American Sociological Association

American Sociological Review
2025, Vol. 90(5) 879–915

DOI:10.1177/00031224251362254
journals.sagepub.com/home/asr
Sage

How Values and Uncertainty Shape Scientific Advanfce in Peer Review

Daniel Scott Smith,[a] Neha Nayak Kennard,[b] Tianyu Du,[c] and Daniel A. McFarland[c]

Abstract

Tens of thousands of scientists contribute to peer review as journal editors and reviewers of the millions of manuscripts submitted every year. How do they decide what is quality work? What values do they apply in evaluating which science merits publication and which does not? How do they respond to dissensus and uncertainty? Who has the greatest influence over the final outcome? This study combines close reading with large language models to analyze 80,000 reviews of 28,000 accepted and rejected manuscripts in engineering and the life sciences. By following reviewers' value judgments and editorial decisions, we come to a different view of how epistemic cultures are practiced in journal science. Instead of a consensual dialogue revealing salient norms, we find reviewers differently weigh ("commensurate") their judgments to attribute value to works. Their pluralistic viewpoints elevate uncertainty about the work, and editors respond by aligning with the most negative of reviewers. Surprisingly, we observe engineers and life scientists find the same epistemic criteria are salient, valued, and influential, with novelty and accuracy being primary. These results underscore how contingency and uncertainty are structural features of STEM peer review and essential to its effectiveness and legitimacy.

Keywords

peer review, epistemic culture, scientific valuation, sociology of science, computational sociology

Nearly every consequential aspect of contemporary science depends on peer review. Governmental and nongovernmental agencies rely on peer review to strategically invest billions of dollars every year. Professional societies, associations, universities, and organizations rely on peer review to determine which works and authors merit prestigious prizes, awards, and life-long faculty appointments. Not to mention the largest and most ubiquitous arena reliant on peer review: journal science. According to the Scopus data published by Elsevier, there were nearly 30,000 peer-reviewed journals in 2024 alone.[1] Peer-reviewed journals are the selection mechanism of modern science: they identify and diffuse promising new ideas (Cheng et al.

[a]Duke University
[b]University of Massachusetts Amherst
[c]Stanford University

Corresponding Authors:
Daniel Scott Smith (daniel.scott.smith@duke.edu) and Daniel A. McFarland (mcfarland@stanford.edu)

2023) and so sustain our collective undertaking of knowledge creation. Impartial and rigorous scholarly evaluation underwrites our trust in science as the basis of the inestimable number of decisions that define modern social life, from how we maintain our personal health to how governments address global challenges like war and climate change.

Hundreds of thousands of scientists voluntarily edit and review millions of manuscripts every year. The diversity of these editors and reviewers abounds. They come from different countries with different personal and professional identities (Smith et al. 2023; Zuckerman and Merton 1971), different disciplinary training and epistemological styles (Mallard, Lamont, and Guetzkow 2009), and different experiences and statuses. Given this plurality, how do they agree on what is quality science? How do they justify their decisions? Whose opinion matters most? How, in short, do scientists' ordinary value judgments shape the advance of new works through peer review? In light of chronic concerns about reviewer reliability and bias (Bornmann, Mutz, and Daniel 2010), proposals to automate peer review (Lin et al. 2023; Zhuang et al. 2025), and rising distrust in science (Druckman 2022), such questions concerning the social character of human judgment and uncertainty in science are timely and essential for sociologists to ask. These questions are all the more important in science, technology, engineering, and mathematics (STEM), where new research is produced and evaluated at far higher rates than most fields, with disproportionate influence on industry, health, and social policy.

To engage these questions, we build on research on the social dimensions shaping scientific evaluation (Kuhn 1977; Zuckerman and Merton 1971), particularly in peer review settings (Lamont 2012; Lamont and Mallard 2005). Central to this work is the importance of consensus (Laudan 1984; Oreskes 2019), where social norms such as impartiality help scientists "self-police," remain objective, and agree on the merit of new works (Merton 1973:276). Yet, a surprising finding from research on grant panels in the natural sciences and medicine is that reviewers are "unreliable" and regularly disagree. Much work has since focused on how biases and inefficiencies distract reviewers from focusing on the quality of scientific arguments, creating unreliable reviews (Herrera 1999; Iezzoni 2018; Langfeldt 2004; Teplitskiy et al. 2018; Wennerås and Wold 2010).

Other work in the sociology of knowledge concerning interdisciplinary grant panels from the humanities and social sciences also focuses on consensus, but from a different angle. Given the remarkable plurality of reviewers, how can they ever agree on what is excellent scholarship in the first place? An economist may find rigorous causal identification to be excellent, whereas a historian may focus on a narrative's style and presentation. A core insight from this literature is that different scholarly fields and individuals emphasize and define merit in distinct ways (Guetzkow, Lamont, and Mallard 2004; Lamont and Mallard 2005; Mallard et al. 2009). Reviewers are "unreliable" not because they are biased, but because they rely on their own training and experiences in their disciplinary cultures to interpret the epistemic merit of new work (Lee et al. 2013). Reviewers overcome their dissensus, despite this plurality, by situating work in the context of their respective field norms and styles or by deferring to others with more expertise (Lamont and Huutoniemi 2011).

We take this prior work as our point of departure. We shift focus away from notions of broad consensus, unreliability, and bias in peer review toward the dissensus, plurality, and uncertainty scientists ordinarily face when determining the quality of scholarship.

In peer review, scholars apply *values* or *criteria* of merit (e.g., "novelty" and "clarity") to determine the worth of knowledge claims.[2] When reviewers take a stand on these claims based on epistemic criteria, we say they make *epistemic value judgments*. These epistemic value judgments are one facet of epistemic culture (Longino 1983), which can include other cognitive content like thought styles

and programs (Fleck 1979), social norms like universalism (Merton 1973), and materials like lab equipment and reagents (Cetina 1999). Like others, we argue that when scholars perform peer review, they draw on the epistemic values most *salient* to their fields' epistemic cultures, and this salience varies meaningfully between fields. But peer review in journals involves more than the commonly held values of one's epistemic culture. Peer review tasks scholars with holistic assessment of a manuscript's worthiness for publication by carefully considering an array of epistemic values. Moreover, it does this not through an egalitarian, consensus-finding process but a hierarchical one. Reviewers are *critics* who evaluate a wide range of qualities to make recommendations to editors, and editors act as *gatekeepers* who face the plurality of often divergent recommendations and must rely on certain reviewers more than others when deciding to publish or reject papers, thus creating potential for influence and bias.

To make their case to an editor, reviewers draw out, differently emphasize, and take stances on qualitatively distinct criteria. In doing so, they construe their various value judgments into a holistic if not unequivocal statement of the work's worth, such as a rating or recommendation. This *commensurated valuation* (Espeland and Stevens 2008; Lee 2015) is different from shared standards (salience). As a persuasive account to the editor, it is also more than a passive description. Instead, commensurated valuation is how reviewers, as critics, establish the epistemic priority of some criteria over others to assign real value to a new work and thereby assert influence over its trajectory in the peer-review process.

Editors, in turn, base their decisions on reviewer recommendations and accompanying evaluations. When recommendations are highly consensual, the editor has a clear signal of the manuscript's merit, and reviewers have comparable discretion or *influence* over the outcome. But reviewers' commensurated valuation of the same manuscript may lead to divergent views on its quality: some reviewers may apply many and varied criteria, or the same criteria with opposite judgments. Prior work finds this is common in peer review (e.g., Miller 2006:427) and reasonable (Lee 2012:863–67). Faced with divergent recommendations, the editor's decision to reject or accept thus may appear "on a par" (Chang 2002): given the reviews, to reject presents as no better or worse than to accept. We argue that this creates epistemic *uncertainty*, and the potential for any one reviewer to influence the editor correspondingly grows. As gatekeepers facing uncertainty, editors pragmatically focus on reducing the risk of publishing dubious work (Latour 1987), even if that increases the risk of rejecting quality science (Zuckerman and Merton 1971). Thus, we argue that a reviewer only ever pulls the editor toward their own valuation and away from that of others by making *negative* value judgments, because these judgements elevate the kind of risk editors seek to minimize.

Our findings support this argument and help shift our understanding of scientific peer review. Among the life scientists and machine learning engineers we study, we find that dissensus and uncertainty are common and asymmetrically influence whether knowledge is valued. Above all others, judgments of accuracy and novelty are decisive, and in uncertain cases where reviewers present many, varied, and mixed value judgments, the *negative* judgments pertaining to accuracy and novelty are the most influential. Indeed, when views of a manuscript are most wide-ranging and divergent, and when we would expect to find the potential for bias to be highest, we find little evidence of it. Our account helps us understand this surprise. Epistemic plurality results from a social process of evaluation that solicits diverse and independent viewpoints from different critics of scientific works. This plurality elevates the epistemic uncertainty about the quality of work for the journal's gatekeeping editors. Instead of equally weighing and consensually reconciling these judgments, editors respond by doubling down on the core values of journal science: is the work right, and is it new?

Although not sufficient in and of themselves, these qualities are necessary if scientific works are to advance in peer review. In the aggregate, we see that peer review depends on precisely the stuff traditional views find epistemically threatening: reviewers' idiosyncratic commensuration, social hierarchy, and epistemic plurality and uncertainty.

