissions that were presented as a talk, the attendance at talks, number of questions asked, ratings of submissions that were presented to posters and rated in the student competition, if the submission was published, and the impact factor of the journal the submission was ultimately published in. For each outcome, we regressed the outcome variable on the the author and submission characteristics simultaneously. This regression evaluated how well the characteristics predict the outcome. We report the regressions using all the submissions with standardized outcome values(Table EC.17), the same regression with the nonstandardized outcome variable (when relevant) (Table EC.20), and the corresponding regressions for multi-author submissions to address multicollinearity (Tables EC.20 and EC.20).

Table EC.17 Coefficients from regressing standardized conference outcomes onto submission characteristics.

Category	Variable	Attendance (z) (N = 84)	Overall (z) (N = 84)	Questions (z) (N = 84)	Poster Rating (z) (N = 44)	Talk/Poster Rating (z) (N = 127)	Publication (N = 440)
	Intercept	-1.067, [-2.628, 0.52]	-0.405, [-1.162, 0.337]	0.768, [-1.135, 2.71]	-1.366, [-2.935, 0.248]	-1.566, [-3.28, 0.147]	0.694, [-0.675, 2.054]
	Male vs Female	-0.255, [-0.859, 0.349]	0.138, [-0.151, 0.431]	-0.123, [-0.875, 0.637]	0.532, [-0.01, 1.081]	0.443, [-0.079, 0.984]	0.235, [-0.359, 0.827]
First Author	No Answer vs Female	2.168, [-0.058, 4.455]	**1.528, [0.447, 2.62]**	2.61, [-0.365, 5.6]		**2.668, [0.146, 5.243]**	**-22.123, [-60.462, -2.373]**
	Years Since Ph.D.	0.682, [0.103, 1.258]	-0.041, [-0.313, 0.235]	-0.208, [-0.919, 0.505]	0.154, [-1.085, 1.418]	-0.035, [-0.663, 0.609]	-0.249, [-0.645, 0.148]
	Domestic (0) or International (1)	**0.536, [0.069, 1.007]**	0.035, [-0.195, 0.269]	-0.099, [-0.719, 0.525]	0.355, [-0.108, 0.819]	0.082, [-0.344, 0.503]	0.124, [-0.332, 0.574]
	Asian vs. White	-0.444, [-1.024, 0.143]	0.139, [-0.148, 0.421]	-0.256, [-1.012, 0.5]	0.196, [-0.194, 0.591]	0.273, [-0.214, 0.758]	-0.211, [-0.751, 0.319]
	Black vs White	0.511, [-0.446, 1.466]	0.283, [-0.187, 0.753]	1.224, [-0.004, 2.434]			-0.611, [-4.359, 3.074]
	Hispanic or Latino vs. White	1.01, [-0.825, 2.903]	0.311, [-0.575, 1.198]	-0.106, [-2.435, 2.238]		0.727, [-0.379, 1.846]	1.25, [-0.183, 2.892]
