Referee Recommendations

Ivo Welch
University of California at Los Angeles
Anderson Graduate School of Management

This paper quantitatively analyzes referee recommendations at eight prominent economics and finance journals, and the SFS (Society for Financial Studies) Cavalcade Conference, where a known algorithm matched referees to submissions. The behavior of referees was similar in all venues. The referee-specific component in the disposition recommendation was about twice as important as the common component. Referees differed both in their scales (some referees were intrinsically more generous than others) and in their opinions of what a good paper was (they often disagreed about the relative ordering of papers). (*JEL* A14)

The editorial process determines not only the evolution of economics and finance but also the incentives and professional fates of academic economists. Yet, its participants do not have much objective knowledge about the process. Authors write only a few papers per year and typically receive only a few referee reports per submission. They usually do not learn which other papers were rejected. They rarely find out why an editor chose a particular referee, much less who the referee was. In turn, they referee only a few papers themselves every year and rarely receive feedback about how their views lined up with those of other referees.

The heterogeneity among referee evaluations is further exacerbated by the fact that the journals themselves have also not explicitly stated their objectives and criteria other than in broad and uncontroversial terms. For example, some referees hold the view that only the submitted paper should influence editorial decisions and that fairness to authors is a main goal. Others hold the view that journals should select submissions to maximize their impact, allowing such factors as the identities or institutions of the authors to play a role. However,

This research was supported in financial terms by no one other than the author. I want to thank all editors that agreed to help me with this study (Daron Acemoglu, Franklin Allen, Harold Cole, Cam Harvey, Christian Hellwig, David Hirshleifer, James Hosek, Larry Katz, Robert Richmond, Matthew Spiegel, Joachim Voth, Fabrizio Zilibotti), the referees and authors from the SFS Cavalcade, the UCLA Office of the Human Research Protection Program (#12-000825), multiple anonymous referees, Jaclyn Einstein, and the editor, Andrew Karolyi. Any mistakes, errors, or misinterpretations are mine alone. Because this paper is about heterogeneity in referees, the author's website will post some remarkably negative earlier referee reports on this very paper.

doi:10.1093/rfs/hhu029 Advance Access publication May 5, 2014

most economists would agree that it should ideally be submission- and author-associated factors in the broadest sense and not referee-associated factors that should determine publication.[1]

My paper studies the extent to which referee recommendations reflect a shared consensus versus the extent to which they reflect referee-specific perspectives. If recommendations are relatively more idiosyncratic, then the evolution of knowledge is likely to be more path-dependent (and the careers of economists more random) than if recommendations reflect a general consensus.

There could be many reasons why referees share perspectives. They could agree not only with respect to the characteristics of the submissions (such as its novelty, interestingness, accuracy, rigor, and polish, as pointed out by Ellison 2002a), but also with respect to other non-submission-related characteristics (such as the identity of the authors). I shall refer to these aspects as the "reliable qualities" of submissions (not to be mistaken for the true scientific quality). Referees could also disagree for other reasons. There could be heterogeneity in their weightings of these characteristics, or there could be referee-specific factors such as noise, skills, time investments, moods, beliefs, ideologies, personal likes, age or cohort, professional networks, vanity, suppression of contrary evidence, or turf motives. Of course, if referees agreed about these characteristics and placed similar weights on them (e.g., if all referees liked pro-free-market papers), these same characteristics would become repeatable commonalities among referees. This would lead my paper to classify these components as "reliable"—again highlighting the difference between reliability (which I can measure) and submission quality (which I cannot measure). It is by this definition that the influence of referee characteristics that are not common (reliable) across referees become idiosyncratic (unreliable). The draw of the referee matters less when the reliable component of referee recommendations plays a more important role than the subjective component.[2]

An immediate concern in any study that seeks to determine the reliability of referees' recommendations is that editors do not choose referees randomly. This makes it difficult to determine whether any observed consensus reflects a reliable component of the referees' views about the submissions or whether it reflects merely the editorial referee selection decision. Therefore, my study examines referee behavior not just in the standard refereeing context (for eight journals: *Econometrica* [ECMTA], the *International Economic Review* [IER], the *Journal of the European Economic Association* [JEEA], the *Journal of Economic Theory* [JET], the *Quarterly Journal of Economics* [QJE], the *Rand*

[1] It is possible that the editorial process is a second-best solution to a moral-hazard problem: editors may have to indulge referee-idiosyncratic opinions in order to incentivize volunteer referees to participate in the editorial process.

[2] In my paper, I sometimes refer to the idiosyncratic referee-specific aspects as the "subjective evaluation" of the submission. This is not meant to imply that the common aspects do not contain subjective but widely shared views, or that the subjective evaluation cannot be based on objective criteria that only one referee considered.

Journal [RAND], the *Journal of Finance* [JF], and the *Review of Financial Studies* [RFS]), but also in a conference venue with an unusual referee selection: In the 2012 Society for Financial Studies (SFS) Cavalcade conference, a known computer algorithm matched referees to submissions based only on shared expertise.

Studying the two venues represents different tradeoffs. On the one hand, human journal editors can presumably match papers better to referee expertise, and journal referees spend more time on journal submissions than on conference submissions. On the other hand, editors may select the number and identities of referees based on their own prior assessments of submission quality or even a desire to influence the referees' recommendation and/or the agreement among multiple referees. My paper will show that referee behavior is very similar in both types of venues. Without the journal data, the conference data could be viewed as too different from the journal settings. Without the conference data, the journal data could be viewed as the result of deliberate editorial selection. Together, Occam's razor suggests that my paper documents behavior that is typical of economics and finance referees, and not an artifact of referee selection.

My paper focuses on the referees' final recommendations to the editors in situations in which two referees evaluated the same paper. The two key findings are straightforward. First, I document that the consensus among referees was modest. The idiosyncratic referee component is stronger than the common reliable component. The following simple statistics put this in perspective. The unconditional probability that a referee at the SFS Cavalcade would recommend accepting a paper was 28.5%. When one referee recommended accepting, the probability that another referee would agree increased only from 28.5% to 38.2%. (At the eight journals, the equivalent figures were 31% and 34%, respectively.) A decomposition model developed in my paper suggests a convenient summary statistic: similar levels of agreement among referees would have been observed if referees had placed about one-third weight on a shared signal and about two-thirds weight on their own idiosyncratic signal. Second, I document that there was significant variation in the "intrinsic generosities" among referees. For example, in the SFS Cavalcade, the probability that a referee would issue a "must accept" recommendation was only 3.2%. However, it increased to 6.9% if this referee judged other papers, *not including the current one*, to be at least of "neutral" average quality. Yet not all disagreement can be explained by differences in the average scorings of referees. When two SFS Cavalcade referees evaluated the same two papers, they agreed which paper was better in 972 cases and disagreed in 702 cases.

Beyond these two core findings, my paper interprets some of the consequences of the observed referee behavior and documents some further empirical regularities, including the behavior of editors in at least one journal.

1. The Data

This section describes the data from the SFS Cavalcade and the eight economics and finance journals used in the analysis.[3]

1.1 The SFS Cavalcade

I was the chair for the 2012 SFS Cavalcade conference. Eighteen distinguished researchers had been chosen as program committee members by the association before my appointment. I solicited additional referees from the set of 663 submitting authors. The assignment of referees to papers was made by a computer program without my intervention using the following algorithm:

1. For each submission-referee combination, the program computed a raw proximity score, based on the number of categories that the referee and paper shared. Categories were based on areas (four large areas like "asset pricing," fifty-one subareas like "options," and *JEL* classification codes), on approach (such as "theoretical"), and on level of complexity (such as "low-tech"). Authors and referees could choose as many designations as they liked. For example, if a referee indicated as expertise "Asset Pricing, International Asset Pricing, Theoretical, Structural, Mid-Tech," and the authors classified a submission to be "Asset Pricing, Empirical, Structural, Mid-Tech," the intersection was "Asset Pricing, Structural, Mid-Tech." The raw score was then the number of intersecting categories squared, divided by one plus the number of categories for the paper times one plus the number of categories for the author. In this example, the proximity score would have been $3^2/[(5+1)\times(4+1)]=0.3$.
2. After excluding authors of their own papers, the program iterated through the proximity-score-ranked list to assign referees to papers, making sure not to assign too many papers to each referee, and not to assign too many referees to each paper. The target number of papers per referee was 21 for the program committee members and 5 for ordinary referees. The target number of referees per paper was 5.

The median proximity score was about 0.2, with an interquartile range of about 0.1 to 0.3. The distribution of proximity scores was similar for ordinary referees and for SFS Cavalcade program committee members. Most important, because I did not intervene subjectively in the referee selection, the only paper-assignment selection bias could be one that was encoded in the computer algorithm—that is, an expertise-related one.

Table 1 shows the final distribution of recommendations used in the analysis. In total, 578 referees returned 3,126 recommendations on 367 papers.

[3] The gathering and analyses programs are available on the website. The data itself is unfortunately too sensitive and confidential to be sharable.

Table 1
SFS Cavalcade number of referees per paper and number of papers per referee

					Number of referees per paper									
#Referees	1	2	3	4	5	6	7	8	9	10	11			
Incidences	1	1	3	22	49	73	59	36	39	24	15			
Reports	1	2	9	88	245	438	413	288	351	240	165			
Cumulative	1	3	12	100	345	783	1,196	1,484	1,835	2,075	2,240			
					Number of referees per paper									
#Referees	12	13	14	15	16	17	18	19	20	21	28			
Incidences	11	9	9	3	2	3	1	1	3	2	1			
Reports	132	117	126	45	32	51	18	19	60	42	28			
Cumulative	2,372	2,489	2,615	2,660	2,692	2,743	2,761	2,780	2,840	2,882	2,910			
					Number of papers per referee									
#Papers	1	2	3	4	5	9	12	17	18	20	21	25	26	28
Incidences	3	7	30	125	395	1	2	1	3	2	5	1	1	1
Reports	3	14	90	500	1,975	9	24	17	54	40	105	25	26	28
Cumulative	3	17	107	607	2,582	2,591	2,615	2,632	2,686	2,726	2,831	2,856	2,882	2,910
				←Submitter-referee			SFS Cavalcade program committee member→							

Explanations: Papers were matched to referees based only on shared expertise. There were 578 referees evaluating 367 papers with 3,126 recommendations. The analysis in my paper is based on the 2,910 reports in which referees rated themselves "not conflicted." Only members of the program committee were asked to referee more than five papers.

