A Kuhnian Critique of Psychometric Research on Peer Review

Carole J. Lee*

Psychometrically oriented researchers construe low interrater reliability measures for expert peer reviewers as damning for the practice of peer review. I argue that this perspective overlooks different forms of normatively appropriate disagreement among reviewers. Of special interest are Kuhnian questions about the extent to which variance in reviewer ratings can be accounted for by normatively appropriate disagreements about how to interpret and apply evaluative criteria within disciplines during times of normal science. Until these empirical-cum-philosophical analyses are done, it will remain unclear the extent to which low interrater reliability measures represent reasonable disagreement rather than arbitrary differences between reviewers.

1. Introduction. The incentives, review processes, and norms of peer review provide the foundation for creating and sustaining the piecemeal contribution and critique of knowledge in scientific communities (Ziman 1969; Merton and Zuckerman 1971; Lee 2012). In light of its central role in determining locally the content of science (Hull 1988, xii), scientists and social scientists have taken it on themselves to undertake "hypothesis-testing research on peer review and editing practices" (Fletcher and Fletcher 1997, 38). There is a growing industry of social scientific research on editorial peer review, about a third of which comes from psychology, and another third, from medicine (Weller 2001, 10).

Of the empirical research available on peer review, one of the "most basic, broadly supported, and damning" aspects is the failure for independent expert reviewers to achieve acceptable levels of agreement in reviews for journals and grant proposals across the physical, social, and human sciences as well as the humanities (Marsh, Jayasinghe, and Bond 2008, 161). I will review this literature and discuss the reflexive felicity of psychometric re-

*To contact the author, please write to: Department of Philosophy, University of Washington, Box 353350, Seattle, WA 98195; e-mail: c3@uw.edu.

Philosophy of Science, 79 (December 2012) pp. 859–870. 0031-8248/2012/7905-0003$10.00

TABLE 1. Single-Rater Reliability for Grant Review

	Single-Rater Reliability
National Science Foundation:	
Chemical dynamics	.25
Solid-state physics	.32
Economics	.37
Australian Research Council:	
Social science and humanities	.21
Physical sciences	.19

Sources.—Cicchetti (1991) and Jayasinghe, Marsh, and Bond (2003).

searchers imposing on themselves as peer reviewers the same standards that they would impose on the content of their research: in particular, high interrater reliabilities.

I will then argue that equating low interrater reliabilities with the invalidity of peer review as a test overlooks the ways in which low interrater reliabilities might reflect reasonable forms of disagreement among reviewers. Although this research focuses on the acceptance of papers and grant proposals as opposed to the acceptance of theories, I will argue that Kuhnian observations about how the definition of epistemic values underdetermine their interpretation and application suggests new empirical hypotheses and philosophical questions about the kinds of reviewer disagreement we would expect to find. It remains an open empirical and philosophical question the extent to which these might account for and rationalize low interrater reliability rates. Still, low interrater reliability rates remain problematic insofar as they cause individual peer review outcomes to result from the "luck of the reviewer draw" (Cole, Cole, and Simon 1981, 885). To close, I will discuss some of the discipline-wide communication structures that can help accommodate low interrater reliability rates. This discussion makes light of less obvious ways in which peer review constitutes a social epistemic feature of the production and communication of knowledge.

2. Interrater Reliability of Expert Reviewers. High correlations between mean reviewer recommendations and final decisions by editors and grant panels suggest that reviewer recommendations are taken very seriously (Cole et al. 1981; Bakanic, McPhail, and Simon 1987; Marsh and Ball 1989; Hargens and Herting 2006). Measures of the intraclass correlation between ratings for two reviewers on a single submission, or the *single-rater reliability* of reviewers, have been found to be quite low. Table 1 presents results from studies on single-rater reliability rates for grant review across disciplines.