	Native American vs. White					0.561, [-1.627, 2.735]	**-42.501, [-118.378, -1.844]**
	Other vs. White	0.237, [-0.906, 1.376]	0.064, [-0.5, 0.616]	-0.631, [-2.109, 0.849]		0.21, [-1.071, 1.501]	0.683, [-0.795, 2.243]
	No Answer vs. White	-0.273, [-1.16, 0.62]	-0.156, [-0.588, 0.281]	-0.619, [-1.747, 0.536]	0.972, [-0.19, 2.143]	0.179, [-0.704, 1.065]	-0.358, [-1.137, 0.411]
	SJDM Apperances	0.144, [-0.106, 0.4]	0.069, [-0.054, 0.193]	-0.051, [-0.37, 0.278]	-0.007, [-0.585, 0.585]	0.138, [-0.131, 0.397]	-0.18, [-0.462, 0.101]
	Prestige Score	0.179, [-0.176, 0.532]	0.152, [-0.019, 0.32]	0.257, [-0.199, 0.705]	0.273, [-0.212, 0.755]	**0.423, [0.043, 0.798]**	0.24, [-0.153, 0.635]
	Economics Ph.D. vs Psychology	0.06, [-0.945, 1.044]	0.453, [-0.027, 0.932]	0.232, [-1.038, 1.48]	-0.228, [-1.657, 1.206]	0.875, [-0.136, 1.877]	-0.282, [-1.234, 0.685]
	Management Ph.D. vs Psychology	0.277, [-0.446, 0.996]	-0.01, [-0.365, 0.341]	-0.397, [-1.286, 0.522]	-0.459, [-1.04, 0.128]	-0.413, [-1.075, 0.241]	-0.359, [-1.077, 0.383]
	Marketing Ph.D. vs Psychology	0.036, [-0.538, 0.609]	0.048, [-0.235, 0.328]	0.53, [-0.207, 1.272]	0.198, [-0.28, 0.675]	0.175, [-0.356, 0.704]	-0.838, [-1.435, -0.245]
	Other Ph.D. vs Psychology	0.234, [-0.415, 0.882]	-0.001, [-0.32, 0.318]	0.424, [-0.439, 1.288]	0.131, [-0.582, 0.837]	-0.04, [-0.728, 0.651]	-0.651, [-1.427, 0.108]
	Undergraduate vs. Full					-1.831, [-4.904, 1.181]	**-41.366, [-118.333, -0.478]**
	Masters vs. Full	1.138, [-1.011, 3.271]	-0.048, [-1.059, 0.961]	-0.52, [-3.156, 2.164]	0.881, [-0.488, 2.215]	-0.747, [-2.873, 1.355]	-1.884, [-4.385, 0.389]
	Ph.D. vs. Full	0.235, [-1.65, 2.117]	-0.086, [-0.973, 0.826]	-0.947, [-3.271, 1.366]	1.134, [-0.308, 2.543]	0.123, [-1.852, 2.079]	-0.39, [-1.92, 1.143]
	Post Doc vs. Full	0.874, [-0.875, 2.615]	-0.007, [-0.831, 0.828]	-0.837, [-3.032, 1.368]		0.343, [-1.526, 2.206]	-0.104, [-1.604, 1.356]
	Research Scientist vs. Full	0.357, [-1.31, 2.009]	0.432, [-0.359, 1.234]	-0.969, [-3.03, 1.103]		0.907, [-0.998, 2.842]	-1.613, [-3.445, 0.102]
	Practitioner vs. Full	1.357, [-1.037, 3.744]	0.152, [-0.982, 1.294]	0.691, [-2.356, 3.707]		0.573, [-2.113, 3.287]	-1.294, [-3.852, 1.055]
	Assistant vs. Full	1.114, [-0.352, 2.59]	0.079, [-0.607, 0.776]	-0.624, [-2.415, 1.158]	-0.625, [-3.295, 2.138]	0.502, [-1.054, 2.062]	0.034, [-1.249, 1.295]