The referees themselves identified 216 recommendations to be conflicted (vis-à-vis the submitting author), leaving 2,910 unconflicted recommendations. Because the mean paper rating for the conflicted reports was significantly higher, the remainder of my paper focuses only on these 2,910 unconflicted recommendations. The 18 program committee members provided between 9 and 28 paper reviews. No ordinary referee evaluated more than 5 papers. The most common number of referees per paper was 6–7.

1.2 Academic journals

I also obtained access to data from six economics journals (*Econometrica* [ECMTA], the *International Economic Review* [IER], the *Journal of Economic Theory* [JET], the *Journal of the European Economic Association* [JEEA], the *Quarterly Journal of Economics* [QJE], and the *Rand Journal* [RAND]) and one finance journal (the *Review of Financial Studies* [RFS]) that used the Editorial Express (EE) web system. In addition, I was given access to redacted data from the *Journal of Finance*. The editors of the EE journals ran a perl program on my behalf in-house on an EE data dump that they downloaded to their own local computers. Thus, I never had direct access to the data itself but was still able to link referees on one paper to their decisions on other papers.

1.3 Frequency of multiple referees

Most of my analysis focuses on referee-pairs—that is, situations in which two or more referees evaluated the same submission. By necessity, the referee-pair unit of analysis excluded both desk rejects and single-referee submissions. When a submitted paper had more than two referees, each possible pair evaluation was entered as one observation in much of the analysis—a paper with three $[n]$ referees yielded three $[n \times (n-1)/2]$ pairs. Table 2 shows the fraction of submissions that were evaluated by more than one referee, the average number of referees per paper, and the number of pairs.

The two finance journals tended to use fewer referees per submission than the six economics journals. The JF used more than one referee in only 20% of their submissions. The mean number of referees was 1.2. The RFS used more than one referee in 31% of their submissions. The mean number of referees was 1.3. At the economics journals, the average number of referees ranged from 1.6 referees per paper at JET to 2.1 and 2.6 referees per paper at ECMTA and the QJE. The number of referee pairs grows quadratically with the number of referees. Thus, while the *Journal of Finance* provided only 1,856 paired referee recommendations, the QJE provided 16,544 and ECMTA provided 15,826. With its unusually large number of referees per paper (an average of 5.1 referees per paper), the 2,910 recommendations in the SFS Cavalcade yielded 24,370 referee pairs. With 87,114 referee-pair recommendations, most statistics reported in my paper have small

Table 2
Multiple referee situations

	Submissions with > 1 referee	Mean referees per submission	Referee pairs
Econometrica (ECMTA)	75%	2.1	15,826
International Economic Review (IER)	85%	2.1	9,702
J of the European Economic Association (JEEA)	74%	2.2	9,922
J of Economic Theory (JET)	50%	1.6	3,024
Quarterly J of Economics (QJE)	88%	2.6	16,544
Rand Economic J (RAND)	81%	1.9	2,322
		6 economics journals	57,340
J of Finance (JF)	20%	1.2	1,856
Review of Financial Studies (RFS)	31%	1.3	3,548
RFS (same time)	28%	1.3	3,028
		2 finance journals	5,404
		8 journals	62,744
SFS Cavalcade	All	5.1	24,370
		All 9 venues	87,114

Explanations: The first column is the venue. The second column is the frequency of papers that had more than one referee. The third column is the mean number of referees per paper. The fourth column is the number of paired evaluations that are available for analysis. Each pair is a unique combination of two referees evaluating the same paper. (Thus, for example, one paper with 5 referees would provide 10 referee pairs.) These statistics exclude desk rejects.

standard errors, allowing the reader to focus on the economic meaning of the estimates.

1.4 Frequency of categorical recommendations

In my sample, referees for EE journals had seven choices: "definitely reject," "reject," "weak revise and resubmit," "revise and resubmit," "strong revise and resubmit," "accept with revisions," and "accept." Referees for the SFS Cavalcade had five choices: "must reject," "should reject," "neutral," "should accept," and "accept." Referees for the *Journal of Finance* had three choices: "reject," "resubmit," and "accept." To make the journals more comparable (and because the differences in meaning between some categories are difficult to understand), most of my analysis collapses the recommendations into four categories (except at the JF, where I only had three categories to begin with), dubbed "Reject" (REJ), "Weak" (WEAK), "Revise" (R&R), and "Accept" (ACC). Table 3 shows how this mapping was accomplished.

More important, Table 3 shows that referees at the six economics journals tended to be more generous than referees at the RFS finance journal in the R&R and better categories. Due to the differences in the number of categories, the two other venues (JF and the SFS Cavalcade) are difficult to compare, although the patterns seemed qualitatively similar. Overall, 56% of the referee reports recommended REJ, 18% were WEAK (very cautious-revise-and-resubmits), 17% were R&R, and 9% were ACC. Note that these recommendations are

Table 3
Frequency of referee recommendations

	SREJ	REJ	WR&R	R&R	SR&R	ACR	ACC
ECMTA		0.64	0.13	0.11	0.06	0.04	0.02
IER	0.05	0.52	0.07	0.25	0.04	0.03	0.03
JEEA	0.09	0.50	0.14	0.15	0.06	0.04	0.01
JET		0.49	0.14	0.17	0.09	0.08	0.04
QJE	0.09	0.57	0.12	0.12	0.05	0.03	0.01
RAND	0.11	0.53	0.14	0.14	0.05	0.02	0.01
6 economics journals	0.06	0.56	0.12	0.15	0.06	0.04	0.02
JF		0.60	0.33		0.06		
RFS	0.17	0.56	0.10	0.11	0.04	0.01	0.00
RFS (same time)	0.17	0.57	0.10	0.11	0.04	0.01	0.00
2 finance journals	0.11	0.58	0.18		0.05	0.01	0.00
8 journals	0.06	0.56	0.13	0.14	0.06	0.04	0.02
SFS Cavalcade	0.07	0.32	0.32	0.25		0.04	
All 9 venues	0.06	0.50	0.18	0.17	0.04	0.04	0.01
Summary categories:	Reject (REJ) 56%		Weak (WEAK) 18%	Revise (R&R) 17%	Accept (ACC) 9%		

Explanations: The first column is the venue. The remaining columns are the unconditional frequencies based on all referee pairs. In the Editorial Express (EE) journals, the "definitely reject" choice was "SREJ," the "weak revise and resubmit" was "WR&R," the "strong revise-resubmit" was "SR&R," and the "accept subject to revisions" was "ACR," The ECMTA and JET programs did not distinguish between SREJ and REJ. Most of the analysis in my paper relies on the four summary categories at the bottom of the table.

referee recommendations and not journal decisions. They do not allow inferring the selectivities of the venues themselves.

1.5 Comparison

The journal data and the SFS Cavalcade data have different strengths and weaknesses. The advantages of the journal setting are (i) journals had multi-year histories (many more refereed papers), (ii) journal editors could match papers better with referee expertise than my computer program could, and (iii) referees probably spent more time evaluating each submission.[4] The advantages of the SFS Cavalcade setting are (i) referee assignments are guaranteed not to be correlated with an a priori assessment of the paper's quality by the editor or with an a priori intent of editors to solicit agreement or disagreement, and (ii) each paper had an unusually large number of referees. In addition, there could be other differences. The same referee may put different weights on different attributes in different venues. For example, conference referees may have put relatively more weight on whether a submission was interesting than whether its proofs were correct.

[4] However, SFS Cavalcade referees that had better-matched expertise, that claimed to have spent more time on the paper, and that had more papers to review, did not show more or less consensus among themselves than other referees.

Table 4
SFS Cavalcade referee recommendations conditional on one other referee's recommendations

		Other referees' recommendations								
		MR	SR	NR	SA	MA		Unconditional		
Own Recommendation		–2	–1	0	+1	+2	Pairs	Freq	Reports	Freq
Must Reject MR	–2	312	718	461	244	32	1,767	0.073	202	0.069
Should Reject SR	–1	718	2,904	2,494	1,520	163	7,799	0.320	905	0.311
Neutral NR	0	461	2,494	2,584	2,027	300	7,866	0.323	933	0.321
Should Accept SA	+1	244	1,520	2,027	1,844	368	6,003	0.246	743	0.255
Must Accept MA	+2	32	163	300	368	72	935	0.038	127	0.044
						Total:	24,370	1	2,910	1

		Translated into conditionals				
		MR	SR	NR	SA	MA
Must Reject MR	–2	0.177	0.41	0.26	0.14	0.018
Should Reject SR	–1	0.092	0.37	0.32	0.19	0.021
Neutral NR	0	0.059	0.32	0.33	0.26	0.038
Should Accept SA	+1	0.041	0.25	0.34	0.31	0.061
Must Accept MA	+2	0.034	0.17	0.32	0.39	0.077
Unconditional		0.073	0.320	0.323	0.246	0.038

Explanations: The predicted referee recommendations are based on 2,910 pairable referee recommendations from the SFS Cavalcade. For each referee recommendation in the first two columns, I tabulate the frequency of recommendations for the same paper by other referees. (Thus, for example, a paper with 5 referees provided 10 referee pairs.)
Interpretation: Referee recommendations on the same paper are significantly positively correlated. They share a reliable component. However, the correlation is modest. The matrix does not approximate the identity matrix. The reliable component is not large.

2. Consensus

2.1 The SFS Cavalcade paired recommendation matrix

To aid intuition, start with the SFS Cavalcade, the venue with the largest number of referee pairs. Table 4 tabulates the observed recommendations. The bottom two rows in the top-right subtable show that of the 2,910 unconflicted recommendations, 127 (4.4%) advised "Must Accept" (MA), and 743 (25.5%) advised "Should Accept" (SA). Thus, about one in three recommendations was positive. About one in three recommendations was neutral (although neutral is widely understood to mean rejection in highly competitive contexts). And about one in three recommendations was negative, "Should Reject" or "Must Reject." These probabilities do not change much when computed for individual reports instead of for paired reports. My discussion focuses on the two highest recommendations, MA and SA. After all, only primarily positive recommendations allow a paper to be accepted into a selective journal or conference. The unconditional probability of an MA recommendation was 3.8% (935 out of 24,370 paired recommendations).[5]

The upper table shows the raw number of paired recommendations. The lower matrix is normalized to conditional probabilities. Inspection of the matrix

[5] There was relatively more consensus for papers that received an MR. Another referee is likely to share this view with 16.7% probability, higher than the 7% unconditional probability of an MR.

reveals that there was modest consensus. Consider a paper that received one rare MA endorsement from one referee:

- The probability that another referee also offered an MA recommendation was $72/935 \approx 7.7\%$—higher than the unconditional 3.8%, but far from 100%.
- The probability that another referee offered the next-best recommendation of SA was $368/935 \approx 39\%$—higher than the unconditional 25%, but still not even a fair bet.
- The probability that another referee recommended not only a reject but a strong reject (MR) was still $32/935 \approx 3.4\%$—lower than the unconditional 7.3%, but not zero.