The finding that reliability measures for reviews in the physical sciences were not better than those in the social sciences and humanities is quite sur-

https://doi.org/10.1086/667841 Published online by Cambridge University Press

TABLE 2. SINGLE-RATER RELIABILITY FOR SPECIFIC EVALUATIVE CRITERIA

	Single-Rater Reliability
Australian Research Council:	
Originality	.17
Methodology	.15
Scientific/theoretical merit	.16
Educational Psychology:	
Significance	.12
Research design	.23
Clarity of problem, hypothesis, assumptions	.22
Journal of Personality and Social Psychology:	
Importance of contribution	.28
Design and analysis	.19

Sources.—Scott (1974), Marsh and Ball (1989), and Jayasinghe (2003).

prising since one might expect less consensus in disciplines with less developed research paradigms (Beyer 1978; Lindsey 1978). Single-rater reliabilities are also comparably low for reviews of manuscripts submitted to top journals such as *American Sociological Review*, *Physiological Zoology*, *Law and Society Review*, and *Personality and Social Psychology Bulletin* (Hargens and Herting 1990b). Even more interesting is research demonstrating low interrater reliabilities for specific evaluative criteria, as presented in table 2.

3. Psychometric Assumptions. Psychometric approaches to studying peer review construe the low agreement between expert reviewers to be deeply problematic.[1] This contention rests on a few key psychometric assumptions. The most fundamental assumption is that submissions have a latent overall quality value along a single dimension of evaluation. Indeed, this assumption is built into the measurement of interrater reliability that measures disagreement along a single, ordinal point scale (used in the evaluation of grant proposals) or by using coefficients requiring the assignment of arbitrary scores to recommendation categories (e.g., "accept," "revise and resubmit," "reject") or to distances between categories (used in the evaluation of journal submissions; Hargens and Herting 1990b, 1). The assumption that there is such a dimension of value is commonplace within psychometrics, which re-

1. Construct validity, the correlation between findings of tests seeking to assess the same property or construct, is also necessary in order to have a valid test or process. Researchers have challenged the construct validity of expert reviewer ratings by measuring how they compare with different proxy measures for manuscript quality (Lee et al., forthcoming).

https://doi.org/10.1086/667841 Published online by Cambridge University Press

TABLE 3. Single-Rater Reliability for Psychometric Tests

	Single-Rater Reliability
Intelligence tests	>.9
Personality tests	>.7
Essay marking	~.6
Rorschach inkblot test	~.2

Source.—Rust and Golombok (2009).

lies heavily on measuring hypothetical entities or constructs, such as intelligence or creativity, along a single dimension (Rust and Golombok 2009, 31).

This turns the role of expert reviewers into identifying the latent quality value of a submission along this single dimension of evaluation with a high degree of reliability and, thereby, interrater reliability (Hargens and Herting 1990a, 92). How high is high? Within the field of psychometrics, different types of tests have different levels of interrater reliability levels (Rust and Golombok 2009, 75–76), as shown in table 3. Some psychometrically oriented researchers suggest that levels of interrater reliability for expert reviewers should be at about 0.8 (or even 0.9; Marsh et al. 2008, 162). Unfortunately, interrater reliability for expert reviewers is perilously close to rates found for Rorschach inkblot tests.

From a psychometric perspective, if we assume that the raters are not in need of retraining, a test with too low a level of interrater reliability is considered invalid—that is, it cannot be said to measure what it purports to measure (Rust and Golombok 2009, 72). According to psychology's own disciplinary standards for valid testing, peer review is a "poor" evaluation tool (Bornstein 1991, 444–45; Suls and Martin 2009, 44). This makes the "critical direction for future research" to be that of improving "the reliability of peer reviews" (Jayasinghe, Marsh, and Bond 2003, 299). Along these lines, Jayasinghe et al. found that some of the variance in ratings resulted from biases related to characteristics of the reviewer: North American reviewers tended to give higher review ratings than Australian ones, reviewers nominated by the researcher gave higher ratings than those nominated by the grant panel, and scientists who reviewed fewer proposals gave higher ratings than those who reviewed more proposals (Jayasinghe et al. 2003). However, these statistically significant biases do not account for very much of the variance in reviewer ratings. Increasing the number of reviewers per proposal (4.1 for social sciences and humanities and 4.2 for science) increased single-rater reliability measures to ~0.47 (Jayasinghe et al. 2003). However, this measure is still low, falling between rates found for essay graders and Rorschach inkblot tests. The obvious empirical conundrum is to figure out what can account for the rest of the variance in reviewer ratings.