	Associate vs. Full	-0.183, [-1.454, 1.087]	-0.268, [-0.862, 0.336]	-1.071, [-2.678, 0.538]		-0.474, [-1.903, 0.97]	-0.556, [-1.756, 0.611]
Authors	Proportion Male Authors	0.013, [-0.304, 0.33]	-0.041, [-0.193, 0.113]	0.349, [-0.057, 0.768]	-0.335, [-0.596, -0.073]	-0.243, [-0.517, 0.028]	0.07, [-0.231, 0.368]
	Ave. SJDM Appearances	0.135, [-0.132, 0.406]	-0.001, [-0.133, 0.128]	-0.164, [-0.507, 0.171]	0.078, [-0.135, 0.291]	0.097, [-0.138, 0.329]	0.072, [-0.198, 0.344]
	Ave. Prestige Score	0.065, [-0.351, 0.482]	-0.042, [-0.243, 0.16]	0.036, [-0.494, 0.559]	-0.353, [-0.882, 0.183]	-0.228, [-0.661, 0.199]	-0.279, [-0.714, 0.157]
	Ave. Seniority	-0.417, [-0.768, -0.075]	0.056, [-0.107, 0.222]	-0.14, [-0.571, 0.293]	0.293, [0.024, 0.563]	0.147, [-0.144, 0.438]	**0.278, [0.017, 0.551]**
Submission	Empirical Studies vs Traditional	0.091, [-0.47, 0.654]	0.094, [-0.177, 0.368]	0.439, [-0.32, 1.186]	0.142, [-0.192, 0.473]	0.311, [-0.146, 0.766]	-0.085, [-0.596, 0.418]
	Psychology of JDM vs Traditional	0.481, [-0.232, 1.204]	0.036, [-0.317, 0.391]	-0.188, [-1.132, 0.733]	0.119, [-0.403, 0.634]	0.303, [-0.329, 0.93]	**-0.665, [-0.325, -0.008]**
	Sentiment	-0.018, [-0.256, 0.217]	-0.055, [-0.172, 0.058]	0.097, [-0.2, 0.387]	-0.084, [-0.273, 0.104]	-0.109, [-0.318, 0.1]	-0.066, [-0.289, 0.156]
Session	Session2	1.052, [0.201, 1.907]	0.038, [-0.381, 0.452]	-0.16, [-1.24, 0.926]		0.113, [-0.886, 1.114]	
	Session3	0.189, [-0.721, 1.089]	-0.078, [-0.522, 0.357]	-0.511, [-1.67, 0.644]		-0.084, [-1.128, 0.955]	
	Session4	0.411, [-0.379, 1.206]	0.258, [-0.124, 0.641]	-0.518, [-1.506, 0.458]		0.585, [-0.327, 1.492]	
	Session5	0.175, [-0.611, 1]	0.213, [-0.175, 0.6]	0.03, [-1.044, 1.128]		0.676, [-0.279, 1.621]	
	Session6	0.637, [-0.226, 1.483]	0.073, [-0.329, 0.489]	-0.524, [-1.606, 0.556]		0.277, [-0.695, 1.265]	
	Session7	-0.224, [-1.074, 0.623]	0.019, [-0.395, 0.433]	-0.386, [-1.453, 0.679]		0.306, [-0.734, 1.373]	
	Session8	-0.872, [-1.727, -0.025]	0.263, [-0.139, 0.667]	-0.28, [-1.337, 0.763]		0.665, [-0.315, 1.65]	
	Session9	-0.474, [-1.303, 0.353]	0.274, [-0.136, 0.673]	-0.918, [-1.98, 0.157]		0.798, [-0.139, 1.722]	
	Poster Session					1.041, [0.252, 1.841]	
	Presentation						0.291, [-0.273, 0.863]
Noise	sigma	0.779, [0.638, 0.96]	0.376, [0.307, 0.464]	0.978, [0.796, 1.214]	0.434, [0.326, 0.589]	0.957, [0.828, 1.111]	