Thus, even given one MA, the chances were better that another referee would offer a negative to neutral recommendation on the submission ($495/935 \approx 53\%$) than that she would offer a second positive recommendation. (A neutral evaluation essentially suggested nonselection of the submission in this competitive a venue.)

The picture is much the same when one referee reported either an SA or an MA ($6,003+935$ out of 24,370 recommendations, 28.5%). The probability that another referee's recommendation was SA or MA was $(72+368+1844+368)/(6,003+935) \approx 38.2\%$, higher than the 28% unconditional probability, but not close to 100%.

In sum, there was more consensus than would have been observed by random chance if referees' opinions had been uncorrelated, but much less than what would have been observed if assessments had been perfectly reliable.

2.2 A decomposition model of referee behavior

It is not easy to interpret pairwise recommendation matrices intuitively. Thus, it is useful to consider a simpler model that maps recommendation matrices into summary statistics.

2.2.1 A low-dimensional model with continuous reports. Assume that referee $R \in (A, B)$ places weights w_R on k different unit-normalized and orthogonalized characteristics c_k of the submission,

$$r_R = w_R \cdot c = \sum_k w_{R,k} \cdot c_k .$$

Specifying the characteristics in this generic linear fashion allows the model to encompass a wide range of decision inputs, such as the submission's true scientific value, its likely future impact, its writing quality and style, the identity of its authors, and so on. It can also include characteristics that most academics would agree should not influence publication decisions (such as the LaTeX format quality or number of vowels in the submission) and multiple noise

terms. If a referee R does not observe characteristic k, her weight $w_{R,j}$ on this characteristic would be zero.

Assume that the editor (and my analysis) observes neither the individual characteristics nor the referees' weightings. The editor observes only the final recommendation r_R. Initially, assume that r_R is not categorical but continuous, with finer gradations perhaps discernable through the reading of the full referee report. The correlation between two referee assessments r_A and r_B for the same paper is

$$\text{Cor}(r_A, r_B) = \frac{w_A' C w_B}{\sqrt{w_A' C w_A w_B' C w_B}},$$

where C is the k-by-k matrix cc'. For example, if referees A and B place weights $w_A = (1/6, 2/6, 3/6, 0)$ and $w_B = (2/6, 1/6, 0, 3/6)$ on the $k = 4$ characteristics, then the correlation among referee recommendations would be 0.286. (In an OLS regression, one referee's opinion would explain 8.2% of the variance in the other referee's opinion.) The correlation is less than one, both because of differences in weighting on characteristics that both referees share (here the first two weights) and because of weights on characteristics for which referees have their own unique views (here the last two weights).[6] Any final correlation in referee recommendations maps into infinitely many higher-dimensional models.

The intent of my decomposition model is to characterize referee behavior with a summary statistic that can then be used in simple thought experiments. This model is calibrated to yield the same correlation as that observed in the data. It maps the higher-dimensional space into a "shared-signal" decomposition with only three characteristics: one shared characteristic (c_S), one characteristic that is unique to referee A (c_A), and one that is unique to referee B (c_B), again with all three characteristics orthogonal and unit-normalized. Lambda is the proportion of weight on the shared characteristic. The referees' recommendations are

$$r_A = \lambda \times c_S + (1-\lambda) \times c_A \qquad r_B = \lambda \times c_S + (1-\lambda) \times c_B \tag{1}$$

and

$$\text{Cor} = \frac{\lambda^2}{\lambda^2 + (1-\lambda)^2} \quad \Leftrightarrow \quad \lambda = \frac{\text{Cor} - \sqrt{\text{Cor} - \text{Cor}^2}}{2 \times \text{Cor} - 1}.$$

This shared-signal model gives the same correlation of 0.286 between referee assessments if both referees had placed weight $\lambda = 0.387$ on the single shared characteristics c_S and weights $1-\lambda = 0.613$ on their unique terms, c_A and c_B, respectively. Any positive correlation between zero and one maps into one

[6] If an editor wanted to place 1/4 weight on each of the four characteristics (e.g., if the characteristics were importance of the paper's contribution to four different subfields), then she could obtain her desired estimate by averaging the two referees' recommendations. More generally, with as many referees as characteristics, an editor who knows the weights of referees on characteristics could uncover the signals and thus determine an optimal linear combination of the signals to decide on the manuscript.

unique value of λ. The extremes of this model are easy to interpret: if there is no overlap in the full high-dimensional model's weights w_i or if the weights are orthogonal ($w'_A w_B = 0$), then lambda is 0. If there is perfect overlap ($w_A = w_B$), then $\lambda = 1$. If lambda is 1/4, 1/2, or 3/4, then the correlation among referee reports (r_A, r_B) is 1/10, 5/10, and 9/10, respectively.[7]

Even if a true scientific paper quality existed upon which the paper should ideally be decided, the observed referee commonality is *not* the referees' agreement about this true paper quality. To see this, assume that this true quality is the first element in the vector c, and both referees observe it (perhaps with modest noise) but do not place 100% weight on it. The observed consensus among referees could then be higher than their agreement about the true quality—for example, if both referees based their recommendation only on a reliable but unimportant metric, such as the number of spelling errors. The observed consensus could be lower if one referee believed that recommendations should be based on the citation counts of authors (to maximize future citation impact of the submission), while the other believed that her recommendation should be based purely on the rigor of the paper or the fact that the paper contradicts some of her own earlier research. However, because this true paper quality is one among a number of reliable characteristics, if both referees placed positive weight on it, it would contribute to a positive correlation. If the recommendation correlation were zero, it would seem unlikely that a true paper quality would be an important input into referee recommendations.

2.2.2 Inferring lambda from observed discrete categories. Although editors may be able to interpret the contents of the referee report and the letter to the editor to infer smooth gradations in recommendations, the final referee recommendations on which my own paper is based are the discrete categories reported by referees. This makes it more difficult to infer lambda.

To fit the best lambda, I first simulate the reduced-space model based on the number of observed pairings for different lambdas. I assume that the three characteristics, c_S, c_A, and c_B, are independent normals. For example, for the SFS cavalcade, I draw 24,370 c_S characteristics, 24,370 c_A characteristics, and 24,370 c_B characteristics.[8] This gives the two "raw" referee scores r_A and r_B according to Equation (1). Next, I need to discretize them. As Table 3

[7] The discussion in my paper is phrased in terms of this agreeable shared-signal summary model for its decomposition intuition, not for the presumption that referees do not make choices based on higher-dimensional evaluation functions. My paper assumes that there is a true unknown summary parameter lambda (mapped from its true higher-dimensional function), and I observe a random draw with a sample lambda from which I infer the true lambda.

[8] This is to preserve the use of lambda as a summary statistic for the pairwise matrix. For the SFS Cavalcade, I also knew the referee-paper pairings, which makes it possible to preserve the structure by simulating the 578 referees' 2,910 recommendations on 367 papers. This provides some additional restrictions. Not surprisingly, it barely changes the inference. It changes the reported coefficient from 0.364 to 0.363. The estimates in Table 7 are based on this full-pairing estimation.

showed, the frequency of categories in the recommendations varied with the venue. For example, referees gave the highest grade in about 20% of the JET submissions, but only in about 3.8% of the SFS Cavalcade submissions. To match each venue's recommendations, I translate the raw referee signals r_R into the observed frequencies of discrete recommendations. For example, because the SFS Cavalcade referees offered 935 "must accept" recommendations out of a total of 24,370 recommendations, the highest 935 simulated recommendations for each of the two referees is assigned into the "must accept" category. My discretization procedure can be thought of as an in-sample estimator for the true unobserved frequencies of categories based on the realized frequencies of categories.

The simulated discretized referee-pair recommendation matrix is then calculated from r_A and r_B. If lambda is one (i.e., referees report the reliable characteristic c_S), the simulated matrix is diagonal. If lambda is zero (referees report c_A and c_B, respectively), the simulated matrix is (on average) the unconditional probability vector product.

The reported estimate of lambda, λ^*, is the one that has the least mean-squared equal-weighted probability difference (distance) between its simulated four-by-four referee-recommendation pair matrix and the empirically observed four-by-four matrix.[9] The model typically provides a good fit for all venues, with cell-averaged root-mean-squared probability distances between the observed and simulated optimal probability matrix of less than 0.001%. This can also be translated into a number of referee recommendations that would need to be changed to perfectly match the observed matrix. (One disadvantage of this metric is that a change from an REJ to an ACC matters as much as a change from an REJ to a WEAK. The advantage is the intuitive nature of this metric.) For example, for the SFS Cavalcade, there were 24,370 referee pairs. If referees had chosen independently ($\lambda = 0$), then to match the empirical distribution of recommendations in Table 4 would require changing 448 recommendations. If referees had chosen identically ($\lambda = 1$), it would require changing 2,458 recommendations. At the best estimate of λ, $\lambda^* = 0.364$, it would require changing only 16 referee recommendations. The model can almost (but not quite) provide a sufficient statistic for the information in the paired data.[10]

2.3 Estimates of consensus and lambda

The left columns in Table 5 show the pairwise referee correlations and inferred lambda parameters if a distance of one is assigned between the four categories

[9] This is a simple simulated method of moments (SMM) estimator. If I minimize the absolute misclassifications in Table 5, the lambda inference typically changes by no more than 0.01. The reported confidence interval is based on the critical lambdas that do not cover the observed distance to that provided by the optimal estimate of lambda λ^* within their 90% confidence intervals.

[10] The results are similar for other distance statistics based on the two matrices, such as the mean-squared error in the diagonal agreement vector, the agreement only among the best recommendation, or the number of recommendations that would need to be changed to achieve perfect fit. The normal distribution on the three c characteristics fits the data modestly better than either a Cauchy or a uniform distribution.