4. Interrater Reliability and Normatively Appropriate Disagreement. There is a reflexive felicity in psychometric researchers imposing on themselves as peer reviewers the same methodological standards (i.e., high interrater reliabilities) that they impose on the content of their research.[2] However, equating low interrater reliabilities with the invalidity of peer review as a test overlooks the ways in which low interrater reliabilities might reflect reasonable forms of disagreement among reviewers. When we shift focus from the numerical representation of a reviewer's assessment to the content on which such assessments are grounded, we can identify cases in which interrater disagreement reflects normatively appropriate differences in subspecialization, as well as normatively appropriate differences in the interpretation and application of evaluative criteria.

Differences in subspecialization and expertise can lead to low interrater reliabilities. Editors might choose reviewers to evaluate different aspects of a submission according to their subspecialization or expertise. For example, some reviewers might be sought for their theoretical expertise, while other reviewers might be sought for their technical expertise in, for example, statistics, modeling, or a special sampling technique (Hargens and Herting 1990a, 94; Bailar 1991). Additional reviewers might be sought to review the domain-specific application of those techniques. In cases in which quality along these different aspects diverges, we would not expect high interreviewer reliability scores (Hargens and Herting 1990a, 94). It is normatively appropriate for editors and grant panels to rely on differences in reviewer expertise in the evaluation of a submission. Note that, in these cases, the discrepancy between reviewer ratings does not reflect disagreements about the same content since reviewers are evaluating different aspects of the research.

There are other cases that can involve more direct disagreement between reviewers. Reviewers can disagree about the proper interpretation and application of evaluative criteria. This possibility may have been overlooked because of long-standing work suggesting expert agreement about evaluative criteria within and across disciplines. Studies on editors of journals in physics, chemistry, sociology, political science, and psychology discovered strong agreement within disciplines about the relative importance of different criteria in the evaluation of manuscripts (Beyer 1978; Gottfredson 1978). And surveys of editors for the top physical, human, and social sciences journals ($n = 73$) indicate agreement most especially about the importance of the significance, soundness, and novelty of submitted manuscripts (Frank 1996).

2. Reflexive critique is a leitmotif throughout empirical research on peer review. For example, research on peer review's impact for medical manuscripts has begun to undertake double-blind, randomized, controlled trials to identify, for example, the effects of masking identities of authors and reviewers on proxy measures for the quality of reviews (Godlee, Gale, and Martyn 1998; Justice et al. 1998; van Rooyen et al. 2010).

Editor opinions about the relevant criteria of evaluation are important since, in 92.5% of the cases, reviewers receive forms with instructions about evaluating manuscripts along these dimensions.

However, there are reasons to think that interdisciplinary and disciplinary agreement about evaluative criteria lies only on the surface. Lamont's interviews of interdisciplinary grant panelists show that disciplines attach different meanings to evaluative criteria such as originality and significance (Lamont 2009). Quantitative sociological research on discipline-specific publication biases corroborates her insights about how differently these criteria are interpreted and applied across disciplines. Consider, for example, the ubiquitous quest for novelty. In medicine, the interest in novelty is expressed as the preference for results of randomized, controlled trials that favor a new therapy as opposed to the standard therapy (Hopewell et al. 2009). In contrast, for social and behavioral scientists, the emphasis on novelty gets expressed as a preference for new effects over replications or failures to replicate an existing effect (Neuliep and Crandall 1990), regardless of whether these effects constitute an "improvement" in normative outcome.[3]

We have Kuhnian reasons to think experts within disciplines might disagree about how best to interpret and apply evaluative criteria. Recall Kuhn's observation that how different scientific values are applied in the evaluation of competing theories is underdetermined by their definitions and characterizations (1977, 322). Likewise, evaluative criteria in peer review are not sufficiently characterized to determine how they are interpreted and applied in the evaluation of papers and projects. Just as two scientists agreed about the importance of accuracy can disagree about which theory is more accurate, two expert reviewers agreed on the importance of novelty can disagree about whether a peer's paper or project is novel. This is because scientists and expert reviewers can come to different antecedent judgments about the significant phenomena or respects in which a theory or submission is thought to be accurate or novel. These Kuhnian considerations challenge the ideal that peer review is impartial in the sense that reviewers see the relationship of evaluative criteria to submission content in identical ways (Lee and Schunn 2011; Lee et al., forthcoming). This is a basic theoretical problem about value-based evaluations that applies, not just in the interdisciplinary contexts Lamont studies, but in disciplinary contexts as well.