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 EC.19 provides the regression coefficients when subsetting the data on multiple author submissions.

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

Category	Variable	Attendance	Questions
	Intercept	115.4, [17.2, 214.1]	4.3, [-1.5, 10.1]
	Male vs Female	-8.7, [-30.4, 12.4]	-0.2, [-1.5, 1.1]
First Author	No Answer vs Female	75.9, [-1.3, 156.2]	4.3, [-0.6, 9.3]
	Years Since Ph.D.	**2.9, [0.5, 5.3]**	0, [-0.2, 0.1]
	Domestic (0) or International (1)	**18.8, [2.1, 35.3]**	-0.2, [-1.2, 0.9]
	Asian vs. White	-15.4, [-35.7, 5]	-0.4, [-1.7, 0.8]
	Black vs White	17.5, [-15.3, 50.3]	**2, [0, 4]**
	Hispanic or Latino vs. White	34.8, [-29.3, 98.9]	-0.2, [-4, 3.7]
	Native American vs. White	-	-
	Other vs. White	8.1, [-32.3, 48.7]	-1, [-3.4, 1.4]
	No Answer vs. White	-9.6, [-41.2, 22.3]	-1, [-2.9, 0.9]
	SJDM Apperances	1.6, [-1.2, 4.3]	0, [-0.2, 0.1]
	Prestige Score	4.2, [-4.2, 12.5]	0.3, [-0.2, 0.8]
	Economics Ph.D. vs Psychology	2.2, [-32.6, 36.6]	0.4, [-1.7, 2.5]
	Management Ph.D. vs Psychology	9.6, [-16.1, 35.6]	-0.7, [-2.1, 0.9]
	Marketing Ph.D. vs Psychology	1.4, [-18.4, 21.6]	0.9, [-0.3, 2.1]
	Other Ph.D. vs Psychology	8.4, [-14.4, 30.8]	0.7, [-0.7, 2.1]
	Undergraduate vs. Full	-	-
	Masters vs. Full	39.7, [-33.2, 114.3]	-0.9, [-5.2, 3.6]
	Ph.D. vs. Full	8.3, [-55.5, 73.4]	-1.6, [-5.4, 2.4]
	Post Doc vs. Full	30.5, [-30.3, 91.2]	-1.4, [-5, 2.3]
	Research Scientist vs. Full	12.8, [-43.7, 70.3]	-1.6, [-4.9, 1.8]
	Practitioner vs. Full	47.4, [-36.1, 130.3]	1.2, [-3.7, 6.2]
	Assistant vs. Full	38.8, [-11.8, 90.4]	-1, [-4, 2]
	Associate vs. Full	-6.2, [-48.7, 37.6]	-1.8, [-4.4, 0.9]
Authors	Proportion Male Authors	1.4, [-31.8, 34.4]	1.7, [-0.3, 3.7]
	Ave. SJDM Appearances	4.2, [-21.4, 29]	0.1, [-1.4, 1.6]
	Ave. Prestige Score	2.9, [-2.9, 8.7]	-0.2, [-0.5, 0.2]
	Ave. Seniority	**-11.8, [-21.5, -2]**	-0.2, [-0.8, 0.4]
Submission	Empirical Studies vs Traditional	3.2, [-16.6, 23.1]	0.7, [-0.5, 1.9]
	Psychology of JDM vs Traditional	16.9, [-8.1, 41.6]	-0.3, [-1.9, 1.2]
	Sentiment	-2.1, [-29.8, 26.3]	0.5, [-1.1, 2.2]
Session	Session2	**36.9, [6.2, 67.1]**	-0.3, [-2.1, 1.5]
	Session3	6.5, [-24.7, 37.7]	-0.8, [-2.7, 1]
	Session4	14.3, [-13.3, 41.5]	-0.9, [-2.5, 0.8]
	Session5	6.1, [-21.7, 34.1]	0, [-1.7, 1.8]
	Session6	22.2, [-7.6, 52.1]	-0.9, [-2.6, 0.9]
	Session7	-7.9, [-37.9, 21.9]	-0.7, [-2.4, 1.1]
	Session8	**-30.4, [-59.3, -1.2]**	-0.5, [-2.2, 1.2]
	Session9	-16.6, [-45.7, 13.1]	-1.5, [-3.2, 0.2]
	Poster Session	-	-
	Presentation	-	-
Noise	sigma	27.1, [22.2, 33.5]	1.6, [1.3, 2]

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.

Table EC.19 Coefficients from regressing standardized conference outcomes onto submission characteristics for multiple author papers.