Table 5
Paired referee correlations and λ parameter estimates

	Correlation			Agreement λ			Misclassifications		
Journal	Spear	Pear	McFd	5%	Mean	95%	0	λ^*	1
ECMTA	0.18	0.20	0.12	0.343	0.368	0.394	285	12	1,539
IER	0.26	0.22	0.16	0.393	0.437	0.467	280	10	903
JEEA	0.22	0.18	0.13	0.346	0.381	0.408	225	15	978
JET	0.25	0.26	0.16	0.356	0.405	0.445	65	13	313
QJE	0.19	0.15	0.12	0.328	0.352	0.375	261	12	1,585
RAND	0.25	0.24	0.15	0.368	0.424	0.465	56	5	214
6 economics journals	0.25	0.20		0.370	0.388	0.403	1,233	23	5,531
RFS	0.14	0.14	0.10	0.236	0.309	0.365	32	5	313
JF (not EE)	0.12	0.15	0.09	0.229	0.319	0.376	25	4	197
2 finance journals	0.15	0.14		0.267	0.324	0.371	68	7	510
SFS Cavalcade	0.21	0.20	0.13	0.344	0.364	0.378	448	16	2,458
All 9 venues	0.27	0.22		0.385	0.405	0.422	2,235	49	8,497

Explanations: All statistics are based on referee-pair observations on the same submission. Except for the *Journal of Finance*, where referees could only recommend one of three dispositions, the four recommendation categories allowed for 16 possible pair values. A submission with more than two referees created multiple pairs. (This matters primarily in the SFS Cavalcade, where the average paper had about five referees.) The correlation columns contain the Spearman (Spear) and Pearson (Pear) correlation coefficients if a distance of one is assigned between the four categories (REJ, WEAK, R&R, ACC). They also contain the square root of the McFadden (McF) R^2 from a probit explaining the recommendation of one referee with dummies spanning the recommendation of the other referee. Lambda is a summary statistic that estimates a weight that referees place on the reliable characteristic if referees report a discretized net signal of $\lambda \times c_S + (1-\lambda) \times c_R$, where c_S is the shared reliable signal, c_R is the referee's own signal, and the two signals are orthogonal unit-normals. λ minimizes the equal-weighted probability deviation in the simulated versus the empirical squared probability matrix. The last three columns show the number of recommendations that would have to be changed to achieve perfect fit. (A lambda λ^+ that would minimize these misclassifications is very similar to λ^* but not identical.)
Interpretation: Referees at ECMTA, the QJE, the RFS, and the JF had lambda statistics between 0.30 and 0.37. Referees at other economics journals had lambda statistics around 0.40.

(REJ,WEAK,R&R,ACC). The Spearman and Pearson correlation coefficients are around 20% to 25% for the six economics journals, 14% to 15% for the two finance journals, and 20% for the SFS Cavalcade. The square root of the McFadden R^2 from an ordered probit predicting one referee's decision with that of the other (either as a single value or a set of dummies) is about half the size of the more common correlation coefficients.

The right columns show the estimates of lambda. They seem high relative to the correlation coefficients because even relatively strong referee agreement results only in modest extra consensus.[11] The lambda estimate is 0.39 for the six economics journals, 0.32 for the two finance journals,[12] and 0.36 for the SFS

[11] For intuition, if referees either observed or did not observe the true characteristic of the submission with probability λ, then consensus above the expected frequency under the null would occur only when both referees get to see the true characteristic. One random and one correct draw do not yield more consensus. Thus, the consensus would increase with probability λ^2. If $\lambda = 0.5$, then each referee has a 50% probability of observing the reliable characteristic. If only one referee observes the reliable quality and the other observes a random draw, the consensus is the same as if both referees observe a random draw. Thus, "excess consensus" occurs only $50\% \times \% = 25\%$ of the time.

[12] For the RFS, the estimates are very similar when I focus only on referees that were solicited on the same day.

Cavalcade. The 5% to 95% ranges that cannot be rejected are narrow, because the number of observations is large and the model has only 10 probabilities to fit with three cutoff and one lambda parameters. The last three columns show the number of reports that would have to be changed to fit the observed empirical pairwise matrix perfectly. Over all nine venues, the model would have fit perfectly if 49 referee recommendations (out of 87,114) were changed. This is much better than the 2,235 recommendations that would have to be changed if there was no consensus ($\lambda = 0$) or the 8,497 recommendations if there was perfect consensus ($\lambda = 1$).

2.4 Counterfactuals

The distribution of recommendations (Table 3) and the estimates of lambda (Table 5) make it possible to consider thought experiments (albeit assuming that the pairwise recommendations are representative of recommendations for all submitted papers, including single-refereed submissions and zero-referee desk rejects). I estimate the referee recommendations that a submitting author and the editors can expect for submissions with a given reliable common characteristic (not true quality). To do so, I consider two values for lambda, $\lambda \approx 0.33$ (roughly representative of ECMTA, the QJE, the RFS, and the JF), and $\lambda \approx 0.40$ (roughly representative of the full set of venues, including the IER, JEEA, JET, and RAND). Similarly, I use recommendation category frequencies in line with the empirical data. They are not intended to match any one venue perfectly. In these hypotheticals, I presume that the editor has knowledge of the raw referee recommendations, and not just the discretized recommendations.

2.4.1 Hypothetical referee recommendation probabilities. Figure 1 plots the recommendations of a single referee r_R as a function of the true reliable characteristic of the paper, c_s. The top graphs show sample random draws that are intended to give a visual impression for how the reliable c_S ranks of 1,000 submissions tend to map into referee reports r_R. The recommendations are neither random nor strongly associated with the true reliable characteristic. The bottom graphs show the probabilities. The 50% horizontal line can be used to assess the asymptotic consensus of referees. If the probability of receiving a given report is below 50%, then as the number of referees increases (goes to infinity), the probability that a majority of referees will recommend it decreases with the number of referees (goes to zero). The graph and table below shows that if $\lambda = 0.33$, the top one percentile paper (the 10th best submission out of 1,000) should expect the majority of its referee reports not to recommend a (strong) revise-and-resubmit or accept, even though such recommendations constitute 13% of the referee reports. Another striking observation is that the more referees the editor consults for the top two percentile paper, the lower is the probability that its *average* review will be an R&R or better.

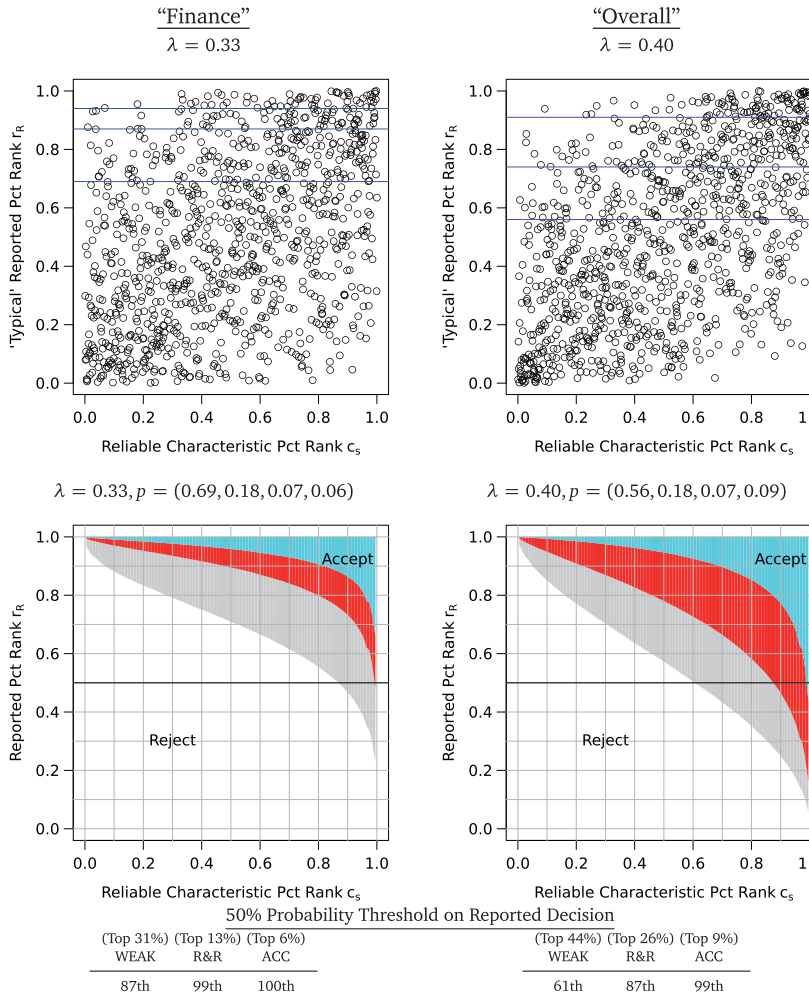


Figure 1
Conditional probabilities of referee recommendations based on estimated λ
Explanations: The left side (right side) assumes a lambda of 0.33 (0.40). The top graph plots the typical relation between the true reliable characteristic c_S percentile rank and the referee-reported percentile rank r_R for 1,000 hypothetical submissions. (The "reliable" characteristic c_S is *not* a measure of the true scientific quality of the submission.) The horizontal lines in the top graphs indicate the frequencies of the four categories in the data, which are also inputs into the bottom graphs. These probabilities of REJ, WEAK, R&R, and ACC are 0.69, 0.18, 0.07, and 0.06 (0.56, 0.18, 0.07, and 0.09), respectively. The lower graphs show the probabilities of different categorical recommendations. The numerical 50% horizontal crossings are listed below. For example, in the left bottom graph, a submission that was truly ranked at the 87th percentile had a 50% probability of receiving a WEAK referee recommendation or better. To expect an R&R with greater than 50% probability required a submission with a true c_S characteristic rank of 99% or better.