In what follows, we relate this account by systematically analyzing the content of 43,000 reviews of 12,000 manuscripts in engineering and 36,000 reviews of 16,000 manuscripts in the life sciences. We combine close reading, qualitative coding, and machine learning to operationalize reviewers' distinctive epistemic value judgments. Using panel regression, we trace how the epistemic value judgments of reviewers are commensurated in their valuations, and how they lend reviewers a commanding influence over editors, particularly in contexts of uncertainty. We do all this by way of accounting for traditional sources of reviewer bias against authors and, as a novel design feature, editorial bias against reviewers.

VALUES, UNCERTAINTY, AND SCIENTIFIC ADVANCE

Most journals have reviewer guidelines, or conventions, that guide peer review (Becker 2008:56; Merton 1973). These conventions are interpreted and negotiated by scientists who apply them (Lamont 2015), and they are heavily theorized by epistemologists, philosophers, and historians (Longino 1983; Machamer and Osbeck 2004; Shapin 1995; Steel 2010) who try to make sense of them. These conventions do three interesting things.

First, conventions establish norms and rules (Horne and Mollborn 2020) intended to constrain scientists in their evaluation of new works by defining appropriate behaviors. Classically, one such social norm in science is universalism: scientists should evaluate works objectively, instead of factoring in personal taste and preference (Merton 1973).[3] More contemporary norms, such as ethical guidelines, regulate how scientists "police" each other when evaluating new works.

Second, these conventions define specific criteria against which reviewers should judge new scientific works. Classically (Kuhn 1977), such values in science include the following: *Accuracy:* Does the theory agree with the results of experiments and observations? *Consistency:* Is it consistent with itself but also with other currently accepted theories applicable to related aspects of nature? *Fruitfulness:* Does it reveal new phenomena or previously unnoted relationships among those already known? *Scope:* Does it extend far beyond the particular observations, laws, and subtheories it was initially designed to explain? *Simplicity:* Does it bring order to phenomena that would otherwise be isolated, confused, or incoherent?

These values are deduced from the epistemological theories and rationales that distinguish modern science (Becker 2008; Laudan 1984; Longino 1983; Steel 2010) from other fields in its aim to validate new knowledge claims (Cetina 1999). As such, we call these criteria *epistemic values* and emphasize their constitutive role in defining scientific culture and legitimating the peer review process (Lee et al. 2013; Longino 1983; Searle 2018).

Third, although abstract and ubiquitous—and despite being "*the* shared basis for theory choice" (Kuhn 1977:322, emphasis in original)—these conventionalized values are neither monolithic in meaning nor uniformly practiced among individuals within and between scientific cultures (Lamont 2015). Instead, they vary in their *salience* among scientists, in the epistemic priority scientists give them when attributing *value* to new works, and in the *power* they have to sway the judgments of others and ultimately to determine the outcome of peer review.

Salience

A core insight from the literature on theory choice is that epistemic values are "imprecise," "vague" (Kuhn 1977:322, 335), and variably salient (Guetzkow et al. 2004; Lee

2012). For example, in grant panels, "excellence" or "beauty" might mean rigorous causal inference to an economist, but incisive and elegant prose to a literary theorist (Kagan 2009:5; Lamont 2015). Empirical work shows that scholars cite some values more or less frequently, and their meaning depends not only on fields but also on individuals and the social context of evaluation (Lamont 2015). Salience, in other words, is the relative expression of scholarly qualities that differentiate how epistemic cultures are practiced in peer review.

There is less conclusive research on the salience of particular values in STEM fields and none on the peer review of engineering and life sciences (see, e.g., Marsh and Ball 1989; for psychology, see Scott 1974). From the empirical work that does exist and from more conceptual accounts, we have reason to expect there are distinctive STEM subcultures. For example, Snow ([1959] 2012:31–32) argues that "pure scientists and engineers often totally misunderstand each other," and whereas basic scientists focus on abstract and causal explanations, engineers are more "absorbed in making things" and solving problems. Kuhn (1977:335) states that "such values as accuracy, scope, and fruitfulness are permanent attributes of science," but they nonetheless vary by time and field. Cetina (1999) finds that molecular biologists focus on direct observation and exogenous manipulation, whereas high-energy physicists rely on indirect observation. Empirically, there is suggestive evidence of variability among STEM fields along the applied–basic science continuum (Barlösius 2019; Terama et al. 2017).

We reason that criteria can vary meaningfully between subfields within STEM peer review because they express differences in subfields' respective theoretical and research problems, sources of evidence, and the degrees to which they intervene in society (Kagan 2009:2–3). We expect "applied" engineering fields, such as deep learning and data science, to show an epistemological style more oriented toward new tools and methods that reliably solve problems in ways that improve on agreed-upon benchmarks. Life science subfields, such as neuroscience and genetics, tend to have more basic research. So, we expect causal accounts that provide new mechanistic explanations and thorough falsifications are more salient. Formally, we expect these dispositions to lead to differences in the relative salience of epistemic values:

> *Hypothesis 1:* The relative *salience* (presence) of epistemic values varies between engineering and life sciences.

Valuation, Commensuration, and Partiality

In reviewing submitted works, reviewers are critics tasked with writing evaluations that consider a full range of criteria, draw out positive and negative aspects, and present balanced accounts that ultimately justify to the editors the value they see or do not see (Glonti et al. 2019; Sizo et al. 2019). If salience describes the epistemic criteria commonly held among reviewers of different fields, then *valuation* is the social act of ascribing discrete value to works based on those criteria.

To justify their recommendations, reviewers must apply and then construe such categorically distinct values as "fruitfulness," "consistency," and "scope" on a single latent scale that summarizes the work's overall quality. At stake here is the relative weight or priority reviewers apply to *commensurate* their various value judgments to come to a reasoned and justifiable recommendation to the editor (Espeland and Stevens 2008; Hsieh and Andersson 2021; Lee 2015). From both classic and contemporary philosophical accounts of epistemic values in science, we argue there is important variation between and within epistemic cultures in these acts of commensurated valuation. For example, Kuhn (1977:322–23) relates that some values are more or less "decisive," emphasizing that accuracy may be most indicative of theory

choice. Yet he emphasizes that other values, such as consistency and simplicity, also play prominent roles, depending on the field and its state of the art.

As a result of commensurated valuation, all criteria are not equally decisive. We reason that such *partiality* in applying these criteria and attributing value to works is "normatively appropriate" and not necessarily an indication of unreliability or bias (Lee 2012). Different reviewers have different expertise that may lead them to different interpretations of the work and to differential weightings of criteria. Tangibly for sociologists, editors may assemble a tableau of reviewers with complementing expertise in theory, methods, and knowledge of the subfield to piece together a composite view of the work under consideration. Such eclectic constellations of expertise are a hallmark feature of interdisciplinary panels (Lamont 2015). In settings where no one reviewer has complete authoritative knowledge over all claims made in a work, it is appropriate that reviewers draw on their own relevant expertise to present partial and commensurate recommendations. Moreover, if criteria were overdetermined rules (instead of vague values), then any and every reviewer could arrive at the same decision; there would be perfect consensus and, ultimately, no scientific debate or controversy or need for scientists. Such "algorithmic" assumptions about choice and values in scientific induction are widely understood as untenable (McMullin 1982:17). Not only is this partiality appropriate, but it is necessary to science.

Much work investigates either reviewers' ratings or their evaluative text, but less sociological analysis has been devoted to studying both jointly. Some work looks at semantic variation within criteria, such as "impact," and traces whether different definitions lead to higher reviewer scores (Terama et al. 2017). Other work looks at the sentiment of reviewer judgments on various aspects of research (e.g., "results," "theory," "style") but is inconclusive about which are most indicative of reviewers' overall ratings (Bakanic, McPhail, and Simon 1989). Other work explores correlations between criteria and outcomes in psychology, showing "clarity" is the weakest correlate, and "significance of conclusions" the strongest (Marsh and Ball 1989:159). To our knowledge, no study analyzes and compares between different fields the value reviewers attribute to works based on their commensurations. We argue that not all criteria will be equally relevant to reviewers as they attribute value to new works, and the priority reviewers place on different criteria will reflect their own interpretations of the work.

Hypothesis 2: Reviewers differently *commensurate* epistemic criteria in their ratings.

Power, Plurality, and Uncertainty

Editors play a distinctive yet interdependent role as gatekeepers (Crane 1967; Siler, Lee, and Bero 2015). Their decisions depend on reviewers' independent and private critiques and recommendations. But they face uncertainty in these decisions when reviewers make divergent recommendations based on differently commensurated value judgments.

Ordinarily, the most expeditious strategy for editors is to align with reviewer consensus on a manuscript. In these cases, a unanimous recommendation from a panel of independent expert reviewers is a strong signal to the editor about the manuscript's quality. But reviewers are famously "unreliable," with many studies finding they rarely agree when following the same directions to evaluate the same work (Bornmann et al. 2010; Jayasinghe, Marsh, and Bond 2003; Miller 2006; National Research Council 1978).