Category	Variable	Attendance (z) (N = 83)	Overall (z) (N = 83)	Questions (z) (N = 83)	Poster Rating (z) (N = 42)	Talk/Poster Rating (z) N = 124)	Publication (N = 413)
	Intercept	-1.84, [-3.266, -0.423]	-0.348, [-1.065, 0.375]	0.476, [-1.368, 2.353]	-1.735, [-3.266, -0.165]	-1.276, [-2.931, 0.398]	1.193, [-0.175, 2.566]
	Male vs Female	-0.281, [-0.726, 0.159]	0.067, [-0.156, 0.289]	0.273, [-0.312, 0.848]	0.175, [-0.187, 0.535]	0.113, [-0.295, 0.517]	0.302, [-0.144, 0.753]
First Author	No Answer vs Female	1.626, [-0.641, 3.89]	**1.466, [0.302, 2.62]**	2.6, [-0.398, 5.639]		**3.07, [0.528, 5.585]**	-29.19, [-78.699, -2.24]
	Years Since Ph.D.	0.684, [0.127, 1.221]	-0.062, [-0.335, 0.213]	-0.177, [-0.888, 0.539]	-0.071, [-1.293, 1.166]	-0.106, [-0.738, 0.515]	-0.437, [-0.917, 0.015]
	Domestic (0) or International (1)	**0.498, [0.033, 0.965]**	0.028, [-0.21, 0.266]	-0.091, [-0.707, 0.53]	0.266, [-0.159, 0.696]	0.172, [-0.25, 0.59]	0.056, [-0.43, 0.538]
	Asian vs. White	-0.381, [-0.958, 0.198]	0.171, [-0.121, 0.464]	-0.278, [-1.046, 0.465]	0.3, [-0.088, 0.693]	0.408, [-0.078, 0.9]	-0.218, [-0.776, 0.335]
	Black vs White	0.503, [-0.41, 1.418]	0.228, [-0.237, 0.696]	**1.372, [0.177, 2.598]**			-0.666, [-4.51, 3.188]
	Hispanic or Latino vs. White	0.982, [-0.803, 2.772]	0.205, [-0.693, 1.105]	-0.035, [-2.388, 2.297]		0.607, [-0.5, 1.713]	1.177, [-0.307, 2.824]
	Native American vs. White					0.354, [-1.733, 2.497]	**-40.062, [-112.352, -2.228**
	Other vs. White	0.137, [-0.975, 1.257]	0.031, [-0.529, 0.599]	-0.662, [-2.137, 0.806]		0.259, [-1.025, 1.526]	0.847, [-0.671, 2.544]
	No Answer vs. White	-0.217, [-1.104, 0.663]	-0.119, [-0.573, 0.322]	-0.686, [-1.841, 0.475]	-0.039, [-1.397, 1.294]	0.09, [-0.765, 0.948]	-0.312, [-1.09, 0.461]
	SJDM Apperances	**0.246, [0.02, 0.474]**	0.089, [-0.026, 0.201]	-0.121, [-0.415, 0.172]	-0.312, [-0.863, 0.241]	0.173, [-0.075, 0.415]	-0.151, [-0.424, 0.125]
	Prestige Score	0.156, [-0.13, 0.439]	0.096, [-0.046, 0.238]	0.291, [-0.077, 0.66]	0.44, [0.016, 0.854]	**0.326, [0.05, 0.601]**	0.093, [-0.209, 0.399]