2.4.2 Hypothetical journal decisions. Assume now, in addition, that a journal wants to continue with the 10% of its submissions that have the highest reliable characteristics c_S (not necessarily the best papers, but those

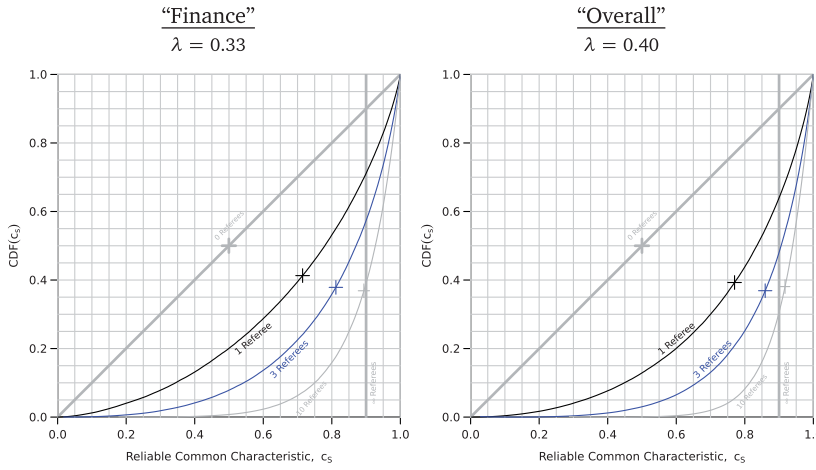


Figure 2
Cumulative distribution of c_S among the top-10% r_R ranked submissions
Explanations: Assume that journals accept the submissions that gather the top-10% highest average referee reports (mean r_R). These graphs then plot the cumulative probability density of the accepted submissions. The cross-hatch is the mean. For example, if $\lambda = 0.33$, if a journal evaluates each submission with three referees, then about 36% of the accepted papers have a true common characteristic c_S that is below the 80th percentile. Because the common characteristic c_S is normally distributed, the 80th percentile corresponds to a true common characteristic $Z = 0.8416$.

that would elicit the most referee approval if repeated many times), that the journal considers each referee to be equally informed, and that the editor has no signal of her own. (In Section 2.5, editors will be assumed to have their own signals, too.) In this case, the editor would continue with the 10% of the submissions that have the highest ranked mean recommendations, $\bar{r}_R$.

Figure 2 illustrates the journal's selection:

Lambda of 0.33: With one referee, the average true percentile rank of continuing papers on the common characteristic is 0.72, about 20% of the continuations are from the bottom half, and 70% of the continuations are misclassified in that their true common characteristic is less than 0.9. With three referees, the average true percentile rank increases to 0.82, about 8% rank in the bottom half, and 57% are not truly top-10%. Even with ten referees, the average percentile rank is 89%, almost no paper from the bottom half is continued, and about 40% of the accepted submissions are not top-10%.

Lambda of 0.4: With one referee, the average true percentile rank of continuing papers on the common characteristic is 0.77, about 13% of the continuations are from the bottom half, and 65% of the continuations are misclassified in that their true common characteristic is less than 0.9. With three referees, the average true percentile rank increases to 0.86, about 3% rank in the bottom half, and 48% are not truly top-10%.

> Even with ten referees, the average percentile rank is 92%, almost no paper from the bottom half is continued, and about 30% of the accepted submissions are not top-10%.

Referee reports are informative but not very reliable.

It is not obvious what the decision rule should be if there is variation in the number of referees that are assigned to submissions. For example, assume that all papers are drawn from the same common characteristics distribution and that lambda is 0.33. If one half of the submissions randomly receive one referee report, while the other half receives two referee reports, then two-thirds of the accepted (i.e., top-10% ranked) papers will come from the first (single-referee) set. This is because it is more common to draw a far-positive outlier with fewer referee reports. The problem is isomorphic to the discrimination-inference problem in Cornell and Welch (1996).[13] The decision problem does not become easier if editors follow a known deterministic rule that assigns multiple referees to some but not other submissions.

2.5 Editorial decisons and referee reports

One journal editor remarked that he ignored the categorical referee ratings and focused only on the detailed reports. It is therefore an interesting question whether the categorical referee reports are generally congruous with the editorial decisions. This journal provided me with a second data set to consider its editorial decision statistics.[14]

In the following analysis, I consider only submissions in which editors consulted one or two referees. (Three-referee submissions were too rare.) In the data, the referees were more favorably disposed toward dual-refereed papers, hinting that editors deliberately assigned two referees to papers for which they had more favorable priors:

	Papers	REJ	WEAK	R&R	ACC	Editor continued
Solo referee	3,272	80.7%	7.2%	9.2%	2.9%	17.2%
Dual referee	1,386	75.5%	10.0%	10.8%	3.8%	20.1%

[13] See Cornell and Welch (1996) for more detail on the model. The problem is worse with discrete categories. Consider an example in which a journal consults two referees for some papers and one referee for others. If a journal is so selective that it only publishes papers with the highest implied inference, it may only publish papers with two (good) reviews. One-review papers cannot reach as high an inference. If the journal publishes more papers, so that it accepts papers with one (good) review, the probability that a two-review paper is published is less than the probability that a one-review paper is published. (The comparison is between getting two-out-of-two good reviews and one-out-of-one good reviews.)

[14] Below, I report statistics based on referee reports that were solicited at the time of the original paper submission, but the results are almost the same if I also include reports solicited later (sequentially, i.e., after the first reports came in).

Table 6
Editorial congruance with referees and editorial inference weights, given $\lambda = 0.33$

Solo referees		Editor rejects (83%)				Editor R&R (17%)			
		Empirical		If $w_E =$		Empirical		If $w_E =$	
Referees	Papers	#Obs	Freq	0.5	0.35	#Obs	Freq	0.5	0.35
REJ	2,641	2,529	0.957	0.928	0.961	112	0.042	0.071	0.038
WEAK	236	117	0.495	0.619	0.577	119	0.504	0.380	0.422
R&R	301	52	0.172	0.386	0.221	249	0.827	0.613	0.778
ACC	94	10	0.106	0.120	0.014	84	0.893	0.879	0.985
Dual referees		Editor rejects (80%)				Editor R&R (20%)			
		Empirical		If $w_E =$		Empirical		If $w_E =$	
Referees	Papers	#Obs	Freq	0.5	0.35	#Obs	Freq	0.5	0.35
Agreeing referees — negative									
REJ,REJ	817	794	0.971	0.964	0.982	23	0.028	0.035	0.017
Agreeing referees — positive									
R&R,WEAK	28	4	0.142	0.296	0.187	24	0.857	0.703	0.812
R&R,R&R	25	1	0.040	0.130	0.023	24	0.960	0.869	0.976
ACC,WEAK	11	1	0.090	0.247	0.174	10	0.909	0.752	0.825
ACC,R&R	23	0	0.000	0.049	0.011	23	1.000	0.950	0.988
ACC,ACC	3	0	0.000	0.000	0.000	3	1.000	1.000	1.000
Ambivalent referees									
WEAK,WEAK	21	12	0.571	0.518	0.400	9	0.428	0.481	0.599
REJ,WEAK	195	146	0.748	0.751	0.715	49	0.251	0.248	0.284
REJ,R&R	197	121	0.614	0.559	0.507	76	0.385	0.440	0.492
ACC,REJ	66	28	0.424	0.440	0.462	38	0.575	0.559	0.537

Explanations: This table shows the expected frequency of editorial decisions and referee recommendations. Both referees and editors are assumed to see a final signal r that has a weight of $\lambda = 0.33$ on the common characteristic c_S and weight 0.67 on their own idiosyncratic signals (c_R or c_E). The editor continued with 18.1% of the submissions. This means the editor would accept the 18.1% of submissions for which her c_S inference was highest. An editor may put less weight on her own signals w_E if she believes it to be less informative. If $w_E = 0.5$, then with one referee, an optimizing editor would place as much weight on the referees' signals as on her own signal. More generally, with one referee (two referees) she would base her decision on the rank of $[w_E \times r_E + (1-w_E) \times r_A]$ ($[w_E \times r_E + (1-w_E) \times (r_A + r_B)]/[w_E + (1-w_E) \times 2]$). In the "EW" column, the editor places equal weight $w_E = 0.5$ on her own signal as on a referee's signal. The behavior of editors seems to match a $w_E \approx 0.35$; that is, the editors placed about twice as much weight on referees' signals as on their own. The resulting probabilities of rejection and continuation are in the 0.35 columns.
Interpretation: The empirical evidence suggests that it was rare for the editors of this journal to override the referees' recommendations.

In the 1,386 dual-referee situations, if referees had had perfect agreement, there should have been $0.755 \times 1,386 = 1,046$ REJ,REJ pairs. If referees had been uncorrelated, there should have been $0.8072^2 \times 1,386 \approx 790$ REJ,REJ pairs. Table 6 shows that there were 817 REJ,REJ pairs in the data, in line with my earlier findings of modest agreement among referees.

Unlike most other results in my paper, the main units of analysis in this section are not referee pairs but editor-referee pairs and editor-referee-referee triplets. Table 6 shows that the decisions of editors aligned closely but not perfectly with the recommendations of the referees. Solo referees recommended 2,641 rejections, 236 weak revisions, 301 strong revisions, and 94 accepts. When a solo referee recommended an REJ, the editors rejected the paper in 2,529 out of 2,643 cases (95.7%) and issued an R&R in the other 112 cases (4.3%). When two referees recommended rejection, the editor rejected in 794 out of 817 cases (97.2%) and continued in 23 cases (2.8%). In the 3,916 cases in which at least

one out of two referees recommended rejection, the editor rejected the paper in 3,618 cases (92.4%) and continued in 298 cases (7.6%).

Given that that only 18.1% of the submissions were continued, it is not surprising that most negative referee reports result in rejections. This would be the case even if editors believed that they had independent information that was as relevant as that of the referee: after one negative referee report, even very positive editors' signals would rarely be sufficient to rank a submission among the top 18%.

However, the evidence suggests editors followed referees even more than would have been expected if they had had their own signal of equal quality and placed equal weight on it. This estimate comes from a model analogous to the paired-referee model in which editors place some weight on their own private signal and some weight on the referees' reports. In this case, the editorial decision would be based on

$$\text{Final Editor's } c_S \text{ Inference} = \begin{cases} w_E \times r_E + (1-w_E) \times r_A & \text{with one referee} \\ \frac{w_E \times r_E + (1-w_E) \times (r_A + r_B)}{w_E + (1-w_E) \times 2} & \text{with two referees} \end{cases}$$

where w_E is the weight to be estimated; r_E, r_A, and r_B are generated as before (each with a lambda of 1/3, i.e., $r_R = \lambda c_S + (1-\lambda) c_R$ and $R \in (A, B, E)$); and editors continue with the top 18.1% of submissions based on their best final inference. When simulated with $w_E = 1/2$, which would have editors weigh their own signal as strongly as those of referees, Table 6 shows that this model predicts less consensus between editor and referee than was observed. For example, if editors had weighed their own signal and one REJ referee recommendation equally, they should have followed the referee only 92.8% and not the observed 95.7% of the time. Similarly, if they had weighted their own signal and one WEAK recommendation equally, they should have continued only 38.0% and not 50.4% of the time. If they had weighted their own signal and one R&R recommendation equally, they should have continued only 61.3% and not 82.7% of the time. Similarly, with two referees, editors tended to follow referees more often than would have been observed with equal-weight signals. However, excess following was not always the case. For example, editors did follow ACC recommendations roughly as would have been expected with equal-quality signals.