In cases where reviewers' evaluations and recommendations diverge, who has the power to persuade the editor to their side? If the editor goes against the majority opinion of acceptance to side with the reviewer who recommends rejection, they can more easily maintain the fairness and validity of the evaluative process (impartiality, Lee et al. 2013). Skepticism (Merton 1973) and discretional

modesty are a ready-made rationale editors can effortlessly draw on in most cases: "while there is optimism about the manuscript, there is also cause for serious concern which makes this manuscript unacceptable in its current form." Conversely, if the editor goes against the majority opinion of rejection to side with the odd reviewer who recommends acceptance, they face difficulty maintaining the fairness and validity of the evaluative process, even though the reviewers' positions are relationally identical. It is more difficult because the editor needs to actively muster a convincing, case-specific justification to go against the majority opinion so as not to appear arbitrary or partial: "even though there is a preponderance of doubt and concern, *this* particular work shows exceedingly rare promise because of X, Y, and Z, and therefore merits further consideration." As a result of this conventionalized imbalance between equally weighing all views and a skepticism that errs on the side of caution (and rejection), the classical adage applies—"When in doubt, reject" (Zuckerman and Merton 1971:78)—and the editor sides with the negative reviewer against the positive majority, but not with the positive reviewer against the negative majority.

Such asymmetry in the relative *power* reviewers have over editors' decisions has empirical support and entails mechanisms beyond the norms and customary rules that maintain the fairness, validity, and therefore legitimacy of peer review. One such mechanism is the mundane exigencies of publication in selective STEM journals where space is scarce: a doubt raised is sufficient to justify rejection, given there are more and higher-quality works than space permits. One empirical study of peer review in cardiology, for example, shows that editors aggressively exercise this prerogative, rejecting 99.6 percent of manuscripts when two of three reviewers recommended doing so; yet they still *rejected* 88.6 percent when two of three reviewers recommended acceptance (see Bornmann and Daniel 2010; Coronel and Opthof 1999). Another mechanism is conservatism (Luukkonen 2012), which entails a proclivity to disfavor potentially novel or groundbreaking works that do not conform to editorial expectations (Brezis and Birukou 2020), instead focusing on potential flaws (Hug and Ochsner 2022; Lane et al. 2022). Still another mechanism is generalized risk-avoidance (Carson et al. 2022), which protects the reputation of both editors and journals.

We argue that all such mechanisms indicate a preponderance of editorial concern with maintaining the legitimacy (Johnson, Dowd, and Ridgeway 2006:58; Lee et al. 2013:3) of the peer review process in the face of divergent recommendations. For most manuscripts, no individual reviewer has a commanding influence over the outcome because the editor either aligns with the full consensus, when present, or with the partial consensus to reject the work. Yet, if a reviewer is ever to pull an editor away from full or partial consensus toward their own contrarian position, we argue it is only by making negative value judgments, because editors, as gatekeepers, treat negatively evaluated criteria ("negative modalities") as cause for doubt and therefore compelling evidence to summarily reject (Latour 1987).

Hypothesis 3: Negative reviewer value judgments *most* influence editors' decisions.

Each opposing reviewer recommendation is systematically adduced by a full range of critically evaluated criteria. When these judgments amount to markedly *pluralistic* ("unreliable" or "inconsistent") positions among reviewers of a work (Lamont and Mallard 2005), editors must make their decisions in a context of *epistemic uncertainty*. This uncertainty arises because the opposing reviewer recommendations (reject versus accept) are warranted with an array of differently interpreted, weighed, and judged standards and concerns. For example, a reviewer may recommend the manuscript proceed to revision because the empirical design is remarkably novel, even if some additional tests and

controls are wanting; yet, a different reviewer may recommend just the opposite for the same reasons, or they may bring in additional criticisms, perhaps its failure to engage with the appropriate literatures. The opposing recommendations, revise versus reject, are thus underdetermined by the evaluations and may be considered "on a par" for the editor (Chang 2002; Hsieh and Andersson 2021). In other words, the manuscript's quality and therefore the editor's decision are uncertain.

We argue that epistemic uncertainty is a structural feature of peer review that intensifies skepticism and correspondingly increases editors' proclivity to reject. We argue that editors do not fully consider and equally weigh the expanded and pluralistic range of value judgments that reviewers present to them in these contexts. Instead, editors double down on negative evaluations of the work. After all, manuscripts of truly uncertain quality likely have a sundry mix of both positive and negative value judgments. This heightened uncertainty elevates the risk of making a Type I error (publishing problematic science), threatens the aims of peer review, and therefore opens the opportunity for a critical reviewer to have amplified power over an editor's decision.

Hypothesis 4: In contexts of epistemic *uncertainty*, negative judgments are *more* influential on editors' decisions.

Bias

To the extent that factors outside the epistemic content of scientific work drive reviewers' and editors' decisions, peer review will be biased (Bornmann et al. 2010; Lee et al. 2013). Classic sources of bias include status, country of origin, gender, nepotism, and homophily (Demarest, Freeman, and Sugimoto 2014; Krieger, Myers, and Stern 2023; Merton 1968; Nielsen et al. 2021; Smith et al. 2023; Squazzoni et al. 2021; Teplitskiy et al. 2018; Travis and Collins 1991; Wennerås and Wold 2010). Findings are notably mixed but suggest bias plays some role in the peer review process. Additionally, empirical research overwhelmingly explores the bias evaluators (reviewers and editors) have toward the evaluated (authors). Early work suggests editors may be biased in their selection of reviewers (Bakanic et al. 1987) and against certain reviewers (Chubin and Hackett 1990). But based on the literature concerning bias in peer review and other decision-making domains, we expect that bias may play a mixed or marginal role among reviewers and editors in evaluating new works (Eagly and Karau 2002; Eagly, Makhijani, and Klonsky 1992).

Hypothesis 5: Social biases are related to reviewers' ratings and editors' decisions.

Past work also suggests that uncertainty in science evaluation and theory selection perturbs or even disrupts normal, rules-based evaluation in science (Longino 1983; Machamer and Osbeck 2004; McMullin 1982). For example, scientists may draw more information from the broader social context of scientific practice to aid in decisions and selection. In the case of peer review, when faced with a plurality of epistemological positions and epistemic concerns raised in reviews for and against acceptance, editors may respond to cues about who the reviewers are, their own past experiences collaborating with them, and reviewers' expertise and status in the field in deciding whose recommendation to give more credence to. Epistemic uncertainty and ambiguity would not only drive editors to angularly focus on negative evaluations all the more (Hypothesis 4), but social biases would have more of an impact on editors' final decisions.

Hypothesis 6: In contexts of epistemic *uncertainty*, social biases are *more* related to editors' decisions.

RESEARCH DESIGN

We conceptualize "scientific advance" in our study of peer review as whether scientists

decide manuscripts should be accepted, revised, or rejected. These value-laden judgments have further downstream effects on future work and whether science, as a whole, advances.

Peer Review Data

Our peer review data come from the proceedings of two prestigious venues and represent one of the most comprehensive datasets of STEM peer review. These datasets are ideal because they allow us to look across multiple disciplines within science and engineering across time. Such variation enables us to systematically explore important differences in epistemic cultures, which likely condition whether and how value judgments play a role in the advancement of new scientific work at various decision junctures. The life science journal data are particularly ideal because they enable us to study whether and to what extent social biases, such as nepotism and homophily, may shape the peer review outcome.

Our first source of peer review data comes from a single-masked, interdisciplinary journal in the life sciences, which we name "life sciences journal" (LSJ). Fields within the scope of LSJ include ecology, biology, global health, and genetics, among others. These data contain all identifying information of all authors, reviewers, and editors, as well as the text of every review (n = 36,868) of every manuscript both rejected and accepted (n = 16,096).

LSJ is considered a high-status and selective journal and has a desk-reject rate upward of 60 percent; 42 percent of the remaining submissions are rejected after a first round of review. Peer review begins after a senior editor decides to assign a manuscript to a reviewing editor, who along with one or more other reviewers then independently write evaluations of the work distributed across "major" and "minor" comments. When there is consensus among reviewing scientists that the work should advance, they consult to define a consistent set of required revisions, which is then shared with authors as an editorial letter accompanying the individual reviews. During the period we investigate (2012 to 2022), the peer review proceedings were closed, although reviewers were asked to provide public summaries of the works in addition to their private evaluations shared with the authors and editors. We exclusively analyze the private reviews (both major and minor comments) *before* consultation for manuscripts that have three or more reviewers, including the reviewing editor. All editors and reviewers are professional scientists and well-socialized in their respective fields as experts. Because of the sensitive nature of these data, analyses were conducted only after IRB review and are subject to strict anonymity requirements in reporting through our data-use agreement with the journal.

The second corpus comes from the International Conference on Learning Representations (ICLR), a double-masked peer-reviewed computer science conference focused on the field of representation learning, which is a branch of artificial intelligence.[4] The conference's scope includes fields from data science and statistics to speech recognition, robotics, and machine vision. These data contain the texts of every review (n = 43,785) of every manuscript (n = 12,626) rejected or accepted. Area chairs give about 63 percent of submissions the equivalent of rejections; the remaining submissions get revised during a discussion period and allocated to different parts of the proceedings, with talks representing the highest "accept" decision (see Part A of the online supplement).

ICLR is considered a prestigious and selective conference for AI engineering. Peer review of submissions is conducted by conference participants, with area chairs (reviewing editors) responsible for the final decisions. Peer review unfolds over distinct phases. The first is private: reviewers independently assess the work generally and along dimensions that are varyingly requested from conference to conference. The second phase is the rebuttal period: authors are offered the opportunity to respond to reviewer comments

and requests with a revised submission. After the rebuttal ends, reviewers and area chairs consult to make a final decision, with reviewers sometimes changing their scores in response to the consultation and revisions.[5] Reviewers tend to be much more diverse in their training and experience compared to area chairs because they are recruited from the ranks of submitting participants. ICLR's peer review proceedings are fully open and hosted on OpenReview.net.