5. Empirical and Normative Questions: Kuhnian Considerations. An empirical hypothesis we might propose in light of the Kuhnian considera-

3. This is not to say that all research programs within the behavioral sciences are neutral with respect to whether there ought be a preference for discovering effects involving normatively desirable as opposed to normatively undesirable outcomes (Lee 2008).

tions just raised is that experts can have diverging evaluations about how significant, sound, or novel a submitted paper or project is because they make different antecedent judgments about the relevant respects in which a submission must fulfill these criteria.[4] So far, current empirical research corroborates this kind of empirical hypothesis. Quantitatively, if the hypothesis were true, we would expect low interrater reliabilities along evaluative dimensions, as researchers have discovered (Scott 1974; Marsh and Ball 1989; Jayasinghe 2003). Qualitatively, if this hypothesis were true, we would expert reviewers to focus on different aspects in the content of reviews: their focus on different features of the work, by the Gricean maxim of relevance (Grice 1989), would suggest that they take different aspects of the work to be most relevant in evaluations of quality. Qualitative research corroborates the suggestion that reviewers focus on different aspects of research. An analysis of reviewer comments from more than 400 reviews of 153 manuscripts submitted to American Psychological Association journals across a range of subdisciplines found that narrative comments offered by pairs of reviewers rarely had critical points in common, either in agreement or in disagreement. Instead, critiques focused on different facets of the paper (Fiske and Fogg 1990). A different study found that comments from reviewers who recommended rejecting papers that went on to become Citation Classics or win Nobel Prizes claimed that manuscripts failed to be novel or significant in what reviewers took to be relevant ways (Campanario 1995).

This last example forcefully raises the normative question of whether disagreements about the interpretation or application of evaluative terms should always be counted as appropriate. In times of revolutionary science, these forms of disagreement may be normatively appropriate in the sense that they are reasonable in light of available evidence and methods. As Kuhn observed, the available evidence for competing theories during scientific revolutions is mixed, where each theory has its own successes and failures (1977, 323). Members of opposing camps prefer one theory or approach because they identify as most significant the specific advantages of their theory and the specific problems undermining the competing one, although there are no evidential or methodological means at the time to establish which aspect is most relevant or crucial.

In light of Kuhn's observations, we would expect reviewers in different camps—with different beliefs about what constitute the most significant advantages or disadvantages of competing theories—to have diverging opin-

4. Of course, there may be no respects in which some manuscripts or projects can be said to be significant, sound, or novel. This might partly explain the finding of increased interrater agreement for cases of rejection than acceptance in peer review of grant (Cicchetti 1991) and journal (Hargens and Herting 1990b) submissions.

ions about how significant, sound, or novel a submitted paper or project is. As a result, we would expect reviewers in different camps to arrive at reasonable disagreements about the quality of a particular submission.[5] If editors were to adopt the strategy of choosing expert reviewers from competing camps and mixed these evaluations with those by neutral referees, we would expect low correlations between reviewer ratings (Hargens and Herting 1990a, 94).

However, in periods of normal science, it is unclear whether disagreements about what features of a submission should be counted as most relevant are reasonable in these ways. Philosophical analysis of peer reviews should be undertaken to evaluate this question. By making this suggestion, I am not defending the claim that peer review, as it is currently practiced, functions as it should. Nor am I denying that normatively less compelling factors might contribute to low interrater reliability measures.[6] I am simply suggesting new lines of empirical and philosophical inquiry motivated by Kuhnian considerations.

Further empirical and philosophical analysis should be undertaken to measure the extent to which the variance in reviewer ratings can be accounted for by reasonable and unreasonable disagreements of various kinds. Until these empirical-cum-philosophical analyses are done, it will remain unclear the extent to which low interrater reliability measures represent reasonable disagreement rather than arbitrary differences between reviewers.