Overall, the behavior of editors is better explained by a model in which editors have less information or place less weight on their own information. The best fit is a model with $w_E \approx 0.35$. It predicts that only 3.8% of submissions with one single rejection recommendation would have been continued (overriding the referee), which actually happened in 4.2% of submissions; that editors should have continued with papers with one WEAK referee recommendation 42.2%

of the time; and that editors should have continued with papers with one R&R recommendation 77.8% of the time.[15]

It is an interesting question whether the editors at this journal preferred papers that elicit strong disagreement to papers that elicit broader but lukewarm support. There are too few observations in Table 6 to generalize, but it hints that editors preferred the latter: they continued with 85.7% of the papers that received two WEAK endorsements, but only with 57.5% of the papers that received one ACC and one REJ endorsement.

The journal also permitted me to look at whether editors prefer multiple referee reports (at least for some papers) because they may reduce the probability that a rejection will be appealed. That is, editors may well be aware of the difficult decision problems across papers with multiple referees, but deem it less important than reducing the number of appeals. The evidence does not suggest that this was the case. Papers with one referee were appealed 1.8% of the time. Papers with two referees were appealed 1.4% of the time. Papers with three referees were appealed 4.4% of the time. And papers with four or more referees were appealed 25% of the time. Thus, multiple referees did not serve, on average, as a "firewall" against appeals.[16]

3. Identifying Referee-Specific Factors

The previous section showed that referee-specific views are more important than common views in explaining referee recommendations, but it could not identify the sources of disagreement. The heterogeneity in recommendations could be based on paper-specific attributes. It is interesting to ask whether the heterogeneity in recommendations can be traced back to specific referee characteristics rather than submission characteristics. Unfortunately, the data about referees is limited—for example, I have no information on such referee attributes as their motives or views (Bayar and Chemmanur 2012). However, I do have some limited data from the SFS Cavalcade on how referees evaluated

[15] My findings are in line with speculations about the editorial process in other fields. Armstrong (1997) describes that editors in many disciplines seem to behave similarly. Editors concerned about fairness tend to treat referees' (categorical) recommendations as votes. This means that the most prestigious journals often do not end up publishing papers that receive mixed reviews. Kupfersmid and Wonderly (1994) summarize four empirical studies that led them to conclude that papers with mixed reviews are unlikely to be published. Munley, Sharkin, and Gelso (1988) and Marsh and Ball (1989) find that mixed-review papers have a low probability of being published. Simon, Bakanic, and McPhail (1990) find that a single negative recommendation at the *American Sociological Review* (ASR) from 1977 to 1982 often resulted in rejection. Similarly, Bakanic, McPhail, and Simon (1987), Beyer, Chanove, and Fox (1995), and (Blank 1991, Table 8) find that editors' decisions are highly predictable from the average categorical ratings of the reviewers. Finally, Simon, Bakanic, and McPhail (1986) find that the ASR agreed with authors on only 13% of appeals. Bayar and Chemmanur (2012) produce an academic model of the editorial decision. However, they also consider referee-specific biases that an editor can undo.

[16] Appeals are rare, which means that further inference is difficult. Nevertheless, it was surprising that for papers with more than one referee, there was no strong association between the decision to appeal and the average referee report. That is, papers that were rejected by all referees were about as often appealed as papers that had at least one favorable referee. (It was rare that single-refereed papers with only one very positive referee were rejected.)

themselves, how often they had opposite perspectives, and how generous they were to other submissions.

3.1 Self-reported referee characteristics

In the SFS Cavalcade, 460 referees answered questions about what kind of referee they considered themselves to be.

Of these 460 referees, 33 referees with 168 recommendations considered themselves generally to be "more critical," 91 referees with 428 recommendations considered themselves to be "less critical." The 168 self-declared more-critical referee reports suggested that only 3.6% of the submissions that were assigned to them fell into the "must-accept" (MA) category, and that only 24% fell into the "must-accept" or "should-accept" categories (MA/SA). The equivalent figures were 4.0% and 27% for 1,636 referee reports that considered themselves typical, and 4.4% and 40% for 428 referee reports that considered themselves less critical.[17]

Referees who considered themselves "more discriminating" provided 180 referee reports. These referees suggested that only 3.3% of papers were MUST ACCEPT and only 24% were MUST ACCEPT or SHOULD ACCEPT. The same figures were 3.9% and 29% for referees that considered themselves typical, and 4.8% and 32% for referees that considered themselves less discriminating. The evidence suggests that "discriminating referees" were more negative. Not reported, separate lambda estimations in the most-discriminating referee group and regression-interaction specifications suggest that when both referees considered themselves to be more discriminating, they tended to agree less with one another. However, the difference is not statistically or economically significant. Thus, it is safe to say that the evidence suggests that more discriminating referees were no more in agreement than other referees, but not that they were less in agreement.

The Appendix briefly describes some additional referee-specific attributes associated with their recommendations. For example, there is evidence that younger faculty (having receiver their PhDs more recently) and faculty from more prestigious universities were more negative.

In sum, the referees themselves proclaimed that their identities would influence their recommendations, above and beyond the submission itself, and these claims were backed up by data on their behavior.

3.2 Referee-specific fixed effects

To measure the magnitude of the extent to which scale effects (differences in referees' generosities) can help explain referee recommendations, it would be ideal to predict each referee's recommendation r_R with two variables: the

[17] Unfortunately, I had asked referees to rate themselves when they returned their reviews, not at the time when they specified their expertise. Thus, it is possible that their self-assessments were contaminated by the specific papers they refereed.

paper's true reliable characteristic c_S and the subjective generosity of referee R, which can be viewed as a mean shift m_R of her signal. Unfortunately, neither is observed. Therefore, my analysis explains referee recommendations with two noisy proxies: the recommendation of other referees on the same paper (which presumably reflects the reliable component of referee judgments about paper characteristics) and the referees' own mean recommendations on other unrelated papers (which presumably reflects the mean scales of the referees).

The analysis is most intuitive if we first estimate referee recommendations according to an admittedly arbitrary scale. For the SFS Cavalcade, I coded recommendations as –2 (MR), –1 (SR), 0 (NR), 1 (SA), and 2 (MA).[18] Table 7 shows the estimation results using different specifications. Specification (A) uses all referee pairs and predicts the recommendations of one referee with that of the other, using simple ordinary least squares (OLS) regression. The standardized OLS coefficients (in parentheses) that express both X and Y in terms of standard-deviation normalized quantities make it easy to interpret economic meaning. In (A), the standardized coefficient for the other referee's recommendations for a given paper (the "submission mean," the proxy for c_S) is 0.220. The standardized coefficient on this referee's mean recommendations on other papers (not including the current paper—i.e., a measure of the "referee generosity"—the proxy for m_R) is 0.157. This suggests that the common characteristic of the paper embodied in the submission mean is more important than the referee's mean (on other papers), but both are important. Because the correlation between the two independent variables is –2%, specifications A.1 and A.2 show that the two variables have almost the same coefficients by themselves.

As noted, both c_S and m_R are unobserved. A simulated method of moments estimation of the coefficients on a summary model, equivalent to Equation (1) and Table 5 but adding an additional term for the referee mean (also drawn from a unit-normal) that matches the OLS coefficients of model A ($r_R = \lambda_1 \times c_S + \lambda_2 \times m_R + (1-\lambda_1-\lambda_2) \times c_R$), can adjust both for noise caused by the discreteness in observations, and for the fact that the proxies are noisy. Such an estimation suggests that the observed coefficients would have obtained if referees had placed $\lambda_1 \approx 30.6\%$ weight on the common signal c_S, $\lambda_2 \approx 19.6\%$ weight on their own referee-mean m_R, and 49.8% weight on other referee-specific signals c_R.

Variations in specification, (B) through (G), show that the inference is robust. The coefficients are similar using only non-reversed pairs (i.e., if for a given paper, referee 1 is used to predict referee 2's recommendation, I do not predict referee 2's recommendation with referee 1), using a probit, using a logit, using

[18] Different numerical values for the categories, combinations of categories, and full set of X-dummies for different categories yield very similar results. The adjusted R^2 in the OLS regression is lower if I factor the other referees' recommendations into intercepts to avoid scaling. The coefficient ordering remains sensible, though. The coefficient on the referee's own mean on other papers is not affected.