We attained permission from the program chairs and ICLR board members of each conference year to access and analyze the pre-discussion data, which include unchanged reviewer ratings and written evaluations. These data had been open and publicly available through the API, but they were subsequently deleted and replaced with the post-discussion versions. We collaborated with the OpenReview system administrators to systematically identify and collect the last logged version of reviews and ratings before the opening of the discussion period, which were remotely stored on an OpenReview server. Thus, we analyze the pre-discussion reviews and ratings that reflect reviewers' independent evaluations.[6]

Outcomes of Interest: Reviewer Ratings and Influence

Reviewer ratings. For each corpus, our first outcome is a reviewer's quantitative judgment about whether the manuscript should advance, falling on the ordinal scale of 1 = most probably reject, 2 = revise and resubmit (probably reject), 3 = revise and resubmit (probably accept), and 4 = most probably accept (Bakanic, McPhail, and Simon 1987:633).

Ratings are explicitly requested from ICLR reviewers (see Part A of the online supplement). In contrast, LSJ makes no explicit request for ratings. Instead, LSJ editors rely on the texts of reviewers' evaluations, which often include explicit recommendations whether to advance the manuscript. To proxy an ICLR-commensurate score for each review in the life science journal, we took a machine learning approach. First, we elected to use a RoBERTa-based sequence classification model (Liu et al. 2019).[7] RoBERTa is an open-sourced improvement on the influential BERT language model (Devlin et al. 2019). RoBERTa and other transformer models have been pretrained on millions of publicly available texts, including the BookCorpus, Wikipedia, and CC-News datasets, to learn rich language representation. Because our use case is predicting an ordinal rating on a scale of 1 to 4 based on evaluative texts, we chose RoBERTa for Sequence Classification with a regression head on top. By adding a regression head, we adapt RoBERTa so that it learns to map evaluative texts directly to continuous scores between 1 and 4.

Second, we fine-tuned the pretrained RoBERTa model using a dataset of ICLR reviewer evaluations paired with their explicit ratings, allowing the model to learn the mapping from peer reviews to reviewer ratings. Finally, we predicted LSJ reviewer ratings with this model. To evaluate its performance before prediction, we computed several standard benchmark metrics from qualitative coding and machine learning. For example, the high correlation coefficients ($r = 0.87$ for ICLR and $r = 0.84$ for LSJ) indicate a strong linear relationship between the predicted and true ratings, and the low RMSE values (RMSE = 0.55 for ICLR and RMSE = 0.63 for LSJ) demonstrate the model's accuracy in predicting ratings close to the actual values.[8] We provide more context on how we coded the LSJ test set in Part A of the online supplement. We also conduct sensitivity analyses to assess the effects of measurement error on our results, finding that our conclusions remained robust despite potential inaccuracies introduced by the classifier (see Part R1 of the online supplement). We discuss measurement error and noise introduced by classifiers in the limitations section.

Reviewer influence. The second outcome we seek to explain is the relative influence a reviewer has over an editor's decision.

Whereas the first outcome concerns what goes into a reviewer's own valuation, this second outcome concerns the extent that editors are relatively persuaded against the consensus (i.e., the average opinion of the others) by what an individual reviewer voices in their review.[9] With this second outcome, we try to observe how value judgments and epistemic culture can be consequential for shaping real outcomes in peer review. To measure a reviewer's relative influence, we proceed in three steps.

First, we compute the absolute difference between each reviewer's rating and the average reviewer rating for each manuscript. This difference indicates how closely a reviewer aligns with other reviewers. A smaller absolute difference indicates greater similarity among the reviewer and the other reviewers, whether they all agree to reject or accept. Correspondingly, a reviewer's similarity with the other reviewers indicates their lower potential to influence the editor, relative to the other reviewers. In this step, we essentially observe the distinctiveness of the focal reviewer from the others in terms of their valuation of the manuscript.

Second, we compute the absolute difference between the reviewer's rating and the editor's decision. A smaller absolute difference indicates that an editor's decision is closely aligned with a given reviewer's recommendation. Correspondingly, greater alignment indicates greater potential that the editor was influenced by the reviewer; in contrast, editors who made a decision opposite from what the reviewer recommended are less likely to have been influenced by that reviewer. In this step, we essentially observe the alignment of the focal reviewer with the editor.

Finally, we take the difference between these two absolute differences and standardize it to have a mean of zero and a standard deviation of one. This final measure captures how much a reviewer's own recommendation pulled the editor's decision away from the consensus of the other reviewers' recommendations. In this final step, we constrain the measure of influence to be negative or zero when the focal reviewer's rating is the same as the consensus or when it is opposite the editor's final decision. Conversely, we constrain it to be positive when the focal reviewer's rating is different from the consensus but similar to the editor's final decision. In summary, our influence measure expresses an individual reviewer's influence, netting out consensus.

Predictors

Epistemic value judgments. Our main argument concerns the epistemic value judgments reviewers bring to bear in their evaluations of a manuscript, and how these judgments influence editors' final decisions. Historically, observing such judgments has been empirically untenable due to the hidden character of peer review. Empirically, observing these judgments demands a machine learning approach due to the sheer size of our corpora (millions of sentences). Because we have the text data corresponding to all evaluations, we can computationally discover, observe, and measure these judgments and study their relationship with review outcomes. To do so, we first defined an epistemic value judgment as a composite of (1) an evaluative statement (i.e., positive or negative) and (2) the epistemic criterion or value being evaluated (e.g., novelty).

We consulted the philosophical and sociological literature on theory selection and scientific (e)valuation to identify a set of canonical epistemic values (Kuhn 1977; Lamont 2015; McMullin 1982). Then, we drew on previous work in natural language processing (Kennard et al. 2021; Yuan, Liu, and Neubig 2022) to identify training data (i.e., DISAPERE and ReviewAdvisor) with labels of the epistemic criteria and the sentiment/polarity with which they are cited in reviewer evaluations. Table 1 shows the reviewer criteria labeled in the training data as they proxy Kuhn's epistemic values.

Generally, we can loosely proxy each of Kuhn's epistemic values, although with

Table 1. Kuhn's Epistemic Values for Theory Selection as Proxied by the Review Criteria Available in Training Data Sources

Proxy Label[a]	Epistemic Value[b]	Merit Criterion[c]	Example Review Sentence[d]
Accuracy	Accuracy *Does it agree with the results of experiments and observations?*	Soundness/Correctness *Are the claims, conclusions, and interpretations in the paper convincingly supported by the data and (experimental) design? Is the approach soundly executed?*	"However, I also believe that there are some significant flaws in the experimental design and interpretation that weaken the study."
Clarity	Simplicity *Does it bring order to phenomena that would otherwise be isolated, confused, or incoherent?*	Clarity *For a reasonably well-prepared reader, is it clear what was done and why? Is the study explained clearly? Is the paper well-written and structured?*	"The quantification of X in terms of hours post-Y is somewhat confusing."
Consistency	Consistency *Is it consistent with itself but also with other currently accepted theories applicable to related aspects of nature?*	Meaningful Comparison *Are the comparisons to prior empirical and theoretical work sufficient? Are these comparisons carried out well and fair?*	"Please justify why X was given in Y once per day instead of continuously in Z per the standard paradigm in the field."
Novelty	Fruitfulness *Does it reveal new phenomena or previously unnoted relationships among those already known?*	Originality *Are there new topics, technique, methodology, or insights?*	"This is a big topic of interest to many groups studying development and cell biology."
	Scope *Does it extend far beyond the particular observations, laws, subtheories it was initially designed to explain?*	Impact *Does the paper address an important problem? Are other people (practitioners or researchers) likely to use these ideas or build on them? Is the paper a good fit for the venue?*	
Replicability		Replicability *Are there sufficient details to verify the correctness of the procedures? Is the supporting dataset or software provided? Is it easy to reproduce/ replicate the results?*	"In fact, the entire description of the surface area calculation is a single sentence at the end of the methods, providing no details, making it difficult for others to replicate the methods."
Thoroughness		Substance *Does the paper contain substantial experiments to demonstrate the effectiveness of proposed methods? Are there extensive analyses and rigorous controls? Does it contain meaningful ablation or sensitivity studies? Have authors considered alternative explanations?*	"Such data would strengthen the authors' conclusion that 'X specifically controls local Y' (line 144)."

[a]For each criterion, we chose a proxy label that jointly expresses its epistemic and merit definitions, hewing more closely to merit definitions, which guided human annotation of the training data.
[b]We adapted Kuhn's (1977:321–22) definitions following his discussions of epistemic values for theory selection.
[c]Annotation guidelines for the ReviewAdvisor (Yuan et al. 2022) and DISAPERE (Kennard et al. 2021) training data are adapted from the Association for Computational Linguistics (ACL) 2018 reviewer rubrics and further adapted here as we generalized to the life sciences.
[d]Examples are drawn from the human-labeled ICLR and LSJ training datasets to illustrate the focal value judgment.

qualifications. First, his notion of simplicity, which best represents adductive parsimony, is closest to but still conceptually distinct from our measure of clarity. Nonetheless, we treated clarity as a proxy because, when annotating, we often observed that "unclear" or "not clear" jointly indicated reviewers' judgments of the coherence and elegance of the research design and its written presentation. In the actual data and in the criterion's definition, this distinction is underspecified. Second, we combined notions of scope/impact with fruitfulness/originality because "groundbreaking" work was often defined by both its novelty and its ability to define new research avenues and applications; in other words, reviewer judgments of the scope and impact of a work (or lack thereof) were warranted by assessments of its fruitfulness/originality.