Psychometrically oriented researchers might suggest an alternative research program that would accommodate reasonable disagreement among reviewers while preserving the idea that low interrater reliability measures (of some kind) render peer review a poor/invalid test for assessing the quality of submissions. Under this refined research program, the task would be to evaluate peer review's well functioning by measuring interreviewer reliability among editors rather than reviewers. After all, it would be reasonable for editors to improve the quality of pooled reviews by choosing reviewers with diverging expertise and antecedent judgments about the significant respects in which a submission should be understood as novel, sound, or significant. This shifts the locus of relevant expert agreement to the editorial rather than the reviewer level.

Unfortunately, there is little to no research on intereditor reliability rates. Note, however, that Kuhnian concerns recur at the editorial level: editors

5. This is not to suggest that reviewers always deem research by allied authors to be sufficiently strong as to merit publication (Hull 1988, 333).

6. Some of these concerns include failures to catch basic statistical mistakes and omissions (Gardner and Bond 1990; Gore and Jones 1992) and short periods of time allotted by reviewers per manuscript across disciplines (Weller 2001, 156).

could disagree with each other about the relevant respects in which a submission should be considered novel, sound, or significant. Along these lines, sociologists Daryl Chubin and Edward Hackett suggest that the editor's task is its own kind of "Rohrschach [*sic*] test," where "both the article and the referee's interpretation are for the [editors] to weigh or discard as they see fit" (Chubin and Hackett 1990, 112).

6. Social Solutions to the Luck of the Reviewer Draw. Regardless of whether we discover reasonable forms of disagreement among reviewers, decisions to accept or reject submissions must be made. Even if the considerations raised by disagreeing reviewers are not arbitrary, low interrater reliabilities can make peer review outcomes an arbitrary result of which reviewer perspectives are brought to bear. Jayasinghe et al. found that the decision to fund a grant submission "was based substantially on chance" since few grant proposals were far from the cutoff value for funding when 95% confidence intervals were constructed for each proposal (Marsh et al. 2008, 162). Cole et al. found that the mean ratings of their newly formed panel of expert reviewers differed enough from the mean ratings of the actual National Science Foundation reviewers that, for about a quarter of the proposals, the funding decisions would have been reversed. They concluded that actual outcomes depend to a large extent on the "luck of the reviewer draw" (1981, 885).

These observations raise important questions about how discipline-wide publication venues should be structured to accommodate these kinds of problems. Hargens and Herting argue that the number of prestigious outlets for publication within a discipline, as well as the thresholds set for these outlets, play an overlooked but crucial role in addressing low interrater reliability rates (1990a, 102–3). Disciplines with very few "core" journals serving as high-prestige outlets (e.g., *Astrophysical Journal* in astronomy and astrophysics and *Physical Review* in physics) are more vulnerable to the possibility of relegating important work to less prestigious and less visible journals as a result of the luck of the reviewer draw. However, these disciplinary journals "minimize this threat" by accepting the great majority of submissions (75%–90%). In contrast, disciplines like psychology and philosophy with many core journals allow for more chances for important work to find a prestigious venue through an iterative process of submission and review. These journals can afford to have substantially higher rejection rates.

The obvious empirical question is whether these considerations can rationalize the large and stable differences observed in acceptance rates in the social sciences and humanities (about 10%–30%) versus the physical sciences (about 60%–80%; Zuckerman and Merton 1971; Hargens 1988, 1990). Structural accommodations of this kind might be constrained by discipline-specific goals, norms about whether to risk accepting bad research

versus rejecting good research (Cole 1992), and norms about whether time and future research (as opposed to peer review) should serve as the central filter for assessing quality (Ziman 1969).

Peer review clearly constitutes a social epistemic feature of the production and dissemination of scientific knowledge. It relies on members of knowledge communities to serve as gatekeepers in the funding and propagation of research. It calls on shared norms cultivated by the community. And it relies on institutions such as journal editorial boards, conference organizers, and grant agencies to articulate and enforce such norms. However, in light of research on low interrater reliabilities and the role that discipline-wide communication structures can serve to address the luck of the reviewer draw, it is clear that we should also analyze and evaluate peer reviewer's social epistemic function within larger communication structures to identify how these structures can accommodate reviewer disagreement.