Table 7
Predicting SFS Cavalcade referee recommendations

Method		R^2		Intercept	Other referees' rec on same refereed paper	Referee's own mean on other papers
24,354 pairs						
A	OLS	7.1%	Coef	–0.0780	0.2196 (0.220)	0.2745 (0.157)
.1	OLS	4.7%	Coef	–0.112	0.2163 (0.216)	
.2	OLS	2.3%	Coef	–0.110		0.2663 (0.152)
			(se)			(0.011)
				$\Rightarrow r_i = 0.498 \times c_A$	$+ 0.306 \times c_S$ (0.012)	$+ 0.196 \times m_i$ (0.014)
12,175 nonreversed pairs						
B	OLS	7.9%	Coef	–0.0784	0.2253 (0.220)	0.3089 (0.180)
			se		(0.009)	(0.016)
C	Probit	3.0% / 8.1%	Coef	–1.6/–0.3/0.5/1.7	0.246	0.349
			se		(0.010)	(0.016)
D	Logit	2.9% / 7.8%	Coef	–2.7/–0.6/0.9/3.1	0.425	0.591
			se		(0.017)	(0.029)
E	OLS	35.1%	Coef	Full Fixed (Paper) Effects		0.3440 (0.201)
			se			(0.014)
2,904 report means						
F	OLS	17.4%	Coef	–0.0030	0.6646 (0.386)	0.3114 (0.176)
			se		(0.028)	(0.031)
1,343 Non-repeated referees (1+2,3+4,...)						
G	OLS	7.7%	Coef	–0.0615	0.2190 (0.214)	0.3072 (0.182)
			se		(0.026)	(0.046)

Explanations: The predicted referee recommendations are for 2,903 referee recommendations from the SFS Cavalcade with data for both variables. For the OLS regressions, the standardized coefficient is in parentheses to the right of the plain coefficient. It multiplies the coefficient by the standard deviation of X over the standard deviation of Y. The OLS standard errors are White-heteroscedasticity adjusted. The Probit and Logit R^2 are the McFadden / Maximum-Likelihood R^2s, respectively. The equation below the A.2 model fits a model equivalent to Equation (1) and Table 5, but adds an additional term for the referee mean m_R (also drawn from a unit-normal) that matches the OLS coefficients of model A.
Interpretation: The mean of recommendations of unrelated papers by the same referee is an important marginal predictor of this referee's recommendation, although the m_R coefficient suggests that it is weaker than the effect of other referees' opinions on the same paper c_S. The statistical significance and economic meaning of the coefficients are largely unaffected by method.

only unique paper-referee combinations (i.e., for a given paper, I use referee pair 1-2, then 3-4, then 5-6, etc.), or using the mean report of other referees instead of the actual report. With thousands of observations, the coefficients are always statistically significantly different from zero or one.

Table 8 reports the results of OLS regressions analogous to model (F) from Table 7 for all venues. The reported coefficients are again standardized to allow the reader to focus on the relative economic significance. The findings are similar. The referee's intrinsic generosity was important in all venues. In fact, at the RFS, the JF, and the QJE (which here seems to follow the finance journal patterns), the OLS coefficient on the own-referee mean was twice as high as the coefficient on the submission mean: to get a good referee report, it was as

Table 8
Consensus and referee mean effects from different venues

Economics journals (all EE)

Journal	Other ref(s) same ppr	Own ref other ppr(s)
ECMTA	0.20	0.19
	(0.01)	(0.01)
JEEA	0.19	0.15
	(0.01)	(0.01)
JET	0.27	0.12
	(0.02)	(0.02)
QJE	0.15	0.32
	(0.01)	(0.01)
IER	0.24	0.18
	(0.01)	(0.01)
RAND	0.25	0.14
	(0.02)	(0.03)
Mean	0.22	0.19

Finance venues

Journal	Other ref(s) same ppr	Own ref other ppr(s)
RFS	0.14	0.29
	(0.02)	(0.02)
JF (not EE)	0.10	0.25
	(0.02)	(0.02)
SFS Cavalcade	0.386	0.176
(in Table 7 under "2,904 report means")	(0.02)	(0.02)

Explanations: These are the standardized coefficients and their standard errors (in parentheses) from OLS regressions explaining referee recommendations with two independent variables: the response of another referees to the same paper, and the referee's own mean response to other papers not under review. The latter can be largely viewed as fixed referee mean effects. The reported coefficients are analogs of the standardized coefficients in Table 7 under "2,904 report means."
Interpretation: Within the six economics journals, referee mean fixed effects are about as important in explaining a referee's recommendation as is a second referee's opinion on the same paper. Within the two finance journals, referee mean effects are even more (twice as) important. The SFS Cavalcade had the most agreement among referees relative to the referee fixed effects.

important to draw an intrinsically generous referee with a high m_R as it was to write a paper whose common reliable component c_S was high. The relative importance of the two factors mattered only mildly across journals. The referee draw was relatively less important at the less selective journals in my sample. In particular, the coefficient on the submission mean at the JEEA, JET, IER, and RAND was higher than the coefficient on the referee mean. In addition, the one venue in which the coefficient on the "other referees' mean on the same submission" was highest relative to the coefficient on the "referees' means on other papers" (0.386 versus 0.176) was the SFS Cavalcade. This suggests that the average recommendation at the SFS Cavalcade reflected relatively more on the paper and relatively less on the referees. However, this was the case primarily because there were more referee recommendations for each paper. Table 7 suggests that under a variety of "only-one-other-referee" specifications, the SFS Cavalcade coefficients were more similar to those observed at the other venues.

In sum, it is safe to conclude that a referee report reflects about as much on this referee's intrinsic generosity as it reflects on what another referee would say about the paper. If an editor wants to extract the reliable common component c_S from the referee reports, it would seem important to adjust referees' recommendations for their average recommendations on earlier submissions.

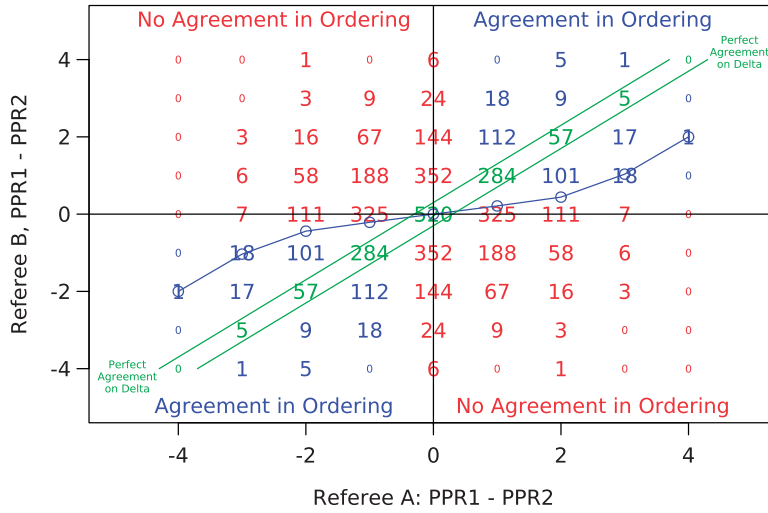
3.3 Scale differences and preference reversals

Referees could have differed because they had different scales (referee-specific means or standard deviations) or because they disagreed fundamentally about which paper was better. In the SFS Cavalcade, there were many referees who shared assignments for the same two papers. (This was because assignments were expertise-based.) I can therefore examine whether referees agreed or disagreed on the relative ordering of the same two papers. For each of the two referees, A and B, I calculate the difference between the referees' two paper recommendations. I then count the instances defined by the difference for A and the difference for B for each overlapping paper.

Figure 3 is a two-dimensional representation of this three-dimensional histogram. (There is symmetry in this graph, because I repeat the calculation with A and B reversed.) The graph suggests the following:

- On the 45-degree diagonal, there are 1,212 cases out of 4,416 pairs in which referees agreed exactly on the relative difference between two papers. This includes 456 cases in which referees were in perfect agreement about the qualities of the two papers. It also includes 756 cases in which referees agreed about the relative qualities but had different mean scales (thresholds). For example, one referee could have ranked the papers "must accept" (2) and "neutral" (0), while the other could have ranked them as "neutral" (0) and "must reject" (–2).
- In the first and third quadrants, but off-diagonal, referees shared the same relative ordering of the papers, with both preferring the same paper, but they could not agree on how much better one paper was relative to the other. One referee ranks one paper much more highly relative to its competitor paper than the other referee (e.g., the eighteen papers at coordinate +1,+3). These cases can be viewed as referees having different scales (possibly on multiple moments) in translating their paper views into numerical scores. Excluding cases in which one referee ranked the quality difference in both papers the same, there were 972 pairings in which referees agreed at least on the relative ordering of two papers. Including these cases increases the agreement to 2,093 pairings.
- In the second and fourth quadrants, referees had different preference orderings. There were 702 cases in which one referee strictly preferred one paper, and the other referee strictly preferred the other paper.

The presence of cases in which referees disagreed about the ordering of papers rejects the hypothesis that referee recommendations are completely due to differences in scales. Because 75% of recommendations shared the relative ordering, while 25% reversed it, there is, however, also more scale agreement than what would have been observed under chance.



Total pairs (duplicated)	4,416
Same distance between papers	1,121
Identical recommendations	456
Same difference, but offset	756
Incl 0 diffs, same orderings	2,093
Excl 0 diffs, same orderings	972
Opposite orderings	702

Figure 3
Relative pairwise ordering of shared referee assignments at the SFS Cavalcade
Explanations: This is a two-dimensional representation of a three-dimensional histogram depicting the relative ordering when two referees evaluated the same two papers. For each referee, I calculate the difference between their two paper recommendations. On the 45-degree diagonal, referees agreed exactly on the relative difference between two papers. In the first and third quadrant, referees shared the same ordering of the papers. The 1,212 pairs on the diagonal shared the same numerical difference in relative ordering. In 456 pairs, referees had the same exact recommendations on both papers; in 756 pairs, one referee had a higher recommendation, but both referees agreed on the numerical distance between the two papers. This is a pure mean effect. In the second and fourth quadrant, referees have different preference orderings.
Interpretation: Referees had modest agreement about relative pairwise ordering. However, preference reversals were common.

4. Related Literature

The peer-review literature is voluminous and spans many disciplines. Cole (1992), Weller (2002), and Armstrong (1997) summarize it. These existing studies suggested that reviewers are enamored with (i) statistical significance, (ii) large sample sizes, (iii) complex procedures, and (iv) obscure writing (verbiage). Reviews do not seem to improve readability and catch many errors, such as in citations, descriptions of the content of prior research, basic abuse of statistics, plagiarism, or previous publication.

Armstrong's most important concern is with the clash between innovative research and existing belief systems, as well as journals' concerns about

"quality" over "importance." (Ellison 2002a,b discuss how the publication process has slowed down over the years due to increased polishing and revising.) Armstrong describes various experimental papers that show that scientists rate controversial research more highly if it corresponds to their own beliefs. Indeed, controversial findings almost never receive unanimous support, and other evidence shows that editors rarely publish papers with conflicting reviews. (The evidence in my own paper similarly suggests that one negative review usually results in rejection.) Armstrong concludes that the barriers to publication of controversial or important research may be too great to make it worthwhile to work on such issues. Similarly, Gans and Shepherd (1994) solicited informal but highly entertaining and troubling descriptions of the publication process for now-classic economics papers. Armstrong even notes a silver lining: the lack of reliability in reviews improves chances of publishing papers with innovative findings, but only if such papers are not evaluated by many referees. Lack of reliability then becomes a useful randomizing device.