The ReviewAdvisor and DISAPERE data together contain over 273,992 peer review sentences from ICLR, each labeled for the presence of these various epistemic criteria. Because these represent only engineering, we augmented these training data with an additional 8,500 sentences from LSJ that we labeled using a standardized coding protocol after iteratively calibrating our judgments. This was essential for domain adaptation from machine learning to the life sciences. Table B1 in the online supplement reports our Cohen's kappa among human annotators for each criterion. We augmented these training data with an additional 5,000 review sentences generated by prompting OpenAI's GPT-4 with domain-specific cues. Table B2 depicts examples from GPT-simulated review sentences, which we closely read in conjunction with authentic human sentences in both of our corpora. We validated these augmented data by having a human annotator choose labels for a set of blinded simulated sentences based on the GPT-4 prompts, and we then compared human labels with GPT labels (f1 = 0.86). These data also contained labels for the sentiment of the evaluative sentence: positive, negative, or neutral (see Parts B and R7 of the online supplement).

Finally, to measure value judgments in our data using these two training corpora, we fine-tuned RoBERTa models on the two tasks (average F1 for values across both corpora = 0.800; for polarity = 0.880; see Table B4 in the online supplement). Then we used each to predict both components of the value judgment in each sentence. Finally, we combined components and aggregated up to the review-level by summing the total sentences of each value judgment per the definition in Table 1.

Editorial Uncertainty

We measure editorial uncertainty using the epistemic content of reviewers' written evaluations, *not* variance in their ratings. This is for technical and substantive reasons. Technically, our measure of reviewers' relative influence on editors already nets out the relative alignment (consensus/dissensus) among their recommendations. If we were to measure editorial uncertainty based on variance in reviewer ratings, we would be selecting manuscripts based on our outcome of influence, thereby upwardly biasing our estimates. Substantively, the sociological and philosophical literature underscores that variance in ratings is principally driven by reviewers' pluralistic epistemic viewpoints in their evaluations (Lee 2012:865). From the editor's perspective, these divergent positions make any recommendation divergent from the consensus underdetermined and thus elevate the uncertainty about quality. Therefore, to measure this uncertainty, we focus on the evaluations, not the recommendations themselves.

To do this, we measure the epistemological positions each reviewer takes on a given manuscript and represent their review as a multidimensional vector comprising the 12 distinct value judgments (6 criteria $\times$ 2 for positive and negative). The value of each dimension is the observed number of sentences in which a reviewer takes the respective position in their evaluation. Then, using cosine similarity, we compare each pair of reviewers' epistemological positions, rendered as review vectors projected into this 12-dimensional epistemic

space, and then average these pairwise cosine similarities to create a summative statement of the degree of alignment among reviewers' epistemological positions on the manuscript. We define consensual manuscripts as those whose reviewers' evaluations are in the top quintile of similarity; these represent relatively more certain, determined cases for the editor. We define dissensual manuscripts as those whose reviewers' evaluations are in the bottom quintile of similarity, and these represent relatively uncertain, underdetermined cases for the editor (see Part C of the online supplement).

Social Biases

Social bias is a persistent concern in peer review. We measure bias in LSJ reviewer-author and editor-reviewer relations. These are not observable in ICLR because it is fully anonymized.

Gender. Authors and reviewers may write differently based on their gendered socialization and professionalization, and this may systematically lead to variation in how work is reviewed (Flaminio et al. 2022). Editors may also place different weight on reviewers' input due to reviewers' gender—or because of their own gender (Eagly and Karau 2002). We use genderize.io to classify the gender conventionally signaled by the pen-names of scientists. Importantly, this classifier was trained on a polyglot corpus of names, so there are no systematic biases introduced when names are not overtly from, for example, Anglo-American contexts. Probabilities that fell within the range of 0.35 to 0.65 we treated as gender neutral. We use dichotomous measures indicating the reviewer's first name is conventionally female (see Parts D and R3 of the online supplement).

Place. Authors and reviewers from places contextually (culturally) more similar to LSJ's (i.e., Anglo-American countries) may, on average, write more conformably and therefore "better" (i.e., in standard English). Editors may also place more weight on these reviewers compared to others because of their conformability and higher cultural alignment (Demarest et al. 2014; Smith et al. 2023). The reverse is likely for the opposite reasons. For the LSJ corpus, we measure place of origin dichotomously, with a 1 indicating the scientist (reviewer, editor, or author) writes from an institution outside the United Kingdom, United States, Canada, New Zealand, Australia, or South Africa (Anglo-U.S. context) (see Part E of the online supplement).

Reviewer status. Different reviewers occupy different positions in the global network of science with respect to their relative scholarly preeminence among other scientists. Field-theoretic accounts underscore how different positions in this social structure correspond to differences in resources and power, which in turn enable them to have a commanding influence over such matters as determining who and which work is recognized or "consecrated" (Lamont 2012). In particular, reviewers with higher status may evaluate work differently (Gallo, Sullivan, and Glisson 2016), and editors may trust and be influenced by such reviewers more than others with comparatively lower status. We use two indicators of reviewer status, measured as their relational positions in the global field of science. The first indicator is the extent to which other scientists cite a reviewer's published works, proxying both their renown and their work's centrality in the field. To compute this, we first construct a network graph with scientists as nodes and draw weighted directed edges between them, indicating if and how often they cite each other (see Part F of the online supplement). We then compute the in-degree centrality of the reviewer. The second indicator is the extent to which the reviewer is collaborative with other scientists. To compute this, we draw weighted undirected edges between scholars, indicating if and how often they co-author papers together. Then, we compute the degree centrality of the reviewer. Because these variables are collinear with variance inflation factors > 10, we create and standardize a latent factor summarily describing the reviewer's status based on these variables (see Part G of the online supplement).

Collaborative proximity. We next double down on the interrelationships of scientists and how that might have a commanding influence on the final editorial outcome (Sandström and Hällsten 2008). We identify four indicators that proxy the collaborative proximity (i.e., nepotism) of reviewers and first authors, as well as between editors and reviewers (see Part F of the online supplement), and we then create a standardized composite measure (see Part G of the online supplement). The first indicator is their total collaborations. The second is the total number of shared collaborators, and the third is the unique number of shared collaborators. Finally, the fourth indicator is constructed using the embeddings from the Node2Vec algorithm on the collaboration graph (Grover and Leskovec 2016). This indicator measures how similar the two scientists' communities of collaborators are. We create and standardize a composite measure describing this collaborative proximity for each reviewer-author and editor-reviewer dyad as a proxy of nepotism.

Citation proximity. Next, we measure the extent to which scientists are intellectually similar (Travis and Collins 1991), as self-similarity is highly related to personal preference and attachment (McPherson, Smith-Lovin, and Cook 2001). In particular, reviewers may be partial to work similar to theirs, and editors may be especially persuaded by reviewers who share similar perspectives or work in adjacent areas because they mirror their own understanding, interpretations, and scientific habitus (Bourdieu 1974, 1988). We use four indicators that proxy the homophily among scientists (see Part F of the online supplement) to create a standardized composite measure (see Part G of the online supplement). The first two indicators are frequencies one scientist cites the other's work and vice versa, indicating the relative importance of one's work for the other (i.e., how many times a reviewer may cite a first author). The third is the number of works the scientists co-cite, indicating shared problem areas. The fourth indicator is the similarity of scientists' cited works, indicating the extent to which they work on similar research problems and topics.

Additional controls. Epistemic values expressed in reviewers' evaluations are wrapped with textual artifacts. These may influence an editor above and beyond social biases. Therefore, we control for the review's politeness and readability (see Part H of the online supplement), although we do not report these in the main results. Review length is collinear with frequency of value judgments, so we controlled for it by dividing each value judgment by the total number of sentences in the review to represent them as relative proportions (e.g., 0.09 means 9 percent of a review's sentences).

Analytic Strategy

We approach these observational data through rigorously controlled analyses and do not present causally identified parameters. Due to this design, our estimates are necessarily correlational, and our controls help us obtain less biased estimates but do not fully rule them out.

In terms of both corpora, our construction of the dependent variable measuring reviewer influence is, by design, a direct violation of the independent and identically distributed assumption of OLS, as values for one reviewer are systematically related to values of the others. Moreover, we do not explicitly measure various dimensions of the submitted manuscripts or ascribed characteristics of their authors (Bakanic et al. 1987), both of which are likely to make reviews of any one manuscript more similar than reviews of different manuscripts. Additionally, these manuscript and author characteristics drive editor decisions. Leaving them out would bias our estimates of reviewers' ratings and influence on editors. Therefore, we account for the violation of OLS assumptions by fitting panel regression models with manuscript fixed effects to our observational data.