	Economics Ph.D. vs Psychology	0.099, [-0.892, 1.08]	**0.488, [0.001, 0.982]**	0.198, [-1.082, 1.47]	-0.79, [-2.383, 0.75]	0.767, [-0.226, 1.751]	-0.465, [-1.473, 0.564]
	Management Ph.D. vs Psychology	0.261, [-0.464, 0.961]	-0.014, [-0.369, 0.335]	-0.368, [-1.279, 0.562]	-0.627, [-1.209, -0.04]	-0.399, [-1.036, 0.226]	-0.434, [-1.208, 0.309]
	Marketing Ph.D. vs Psychology	0.04, [-0.525, 0.617]	0.045, [-0.237, 0.33]	0.492, [-0.248, 1.223]	0.055, [-0.43, 0.535]	0.128, [-0.379, 0.636]	-0.955, [-1.594, -0.337]
	Other Ph.D. vs Psychology	0.294, [-0.356, 0.938]	0.051, [-0.274, 0.366]	0.44, [-0.421, 1.313]	-0.194, [-0.921, 0.53]	0.002, [-0.689, 0.679]	-0.658, [-1.451, 0.119]
	Undergraduate vs. Full					-2.754, [-5.625, 0.068]	**-40.474, [-114.193, -1.326**
	Masters vs. Full	**2.137, [0.183, 4.085]**	-0.109, [-1.075, 0.864]	-0.244, [-2.743, 2.309]	0.963, [-0.352, 2.229]	-1.247, [-3.269, 0.724]	-2.68, [-5.19, -0.454]
	Ph.D. vs. Full	1.311, [-0.333, 2.956]	-0.127, [-0.942, 0.688]	-0.684, [-2.842, 1.453]	1.218, [-0.168, 2.563]	-0.304, [-2.228, 1.558]	-1.098, [-2.645, 0.401]
	Post Doc vs. Full	**1.921, [0.273, 3.52]**	-0.017, [-0.823, 0.776]	-0.681, [-2.792, 1.41]		0.004, [-1.87, 1.833]	-0.753, [-2.218, 0.701]
	Research Scientist vs. Full	1.164, [-0.422, 2.743]	0.43, [-0.371, 1.228]	-0.896, [-2.952, 1.181]		0.9, [-1.016, 2.764]	-2.119, [-3.938, -0.418]
	Practitioner vs. Full	2.154, [-0.185, 4.491]	0.213, [-0.945, 1.377]	0.74, [-2.265, 3.761]		0.329, [-2.337, 2.972]	-1.526, [-4.144, 1.005]
	Assistant vs. Full	**1.645, [0.276, 3.017]**	0.083, [-0.61, 0.776]	-0.599, [-2.454, 1.21]	1.349, [-0.967, 3.71]	0.441, [-1.137, 2.017]	-0.075, [-1.375, 1.219]
	Associate vs. Full	0.203, [-1.038, 1.421]	-0.238, [-0.843, 0.373]	-1.052, [-2.638, 0.558]		-0.616, [-2.059, 0.819]	-0.68, [-1.916, 0.52]
Authors	Proportion Male Coauthors	-0.076, [-0.313, 0.163]	-0.028, [-0.146, 0.09]	0.284, [-0.02, 0.589]	-0.234, [-0.416, -0.046]	-0.18, [-0.391, 0.028]	0.057, [-0.179, 0.289]