7. Conclusion. Reflecting on the various ways in which epistemic values can be interpreted and applied, Kuhn suggested that "essential aspects of the process generally known as verification will be understood only by recourse to the features with respect to which" researchers "may differ while still remaining scientists" (1977, 334). It remains an open empirical and philosophical question whether the same can be said of peer review, namely, that the essential aspects of the process known as expert peer review should be understood by recourse to the features to which reviewers may differ while still remaining experts in their field. Further inquiry into this philosophical and empirical question should be undertaken, with a sensitivity to how reasonable and unreasonable disagreement can be accommodated in discipline-wide communication structures.

REFERENCES

Bailar, John C. 1991. "Reliability, Fairness, Objectivity and Other Inappropriate Goals in Peer Review." *Behavioral and Brain Sciences* 14:137–38.

Bakanic, Von, Clark McPhail, and Rita J. Simon. 1987. "The Manuscript Review and Decision-Making Process." *American Sociological Review* 52:631–42.

Beyer, Janice. 1978. "Editorial Policies and Practices among Leading Journals in Four Scientific Fields." *Sociological Quarterly* 19:68–88.

Bornstein, Robert F. 1991. "Manuscript Review in Psychology: Psychometrics, Demand Characteristics, and an Alternative Model." *Journal of Mind and Behavior* 12:429–68.

Campanario, Juan Miguel. 1995. "Commentary: On Influential Books and Journal Articles Initially Rejected Because of Negative Referees' Evaluations." *Science Communication* 16:304–25.

Chubin, Daryl E., and Edward J. Hackett. 1990. *Peerless Science: Peer Review and U.S. Science Policy.* Albany, NY: SUNY Press.

Cicchetti, Domenic V. 1991. "The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation." *Behavioral and Brain Sciences* 14:119–86.

Cole, Stephen. 1992. *Making Science: Between Nature and Society.* Cambridge, MA: Harvard University Press.

Cole, Stephen, Jonathan R. Cole, and Gary Simon. 1981. "Chance and Consensus in Peer Review." *Science* 214:881–86.
Fiske, Donald, and Louis Fogg. 1990. "But the Reviewers Are Making Different Criticisms of My Paper! Diversity and Uniqueness in Reviewer Comments." *American Psychologist* 45: 591–98.
Fletcher, Robert H., and Suzanne W. Fletcher. 1997. "Evidence for the Effectiveness of Peer Review." *Science and Engineering Ethics* 3:35–50.
Frank, Erica. 1996. "Editors' Requests of Reviewers: A Study and a Proposal." *Preventative Medicine* 25:102–4.
Gardner, Martin J., and Jane Bond. 1990. "An Exploratory Study of Statistical Assessment of Papers Published in the British Medical Journal." *Journal of the American Medical Association* 263:1355–57.
Godlee, Fiona, Catharine R. Gale, and Christopher N. Martyn. 1998. "Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports: A Randomized Controlled Trial." *Journal of the American Medical Association* 280:237–40.
Gore, Sheila M., and Gerald Jones. 1992. "The *Lancet*'s Statistical Review Process: Areas for Improvement by Authors." *Lancet* 340:100–102.
Gottfredson, Stephen D. 1978. "Evaluating Psychological Research Reports: Dimensions, Reliability, and Correlates of Quality Judgments." *American Psychologist* 33:920–34.
Grice, Paul. 1989. "Logic and Conversation." In *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Hargens, Lowell L. 1988. "Scholarly Consensus and Journal Rejection Rates." *American Sociological Review* 53:139–51.
———. 1990. "Variation in Journal Peer Review Systems." *Journal of the American Medical Association* 263:1348–52.
Hargens, Lowell L., and Jerald R. Herting. 1990a. "Neglected Considerations in the Analysis of Agreement among Journal Referees." *Scientometrics* 19:91–106.
———. 1990b. "A New Approach to Referees' Assessments of Manuscripts." *Social Science Research* 19:1–16.
———. 2006. "Analyzing the Association between Referees' Recommendations and Editors' Decisions." *Scientometrics* 67:15–26.
Hopewell, Sally, Kirsty Loudon, Mike J. Clarke, Andrew D. Oxman, and Kay Dickersin. 2009. "Publication Bias in Clinical Trials due to Statistical Significance or Direction of Trial Results." *Cochrane Database of Systematic Reviews*, no. 1.
Hull, David L. 1988. *Science as a Process*. Chicago: University of Chicago Press.
Jayasinghe, Upali W. 2003. "Peer Review in the Assessment and Funding of Research by the Australian Research Council." PhD diss., University of Western Sydney.
Jayasinghe, Upali W., Herbert W. Marsh, and Nigel Bond. 2003. "A Multilevel Cross-Classified Modelling Approach to Peer Review of Grant Proposals: The Effects of Assessor and Researcher Attributes on Assessor Ratings." *Journal of the Royal Statistical Society* A 166:279–300.
Justice, Amy C., Mildred K. Cho, Margaret A. Winker, Jesse A. Berlin, Drummond Rennie, and the PEER Investigators. 1998. "Does Masking Author Identity Improve Peer Review Quality? A Randomized Controlled Trial." *Journal of the American Medical Association* 280:240–42.
Kuhn, Thomas S. 1977. "Objectivity, Value Judgment, and Theory Choice." In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, ed. Thomas S. Kuhn, 320–39. Chicago: University of Chicago Press.
Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.
Lee, Carole J. 2008. "Applied Cognitive Psychology and the 'Strong Replacement' of Epistemology by Normative Psychology." *Philosophy of the Social Sciences* 38:55–75.
———. 2012. "Incentivizing Procedural Objectivity: Community Response to Publication Bias." Unpublished manuscript, University of Washington, Seattle.
Lee, Carole J., and Christian D. Schunn. 2011. "Social Biases and Solutions for Procedural Objectivity." *Hypatia* 26:352–73.
Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Forthcoming. "Bias in Peer Review." *Journal of the American Society for Information Science and Technology.*