Table 2. Reviewer Valuations (Ratings), Editor Decisions, and Reviewer Influence for All Manuscripts

	ICLR				LSJ			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Reviewer rating	2.390	.660	1.000	4.000	2.580	.960	.980	4.150
Editor decision	1.470	.700	1.000	4.000	1.880	.750	1.000	4.000
Ed. decision = Avg(Rev. rating)	.190	.390	.000	1.000	.180	.390	.000	1.000
Reviewer influence	.000	1.000	–3.500	3.810	.000	1.000	–3.540	3.200
N manuscripts	12,626				16,096			
N reviews	43,785				36,868			
N sentences	1,276,216				1,188,409			

Manuscript fixed effects statistically control for observed and unobserved confounders related to the manuscript, including the manuscript's content and characteristics of its authors, as well as reviewers' evaluations of its epistemic merit, their ratings, and their influence on editors. We fit a taxonomy of nested fixed-effects panel regression models to test our research hypotheses, which take the general form

$$y_{rm} = \alpha_m + \beta x_{rm} + \varepsilon_{rm},$$

where y is either reviewer r's rating or their influence on the editor for manuscript m, α is the manuscript fixed effect, x is a vector representing question predictors and controls, and ε represents the idiosyncratic error term.[10] We standardize all reported regression coefficients, which can be interpreted as *on average, a standard-deviation increase in the proportion of sentences a reviewer makes about* <epistemic value judgment> *corresponds to a change in their* <dependent variable> *of* <coefficient> *standard deviations*.

FINDINGS

Reviewer Recommendations and Editorial Decisions

Table 2 reports statistics on the review outcomes for ICLR and LSJ; this includes reviewers' valuations, editorial decisions, and reviewers' influence on those decisions. For both venues, reviewers' valuations (ratings) are higher than editors' decisions by nearly a full point, which ordinally corresponds to a different decision altogether (~2.5 = revise & resubmit, leaning accept versus 1.5 = reject, leaning revise & resubmit). This difference means editors are, on average, more discriminating and less sanguine about manuscripts reflecting their role as gatekeepers.

The statistics in Table 2 do show more variation among reviewer ratings, on average, in LSJ. Life science reviewers in our venue tend to diverge in their recommendations more often than do engineers (0.960 versus 0.660), suggesting more "unreliability."[11] The two venues have comparable ranges of reviewer influence, indicating that editors tend to be similarly responsive to reviewers' recommendations. Editorial alignment with reviewer consensus is relatively rare in each corpus. The editor's decision is identical to the reviewer consensus for only 19 and 18 percent of manuscripts in ICLR and LSJ, respectively. Such lack of consensus means there are far more cases where reviewers have the potential to influence the editor.

The Salience of Epistemic Values

We start descriptively by comparing the relative salience of epistemic values among engineers and life scientists in Table 3 and

Table 3. Describing Salience: Distribution of Epistemic Value Judgments in ICLR (n = 43,785 reviews) and LSJ (n = 36,868 reviews)

Epistemic Value Judgment	ICLR				LSJ				Diff. in Means
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	
Accuracy (–)	.077	.066	0	.750	.130	.089	0	.833	–.053***
Clarity (–)	.083	.091	0	.857	.101	.092	0	.857	–.018***
Consistency (–)	.025	.040	0	.571	.023	.036	0	.500	.002***
Novelty (–)	.070	.068	0	1.000	.025	.045	0	.800	.045***
Replicability (–)	.016	.031	0	.417	.023	.037	0	.474	–.007***
Thoroughness (–)	.120	.082	0	.714	.169	.104	0	.833	–.049***
Accuracy (+)	.021	.035	0	.400	.022	.039	0	.800	–.001***
Clarity (+)	.026	.036	0	.429	.009	.026	0	.545	.017***
Consistency (+)	.004	.014	0	.333	.004	.015	0	.333	.000
Novelty (+)	.072	.069	0	.800	.048	.058	0	.700	.024***
Replicability (+)	.004	.012	0	.250	.001	.007	0	.200	.003***
Thoroughness (+)	.031	.043	0	.500	.017	.035	0	.667	.014***
No judgment	.451	.137	0	1.000	.428	.134	0	1.000	.023***
N sentences	29.147	14.959	4.000	122.000	32.237	17.992	5.000	105.000	–3.090***

Note. *** differences are statistically different from zero at the alpha = 0.001 level based on independent two-sample t-tests of differences in means with unequal variance.

Figure 1. For both groups of reviewers, about half the review is dedicated to epistemic value judgments (the other half accounts for text that structures and summarizes the manuscript without evaluations). These central tendencies and variation in the evaluative text highlight the relative salience of value judgments within and between these two venues. The table and figure illustrate how frequently reviewers independently cite these concerns.

We note three general features of these distributions. First, while all ICLR-LSJ differences are statistically different from zero, none are substantially large in terms of the relative attention given to them. All are less than 0.100, meaning a difference amounting to fewer than 10 percent of an average review's sentences. The most salient and least salient concerns are the same in ICLR and LSJ: thoroughness (–), accuracy (–), and clarity (–) versus replicability (+) and consistency (+). The tight distribution of the average salience of each concern along the $y = x$ line in Figure 1 (r = 0.889) suggests, overall, there is more alignment than divergence between these venues in terms of the salient values they frequently raise. These results are inconsistent with our expectations set out in Hypothesis 1, which advanced the idea that the fields would be more dissimilar than similar. Second, LSJ and ICLR reviewers tend to raise doubts, questions, and concerns more than they tend to praise the works. Although we did not expect this result, it corroborates past work that finds a preponderance of negativity in reviews and suggests peer review in both fields generally skews skeptical and critical (Bordage 2001). Third, there is great variance in the concerns reviewers raise. The standard deviations across the criteria within and between fields are neither the same nor very small. This variation suggests reviewers regularly depart from field norms (the central tendencies) and are very partial in their treatments of the criteria (Lee 2015).

To illustrate, Figure 2 reproduces two reviews that highly load on the commonly salient epistemic values of thoroughness (–), accuracy (–), and clarity (–). Both reviewers take issue with the accuracy of the claims made and the thoroughness of the investigation, while repeatedly asking for more

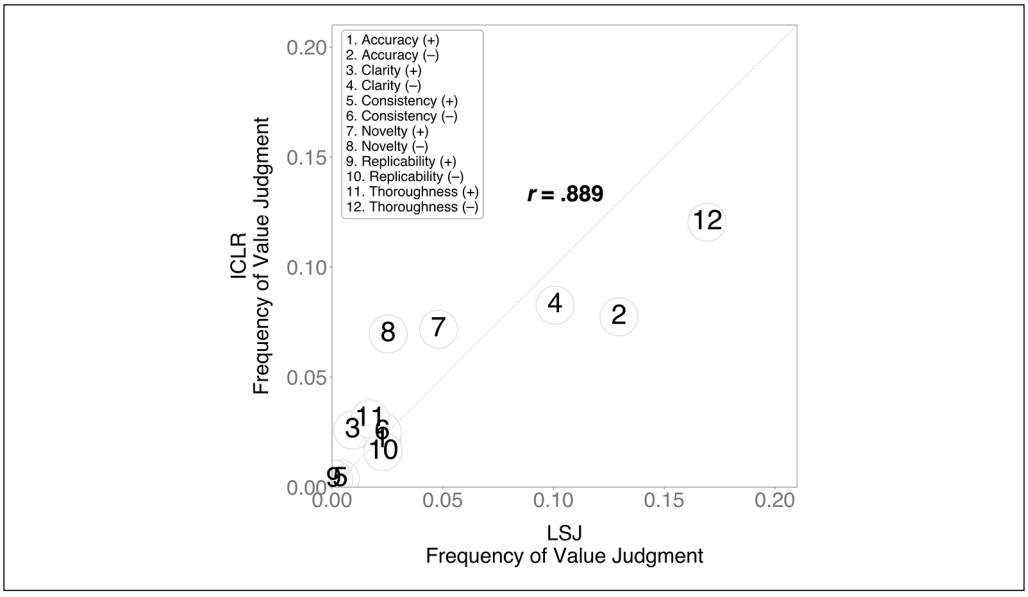

Figure 1. Describing Salience: Proportion of Total Sentences (Salience) in which ICLR versus LSJ Reviewers Cite Epistemic Criteria in Their Written Evaluations
Note: Both axes represent the ratio of total sentences that reviewers cite each criterion (1 to 12) in their evaluations. The diagonal line represents identical salience between fields. The off-diagonal line represents differential salience. Engineers (ICLR) and life scientists (LSJ) find accuracy+ (marker 1) equally salient, but life scientists find thoroughness– (marker 12) twice as salient (even as it is also most salient for engineers).

clarification. In the ICLR text, the reviewer takes issue with the paper's central claims; they argue that the experimental results are not sound evidence, offering alternative interpretations of the reported results. Moreover, the reviewer states the experiments were unclear ("puzzling") and the protocol "flawed." In the LSJ text, the reviewer seems more sanguine ("could enthusiastically be accepted"), but then itemizes several concerns about whether the reported data warrant the paper's main conclusions. Importantly, these are not framed as unsound evidence, just insufficient ("there is not a great deal of evidence to support . . ."). Still, the LSJ reviewer, like the ICLR reviewer, questions the soundness of the approach ("Methods and results seem to conflict") while also asking for clarification on key points. Both examples show the under-specification of "clarity" we use to proxy Kuhn's value of "simplicity." Here, reviewers draw on clarity to express a need for coherence in experimental design ("I find the experimental design puzzling") and straightforwardness of the report ("clearly shown"; "the . . . appears to be . . . + . . . + . . . which one and how much?").