	Ave. SJDM Coauthor Appearances	0.119, [-0.128, 0.363]	-0.003, [-0.125, 0.117]	-0.182, [-0.51, 0.135]	0.154, [-0.035, 0.339]	0.096, [-0.116, 0.312]	0.078, [-0.178, 0.34]
	Ave. Coauthor Prestige Score	-0.004, [-0.298, 0.291]	-0.027, [-0.172, 0.118]	0.01, [-0.369, 0.393]	-0.419, [-0.847, 0.012]	-0.144, [-0.45, 0.161]	-0.194, [-0.508, 0.121]
	Ave. Coauthor Seniority	**-0.359, [-0.616, -0.102]**	0.01, [-0.117, 0.139]	-0.06, [-0.401, 0.274]	0.029, [-0.22, 0.271]	-0.049, [-0.292, 0.194]	0.145, [-0.094, 0.387]
Submission	Empirical Studies vs Traditional	-0.01, [-0.576, 0.555]	0.054, [-0.227, 0.339]	0.397, [-0.331, 1.142]	0.282, [-0.064, 0.63]	0.315, [-0.151, 0.786]	-0.124, [-0.652, 0.405]
	Psychology of JDM vs Traditional	0.319, [-0.41, 1.034]	0.005, [-0.351, 0.354]	-0.29, [-1.206, 0.63]	0.073, [-0.444, 0.589]	0.25, [-0.397, 0.893]	-0.617, [-1.295, 0.061]
	Sentiment	0.015, [-0.215, 0.248]	-0.058, [-0.175, 0.057]	0.088, [-0.211, 0.388]	-0.033, [-0.206, 0.137]	-0.108, [-0.316, 0.1]	-0.099, [-0.332, 0.133]
Session	Session2	**1.206, [0.369, 2.053]**	0.109, [-0.322, 0.528]	-0.206, [-1.278, 0.876]		0.308, [-0.698, 1.291]	
	Session3	0.463, [-0.452, 1.365]	-0.024, [-0.481, 0.438]	-0.518, [-1.711, 0.677]		0.01, [-1.037, 1.057]	
	Session4	0.443, [-0.316, 1.214]	0.268, [-0.114, 0.654]	-0.539, [-1.544, 0.456]		0.722, [-0.203, 1.647]	
	Session5	0.498, [-0.321, 1.324]	0.298, [-0.118, 0.712]	-0.028, [-1.116, 1.062]		0.856, [-0.131, 1.83]	
	Session6	**0.85, [0.015, 1.679]**	0.128, [-0.29, 0.542]	-0.545, [-1.647, 0.545]		0.478, [-0.521, 1.468]	
	Session7	-0.133, [-0.956, 0.696]	0.046, [-0.374, 0.468]	-0.434, [-1.508, 0.659]		0.25, [-0.778, 1.278]	
	Session8	-0.698, [-1.506, 0.106]	0.284, [-0.122, 0.689]	-0.288, [-1.334, 0.768]		0.791, [-0.194, 1.782]	
	Session9	-0.344, [-1.149, 0.474]	0.318, [-0.092, 0.729]	-0.946, [-1.994, 0.116]		**0.985, [0.058, 1.907]**	
	Poster Session					**1.258, [0.431, 2.065]**	
	Presentation						0.327, [-0.229, 0.898]
Noise	sigma	0.752, [0.617, 0.928]	0.376, [0.307, 0.464]	0.978, [0.799, 1.211]	0.411, [0.306, 0.565]	0.932, [0.801, 1.091]	