Lindsey, Duncan. 1978. *The Scientific Publication System in Social Science*. San Francisco: Jossey-Bass.

Marsh, Herbert W., and Samuel Ball. 1989. "The Peer Review Process Used to Evaluate Manuscripts Submitted to Academic Journals: Interjudgmental Reliability." *Journal of Experimental Education* 57:151–69.

Marsh, Herbert W., Upali W. Jayasinghe, and Nigel W. Bond. 2008. "Improving the Peer-Review Process for Grant Applications: Reliability, Validity, Bias, and Generalizability." *American Psychologist* 63:160–68.

Merton, Robert K., and Harriet Zuckerman. 1971. "Institutional Patterns of Evaluation in Science, 1971." In *The Sociology of Science: Theoretical and Empirical Investigations*, ed. Norman W. Storer, 460–96. Chicago: University of Chicago Press.

Neuliep, James W., and Rick Crandall. 1990. "Editorial Bias against Replication Research." *Journal of Social Behavior and Personality* 5:85–90.

Rust, John, and Susan Golombok. 2009. *Modern Psychometrics: The Science of Psychological Assessment*. 3rd ed. New York: Routledge.

Scott, William A. 1974. "Interreferee Agreement on Some Characteristics Submitted to the *Journal of Personality and Social Psychology*." *American Psychologist* 29:698–702.

Suls, Jerry, and Renee Martin. 2009. "The Air We Breathe: A Critical Look at Practices and Alternatives in the Peer-Review Process." *Perspectives on Psychological Science* 4:40–50.

van Rooyen, Susan, Fiona Godlee, Stephen Evans, Richard Smith, and Nick Black. 2010. "Effect of Blinding and Unmasking on the Quality of Peer Review." *Journal of the American Medical Association* 280:234–37.

Weller, Ann C. 2001. *Editorial Peer Review: Its Strengths and Weaknesses*. Medford, NJ: American Society for Information Science and Technology.

Ziman, J. 1969. "Information, Communication, Knowledge." *Nature* 224:318–24.

Zuckerman, Harriet, and Robert K. Merton. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System." *Minerva* 9:66–100.