How Reviewers Commensurate Epistemic Values

When a reviewer evaluates a manuscript, they make an array of value judgments and then prioritize these to make a recommendation to the editor. Not all possible criteria are applied in their review, and those that are raised are not equally weighted in their recommendation. Which criteria matter most, and do engineers and life scientists differ in how they attribute value to works based on these commensurated criteria?

Table 4 reports empirical tests addressing these questions. The first column for each field shows the baseline model with

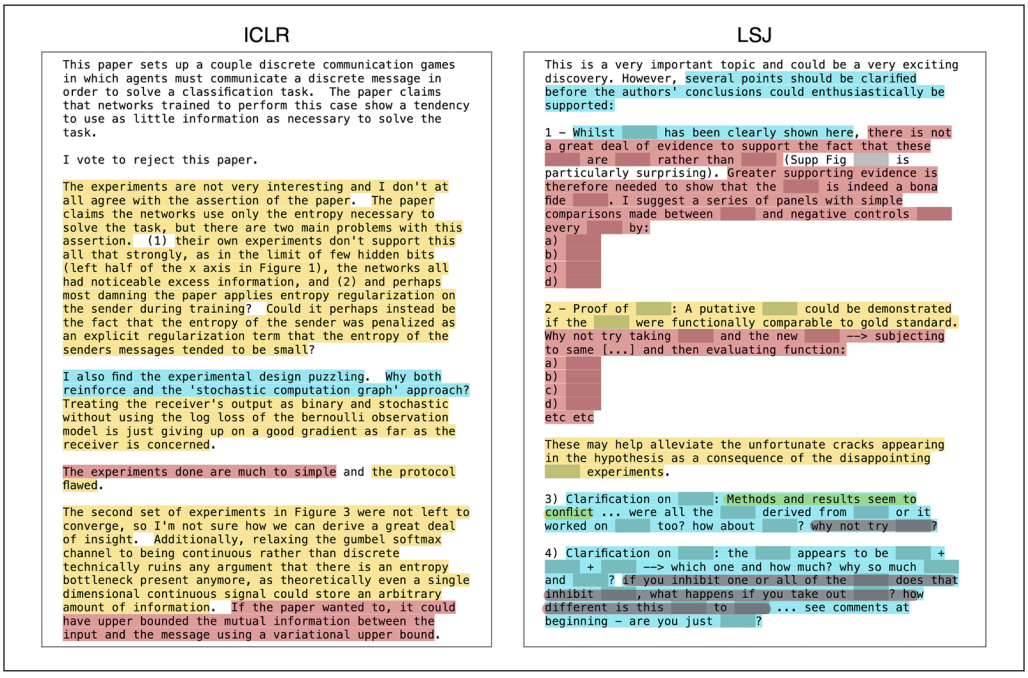

Figure 2. Two Reviews from ICLR and LSJ Showing Alignment in the Salience of Accuracy (–) (Yellow), Clarity (–) (Blue), and Thoroughness (–) (Red)
Note: Text is highlighted in color in the online version.

only manuscript fixed effects. The variation explained in reviewers' ratings by these models is attributable to the manuscript's unobserved and unobservable characteristics (e.g., its quality). We expect most of the variation to be explained by the work itself and interpret high *R*-squared values as strong indications that reviewers' ratings are tightly coupled with the work they evaluate. The second column for each field reports the full array of value judgments present in evaluations, so the effects of one value judgment are net of having made the others. The additional variation explained by these models is specific to reviewers' reasoning and interpretations of the manuscript. The third column for LSJ includes indicators of social bias.

When we include reviewers' value judgments, the proportion of overall variation explained in their recommendations increases by 17.2 and 16.9 percentage points, respectively, over the baseline models with just manuscript fixed effects for ICLR and LSJ. And we explain about 30 to 36 percent of the variation in recommendations between reviewers of the same manuscript ("within *R*-squared"). We interpret these values to mean our models are telling a substantial part of the story about how reviewers ascribe value to new works.[12]

Unambiguously, reviewers in our two venues align in their valuative acts. The tight, linear distribution of coefficients along the $y = x$ line in Figure 3A strongly indicates ($r = 0.957$) engineers and life scientists in these venues similarly commensurate various criteria. This finding is consistent with the similarity in the relative salience and provides further evidence against our expectation that reviewers in different fields would differ (Hypothesis 1). For engineers in ICLR and life scientists in LSJ, whether negatively or positively evaluated, what matters most is whether the work is sound (accuracy) and original (novelty), net of all other value judgments. To put the standardized coefficients

Table 4. Explaining Valuation: Panel OLS Fixed-Effects Regression Results Explaining Reviewer Ratings as an Outcome of Epistemic Value Judgements and Social Biases among all Manuscripts for ICLR and a Life Sciences Journal (LSJ)

	ICLR		LSJ		
Overall *R*-squared	.426	.598	.531	.700	.700
Between *R*-squared	.000	.424	.000	.474	.478
Within *R*-Squared	.000	.299	.000	.361	.361
N reviews	43,785	43,785	36,868	36,868	36,868
N manuscripts	12,626	12,626	16,096	16,096	16,096
Epistemic Judgments					
Accuracy (–)		–.134*** (.005)		–.237*** (.006)	–.238*** (.006)
Clarity (–)		–.044*** (.005)		–.007 (.006)	–.008 (.006)
Consistency (–)		–.106*** (.005)		–.070*** (.005)	–.070*** (.005)
Novelty (–)		–.219*** (.005)		–.217*** (.006)	–.217*** (.006)
Replicability (–)		–.027*** (.005)		–.020*** (.005)	–.020*** (.005)
Thoroughness (–)		–.056*** (.005)		–.047*** (.006)	–.047*** (.006)
Accuracy (+)		.097*** (.005)		.155*** (.006)	.155*** (.006)
Clarity (+)		.106*** (.005)		.101*** (.006)	.102*** (.006)
Consistency (+)		.044*** (.005)		.038*** (.006)	.038*** (.006)
Novelty (+)		.253*** (.005)		.233*** (.006)	.232*** (.006)
Replicability (+)		.009* (.004)		.014* (.006)	.014* (.006)
Thoroughness (+)		.132*** (.005)		.108*** (.006)	.108*** (.006)
Social Biases					
Author-reviewer citation proximity					.020* (.009)
Author-reviewer collaboration proximity					.008 (.007)
Non-Anglo-U.S. reviewer					–.018** (.005)
High-status reviewer					–.014* (.006)
Woman reviewer					–.000 (.005)

Note: Models include manuscript (and thereby author) fixed effects; baseline (null) models include only these, which account for the variation explained. Manuscript-clustered standard errors appear next to standardized coefficients. Epistemic value judgments are first scaled by the total length of the review (*n* sentences) and are proportions. LSJ models include the politeness and readability of the review as controls.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$ (two-tailed tests).

on novelty (–0.217 and –0.219 for ICLR and LSJ) in terms of real points: if the reviewer were to spend one-standard-deviation longer in terms of review length critiquing novelty, on average, that would correspond to a decrease in their rating by about 1.1 points in ICLR and 2.1 points in LSJ.[13] This is sizeable: it is more than enough to change their recommendation from "accept" (= 4) to "revise & resubmit" (= 3 or 2).

We also see some concerns are not as decisive for reviewers' recommendations: for example, replicability (–/+) and, unexpectedly, thoroughness (–) and clarity (–). Recall that these latter two criteria were the most salient in evaluations (Table 3), yet they are the *least* decisive in recommendations. This weak effect is also highlighted in Figure 3B: markers 4 and 12 (clarity– and thoroughness–) are among the farthest from the origin in terms of *x* (salience) but are substantively zero in terms of *y* (effect on recommendations). Holistically, the large variation in coefficient magnitudes seen up and down Table 4's columns is strong evidence that reviewers commensurate their value judgments, placing far greater weight on some (accuracy, novelty) over others (clarity). This

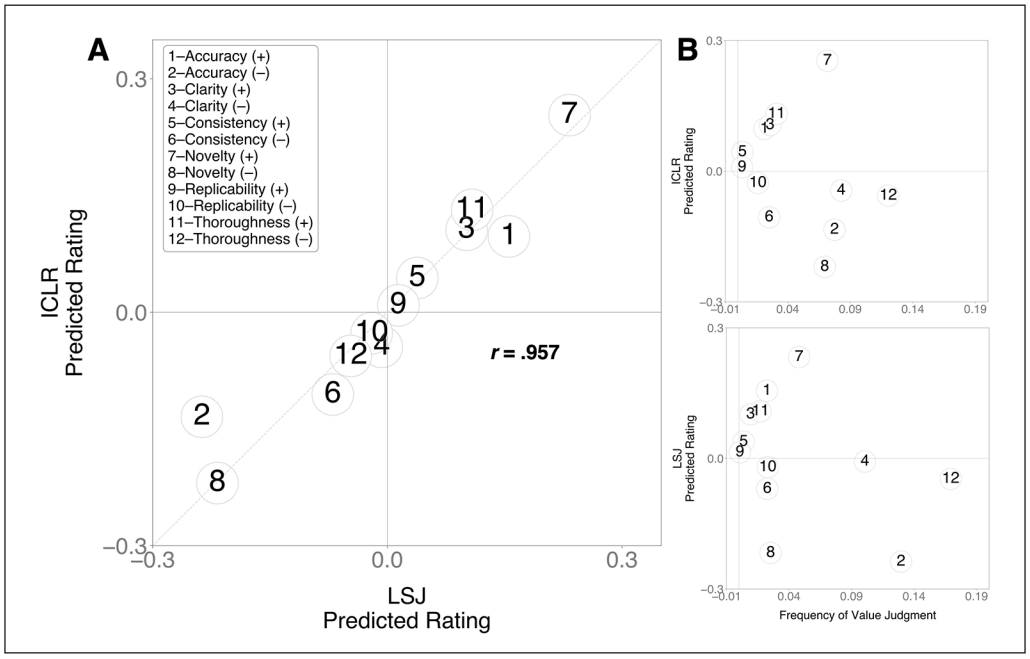

Figure 3. Explaining Valuation: The Strength of Reviewers' Value Judgments in Predicting Their Recommendations in ICLR versus LSJ (Panel A) and versus Their Respective Salience (Panel B)