The most relevant related literature to my paper are articles that study in quantitative terms the reliability (agreement) of referee accept/reject recommendations and the fixed-effect and other biases of referees. This literature is much smaller. Cole, Cole, and Simon (1981) find that a second set of National Science Foundation (NSF) proposal reviewers for 150 proposals had only modest agreement with the first (regular) set of NSF reviewers. Peters and Ceci (1982) resubmitted twelve recently published psychology papers under fictitious names and institutions to the same journal again. They summarize their findings as follows: "Of 38 editors and reviewers, only three (8%) detected the resubmissions. This result allowed nine of the 12 articles to continue through the review process to receive an actual evaluation: eight of the nine were rejected. Sixteen of the 18 referees (89%) recommended against publication and the editors concurred. The grounds for rejection were in many cases described as 'serious methodological flaws."' Fiske and Fogg (1990) found that in the typical paper, two reviews of the same paper had no critical point in common. Blank (1991) provides some evidence of referee consensus and similarly finds only modest association. Cherkashin et al. (2009) study the citation impact of 21 coeditors at the *Journal of International Economics* (JIE) and document editor fixed effects.

Two current papers examine the journal review process in economics. The first is Brogaard, Engelberg, and Parsons (2011), which looks at how editor rotations associate with increased publication rates of "nearby" authors. By showing that these papers are also more cited in the long run, the evidence rejects a nepotism hypothesis in favor of an informedness hypothesis. The second is Card and DellaVigna (2013), which looks at how a recent policy change in maximum submission paper lengths can be used to measure the elasticity of authors to submit papers to the AER and to the JEEA. The AER seems to have enough market power to induce authors to change their drafts to qualify for submission, while the JEEA does not.

5. Conclusion

The editorial process sets not only the direction of economics as a social science but also the incentives and professional fates of academic economists. It consumes significant resources, both in terms of the time and effort by the editor and referees, and in terms of the time of authors adjusting their papers to make them more amenable to publication. Gans and Shepherd (1994) write about how the vagaries of the process have annoyed even the most distinguished economists—and on occasion editors Spiegel (2012).

The evidence in my paper quantified to what extent referee recommendations are influenced by referee-specific idiosyncratic components and by common reliable characteristics of the submissions (including author-related information). The fact that the referee behavior is similar in the editorial-selection journal process and in the automated objective expertise-based referee-selection process suggests that the documented patterns are more likely to be due to referees themselves than to referee selection. The evidence further suggests that the referee-specific component is partly due to the fact that some referees are intrinsically more generous than others, and partly due to the fact that referees cannot agree with one another about how good papers are relative to one another.

Suggestions to improve the refereeing process are always controversial. In fact, it is not even clear that it is desirable to increase the consistency and reliability of reports across referees. Reliability is not necessarily a measure of the quality or the future impact of the paper. (Gans and Shepherd note that many of the [previously] "rejected classic articles" were ultimately published by editors against the advice of the referees.) Nevertheless, my paper suggests a number of changes where the primary cost amounts to little more than a change in the editorial web system:

- It is not clear whether the low or zero active experimentation (common at many journals) is an optimal choice. My paper had to rely on indirect evidence from a non-journal venue to suggest that referee selection effects do not seem to have influenced greatly findings about referee behavior. Journals could gain better and more relevant information if they themselves experimented. If the current process is optimal, then small perturbations from the currently-optimal process would not be very costly.

  For example, it would be interesting to ask some referees for some additional quantifiable information, such as (i) whether they believe that another referee would likely agree or disagree with their assessment of the reviewed paper, and/or (ii) whether another referee would likely give the submission a higher or lower assessment. Lynch (1998) points out that most referees have a distorted view of how much their views represent a general consensus of their peers. Do they?

- The widely used current editorial systems do not provide editors with information about referees' past recommendation behavior. The benefit of an enhanced system with this information easily on hand is that it could make it easier for the editor to extract the referee-specific non-paper-specific signal in the referee report.
- Journals could provide more feedback to referees. Specifically, they could provide automatic notification of what other referees recommended on the same submission. If it were the default to inform every referee of the final recommendation of other referees, it would not place editors in the awkward position of selectively informing referees about how different their views were.
- Editors could "even out" the number of referees per paper. The cost is that such homogeneity reduces the flexibility of editors to deal with papers and authors that are different. The benefit is that a homogeneous number of referees reduces the very difficult problem of extracting the reliable component from (and deciding fairly across) papers with different numbers of referees.

Appendix

Related Findings

The SFS Cavalcade also included a set of additional variables that allowed some more analysis of referee-related behavior patterns. The following inference explaining referee recommendations and consensus[19] among referees is based on simple correlations between variables provided by referees and authors themselves.

Execution quality and interestingness: (Mean) Referees valued execution quality and degree of interesting-ness of the paper roughly equally. (Consensus) Referees had no more agreement about whether papers were interesting than about whether they were well executed.

Author age: (Mean) There was no meaningful difference in referee recommendations for submissions by author teams that were younger or older, where age was measured either as the average year of the PhD or as the age of the PhD of the oldest member.

Author university rank: (Mean) When there was at least one author from a highly ranked university,[20] the probability of the paper receiving an MA was 10.1%. When there was none, it was 2.2%. (Consensus) The lambda statistics for agreement among referees were 0.12 and 0.18, respectively. That is, referees agreed more about papers written by authors from lower-ranked universities.

This contrasts with the findings of Peters and Ceci (1982), who found in a quasi-experiment that reviewers in psychology journals were biased against authors from less prestigious

[19] The consensus inference is based either on lambda estimates in a subset of papers or on cross-variables explaining one referee's recommendation with that of another multiplied by the additional variable

[20] This is a dummy largely based on the ranking in the list of top Arizona PhD program list http://wpcarey.asu.edu/fin-rankings/rankings/results.cfm: NYU, Chicago, Harvard, Penn, UCLA, Michigan, Duke, OSU, Columbia, Northwestern, MIT, Stanford, Cornell, Texas, USC, LBS, Illinois, BC, UNC, Berkeley, Maryland, ASU, Hong Kong, Purdue, Rochester. I added LSE, Princeton, and Yale.

institutions. Perlman (1982), however, argues that authors at higher-ranked institutions write more cited papers, which *should* influence the reviewing process. (He equates citation impact with quality.)

Author U.S. university: (Mean) When there was at least one U.S. author on the team, the probability of an MA was 5.7%. When there was none, it was 1.5%. (Consensus) The agreement among referees about submission quality was similar.

Other referee recommendations: There is some evidence that I could have predicted even better with a more complex function than the referee's mean. When a referee had not offered even one single other high recommendation, such a referee was marginally more likely to recommend an acceptance on the current paper. However, the effect is not strong and could easily have been the result of specification search.

Refereeing frequency: (Mean) Referees who reported to have refereed more papers and accepted more papers in the past (elsewhere) were less likely to accept a paper for the conference. (Consensus) There was modestly greater consensus among referees with more past refereeing.

Referee age: There was no clear pattern related to the referee's age (where age is PhD cohort).

Referee PhD university rank: Referees with PhDs from highly ranked universities were modestly more critical (mean of –0.18), but showed less consensus than their complement.

Referee university rank: Referees currently at highly ranked universities were more critical (mean of –0.29) and showed modestly more consensus than their complement (0.28 versus 0.23).

References

Armstrong, J. S. 1997. Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics* 3:63–84.

Bakanic, V., C. McPhail, and R. J. Simon. 1987. The manuscript review and decision-making process. *American Sociological Review* 52:631–42.

Bayar, O., and T. J. Chemmanur. 2012. A model of the editorial process in scientific journals. Working Paper, Boston College.

Beyer, J. M., R. G. Chanove, and W. B. Fox. 1995. The review process and the fates of manuscripts submitted to AMJ. *Academy of Management Journal* 38:1219–60.

Blank, R. M. 1991. The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *American Economic Review* 81:1041–67.

Brogaard, J., J. Engelberg, and C. A. Parsons. 2011. Network position and productivity: Evidence from journal editor rotations. Working Paper, University of Washington and University of California, San Diego.

Card, D., and S. DellaVigna. 2013. Nine facts about top journals in economics. *Journal of Economic Literature* 51:144–61.

Cherkashin, I., S. Demidova, S. Imai, and K. Krishna. 2009. The inside scoop: Acceptance and rejection at the journal of international economics. *Journal of International Economics* 77:120–32.

Cole, S. 1992. *Making science: Between nature and society*. Cambridge, MA: Harvard University Press.

Cole, S., J. R. Cole, and G. Simon. 1981. Chance and consensus in peer review. *Science* 214:881–86.

Cornell, B., and I. Welch. 1996. Culture, information, and screening discrimination. *Journal of Political Economy* 104:542–71.

Ellison, G. 2002a. Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy* 110:994–1034.

———. 2002b. The slowdown of the economics publishing process. *Journal of Political Economy* 110:947–993.

Fiske, D. W., and L. Fogg. 1990. But the reviewers are making different criticisms of my paper. *American Pscyhological Association* 45:591–8.

Gans, J. S., and G. B. Shepherd. 1994. How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives* 8:165–79.

Kupfersmid, J., and D. M. Wonderly. 1994. *An author's guide to publishing better articles in better journals in the behavioral sciences*. Brandon, VT: Clinical Psychology Publication Company.

Lynch, Jr., J. G. 1998. Presidential address reviewing. *Advances in Consumer Research* 25:1–6.

Marsh, H. W., and S. Ball. 1989. The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *Journal of Experimental Education* 57:151–69.

Munley, P. H., B. S. Sharkin, and C. J. Gelso. 1988. Reviewer ratings and agreement on manuscripts reviewed for the journal of counseling psychology. *Journal of Counseling Psychology* 35:198–202.

Perlman, D. 1982. Review "bias": Do Peters and Ceci protest too much? *Behavioral and Brain Sciences* 5:231–2.

Peters, D. P., and S. J. Ceci. 1982. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences* 187–95.

Simon, R. J., V. Bakanic, and C. McPhail. 1986. Who complains to editors and what happens. *Sociological Inquiry* 56:259–71.

———. 1990. If at first you don't succeed: Review procedures for revised and resubmitted manuscripts. *American Sociologist* 21:373–91.

Spiegel, M. 2012. Reviewing less-progressing more. *Review of Financial Studies* 25:1331–8.

Weller, A. C. 2002. *Editorial peer review: Its strengths and weaknesses*. American Society for Information Science and Technology.