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.

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

Category	Variable	Attendance	Questions
	Intercept	65.1, [-2.4, 134.6]	3.3, [-0.9, 7.5]
	Male vs Female	-9.9, [-25.5, 5.9]	0.5, [-0.5, 1.4]
First Author	No Answer vs Female	55.9, [-23.9, 136.8]	4.4, [-0.5, 9.3]
	Years Since Ph.D.	**2.9, [0.5, 5.2]**	0, [-0.2, 0.1]
	Domestic (0) or International (1)	**17.3, [0.6, 33.3]**	-0.1, [-1.2, 0.9]
	Asian vs. White	-13.3, [-33.4, 6.8]	-0.5, [-1.7, 0.8]
	Black vs White	17.4, [-15.1, 49.8]	2.3, [0.3, 4.2]
	Hispanic or Latino vs. White	34, [-27.5, 95.4]	0, [-3.9, 3.8]
	Native American vs. White	-	-
	Other vs. White	4.9, [-34.4, 44.7]	-1.1, [-3.6, 1.4]
	No Answer vs. White	-7.5, [-38.2, 23.6]	-1.1, [-3, 0.7]
	SJDM Apperances	2.7, [0.2, 5.1]	-0.1, [-0.2, 0.1]
	Prestige Score	3.7, [-3.1, 10.4]	0.3, [-0.1, 0.7]
	Economics Ph.D. vs Psychology	3.6, [-30.7, 37.4]	0.3, [-1.8, 2.5]
	Management Ph.D. vs Psychology	9.3, [-15.1, 33.9]	-0.6, [-2.1, 0.9]
	Marketing Ph.D. vs Psychology	1.3, [-18.7, 20.7]	0.8, [-0.4, 2]
	Other Ph.D. vs Psychology	10.4, [-11.9, 33.2]	0.7, [-0.7, 2.2]
	Undergraduate vs. Full	-	-
	Masters vs. Full	**73.9, [5, 141.9]**	-0.4, [-4.5, 3.8]
	Ph.D. vs. Full	45.4, [-13.1, 102.5]	-1.1, [-4.7, 2.4]
	Post Doc vs. Full	**66.6, [9.5, 122.6]**	-1.1, [-4.6, 2.3]
	Research Scientist vs. Full	40.1, [-16.3, 95.6]	-1.4, [-4.8, 2]
	Practitioner vs. Full	74.8, [-7.9, 156]	1.2, [-3.8, 6.2]
	Assistant vs. Full	**56.9, [7.4, 105.9]**	-1, [-4, 2]
	Associate vs. Full	6.8, [-36.4, 50.5]	-1.7, [-4.4, 1]
Coauthors	Proportion Male Coauthors	-6.6, [-26.7, 13.8]	1.2, [-0.1, 2.4]
	Ave. SJDM Coauthor Appearances	2.3, [-2.4, 7.1]	-0.2, [-0.5, 0.1]
	Ave. Coauthor Prestige Score	-0.3, [-20.5, 19.6]	0.1, [-1.2, 1.3]
	Ave. Coauthor Seniority	-6.9, [-11.8, -1.9]	-0.1, [-0.4, 0.2]
Submission	Empirical Studies vs Traditional	-0.3, [-20.1, 19.6]	0.6, [-0.6, 1.9]
	Psychology of JDM vs Traditional	11.1, [-13.3, 36.2]	-0.5, [-2, 1.1]
	Sentiment	1.6, [-25.4, 28.5]	0.5, [-1.2, 2.1]
Session	Session2	**41.8, [12.3, 71.4]**	-0.3, [-2.1, 1.5]
	Session3	16.2, [-15.8, 47.8]	-0.8, [-2.8, 1.1]
	Session4	15.4, [-11.6, 42.5]	-0.9, [-2.5, 0.8]
	Session5	17.3, [-11.5, 46.6]	0, [-1.8, 1.8]
	Session6	29.6, [-0.2, 59]	-0.9, [-2.7, 0.9]
	Session7	-4.5, [-32.9, 25.2]	-0.7, [-2.5, 1.1]
	Session8	-24.4, [-53.4, 4.4]	-0.5, [-2.2, 1.3]
	Session9	-11.9, [-41.1, 17.3]	-1.5, [-3.3, 0.2]
	Poster Session	-	-
	Presentation	-	-
Noise	sigma	26.2, [21.4, 32.6]	1.6, [1.3, 2]

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 DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022. https://doi.org/10.1162/jmlr.2003.3.45.993.

Goodrich B, Gabry J, Ali I, Brilleman S (2020) rstanarm: Bayesian applied regression modeling via Stan version 2.32.1. https://mc-stan.org/rstanarm.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101(suppl 1):5228–5235. https://doi.org/10.1073/pnas.0307752101.

Hu M, Liu B (2004) Mining opinion features in customer reviews. *Proc. of the 19th National Conference on Artificial Intelligence (AAAI-04)*, pp. 755–760. Available at: https://www.cs.uic.edu/~liub/aaai04-featureExtract.pdf.

Morey RD, Rouder JN (2018) BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-4.7. https://cran.r-project.org/package=BayesFactor.

R Core Team (2019) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria). https://www.R-project.org/.