Note: In Panel A, both axes represent the magnitudes of coefficients from Table 4. The diagonal line represents identical predictive strength. The off-diagonal line represents differential strength. So, replicability+ (marker 9) is equally unimportant to engineers (ICLR) and life scientists (LSJ), but accuracy– (marker 2) is relatively less decisive for engineers and more decisive for life scientists. In Panel B, the *y*-axes measure coefficient sizes, and the *x*-axes measure the ratio of total sentences reviewers cite each criterion in their evaluations. So, for life scientists (bottom plot), thoroughness– (marker 12) is cited very frequently but is not decisive in their overall recommendations.

commensuration reveals that salient criteria are not always decisive and is evidence in support of Hypothesis 2.

We analyze differences between ICLR and LSJ in Table 5, where we report the results of Wald tests comparing coefficient magnitudes from Table 4 and Figure 3. Generally, the differences are neither substantively large nor significant. Of those that are, most correspond to epistemic value judgments that are not decisive in reviewers' recommendations (clarity–, consistency–). For engineers in ICLR, clarity– tends to be relatively more decisive (d = 0.024 points = 0.037 SD × 0.66 points/SD), whereas accuracy– tends to be more decisive for life scientists in LSJ (d = 0.021 points = 0.237 SD × 0.089 points/SD). In terms of real points, however, the differential effect is not large enough to convince an editor in one venue to choose differently than an editor would in the other venue.

Turning to the LSJ columns in Table 4, we see evidence that reviewers are slightly more favorable in their evaluations of manuscripts whose authors are closer to them in citation networks, a rough proxy of intellectual homophily. Comparatively, the strength of this relationship (0.020) compensates for any concerns raised about the work's replicability, but it is an order of magnitude (10x) smaller than the average hit a manuscript takes when reviewers are concerned with accuracy and novelty. We also find evidence corroborating prior research that high-status reviewers tend to be, on average, slightly harsher

Table 5. Comparing Results: Wald χ^2 Tests Comparing Model Coefficients between ICLR and LSJ

	Estimated Effect on Reviewer Rating (Table 4)			Estimated Effect on Reviewer Influence (Table 6)		
Epistemic Value Judgment	$\hat{\beta}_{ICLR} - \hat{\beta}_{LSJ}$	Wald χ^2	p	$\hat{\beta}_{ICLR} - \hat{\beta}_{LSJ}$	Wald χ^2	p
Accuracy (–)	.103	13.210	.000***	–.085	–9.160	.000***
Clarity (–)	–.037	–4.562	.000***	.038	4.212	.000***
Consistency (–)	–.036	–4.874	.000***	.026	3.145	.002**
Novelty (–)	–.002	–.301	.763	–.020	–2.162	.031*
Replicability (–)	–.007	–.965	.335	.007	.834	.404
Thoroughness (–)	–.009	–1.200	.230	.007	.765	.444
Accuracy (+)	–.058	–7.546	.000***	.046	5.626	.000***
Clarity (+)	.005	.596	.551	–.011	–1.322	.186
Consistency (+)	.006	.735	.462	.007	.844	.399
Novelty (+)	.021	2.468	.014*	–.019	–2.112	.035*
Replicability (+)	–.005	–.732	.464	–.002	–.205	.838
Thoroughness (+)	.023	3.014	.003**	–.014	–1.736	.083

Note: Wald $\chi^2 = \frac{\hat{\beta}_{ICLR} - \hat{\beta}_{LSJ}}{\sqrt{SE_{ICLR}^2 + SE_{LSJ}^2}}$. Differences are computed and tested between the ICLR and LSJ coefficients reported in the values-only models in Tables 4 and 6 for reviewer and influence (all manuscripts). For each pair of coefficients, the Wald test statistic is computed with the formula above and the reported inverse probability is drawn using the cumulative density function of the χ^2 distribution with one degree of freedom and corresponds to observing a value as extreme as, or more extreme than, the reported statistic under the null hypothesis.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$ (two-tailed tests).

evaluators (Gallo et al. 2016). And we find reviewers outside the Anglo-American context tend to be slightly harsher. Nonetheless, comparing the second and third LSJ columns, we see the inclusion of reviewer-author bias metrics has virtually no effect on variation explained, nor are the coefficients on their value judgments markedly different. In Part R8 of the online supplement, we explore the interactions between social biases and evaluative acts among reviewers, and we find little evidence of collider effects. In all, this suggests relational biases among reviewers and authors play a very slight role in reviewers' recommendations to editors, which is inconsistent with our expectation (Hypothesis 5), but adds to a growing body of research that finds peer review remains robust to biases (Dahlander et al. 2023; Fox, Meyer, and Aimé 2023; Kern-Goldberger et al. 2022).

Together, we find that most of the epistemic values we observe in evaluations are significantly related to reviewers' ratings, that both positively and negatively evaluated criteria play a role, even when controlling for each in the model, and that reviewers tend to show no relational biases against authors. Moreover, we see in engineering and life science reviewers' discrete acts of valuation that they commensurate their judgments; the full range of criteria is relevant, but critical appraisals of accuracy and novelty are most decisive in recommendations to editors. The alignment between engineers and life scientists in these venues is unexpected and inconsistent with Hypothesis 2, which holds that fields differ in the criteria that are most decisive.

How Reviewers Influence Editors

When reviewers write their evaluations of manuscripts, they not only provide an account of their recommendation, but they also seek to influence the editor. Who is most persuasive? Which reviewer influences the editor most? Recall that our measure of a reviewer's *relative influence* on an editor ranges from negatively influential (i.e., the editor is propelled away from the focal reviewer to side with the others) to positively influential (i.e., the editor is pulled toward the focal reviewer's recommendation against the consensus). A reviewer can only have influence, then, when they make a recommendation opposing the other reviewers, and the editor sides with them and not the others. Table 6 (see columns "All") and Figure 4 report results of tests that explain variation in this relative influence.

The general pattern among positive and negative value judgments in Table 6 indicates that editors filter out positive evaluations when adjudicating peer review reports to decide whether to advance a manuscript. Editors are most influenced by the bad and ugly, even if other reviewers consensually recommend acceptance. This finding confirms Hypothesis 3. In both fields, editors, who already tend to make decisions that are lower than reviewers' recommendations (Table 2), are usually persuaded away from the consensus only when a reviewer makes a case against the manuscript, not vice versa. As with reviewer valuations, the linear distribution of coefficients along the $y = x$ line in Figure 4A strongly indicates ($r = 0.956$) that these two fields are more similar than different.

Some of these value judgments sway editors more than others. For example, engineering and life science editors are not persuaded by reviewers' positions on replicability and thoroughness. Instead, they are persuaded to reject manuscripts when reviewers raise concerns about a work's accuracy or novelty (markers 2 and 8 in Figure 4A). Negative value judgments concerning the originality and validity of new works are core to the epistemic cultures of engineers in ICLR and life scientists in LSJ; they determine the two most critical, decisive inflection points of the peer review process in both fields.

These findings also highlight important variation between the commonly held values reviewers voice in their evaluations and the values that exert the most influence on editors' final decisions (Figure 4B). Looking at frequencies alone (Table 3) would suggest

Table 6. Explaining Influence: Panel OLS Fixed-Effects Regression Results Explaining Reviewer Influence on Editorial Decisions as an Outcome of Epistemic Value Judgments and Social Biases among All and Contentious Manuscripts for ICLR and a Life Sciences Journal (LSJ)

	ICLR			LSJ					
	Consensus	All	Dissensus	Consensus	All	Dissensus	Consensus	All	Dissensus
Overall *R*-squared	.591	.562	.525	.556	.558	.599	.556	.558	.600
Between *R*-squared	.001	.002	.005	.047	.049	.050	.046	.049	.050
Within *R*-squared	.123	.195	.221	.131	.194	.289	.132	.194	.290
N reviews	8,731	43,785	8,724	7,349	36,868	7,362	7,349	36,868	7,362
N manuscripts	2,534	12,626	2,721	3,395	16,096	3,298	3,395	16,096	3,298
Epistemic Judgments									
Accuracy (–)	.092*** (.012)	.118*** (.005)	.156*** (.013)	.176*** (.022)	.203*** (.007)	.263*** (.016)	.176*** (.022)	.203*** (.008)	.263*** (.016)
Clarity (–)	.026 (.014)	.035*** (.006)	.045** (.013)	.023 (.023)	–.004 (.007)	.006 (.015)	.023 (.023)	–.003 (.007)	.009 (.015)
Consistency (–)	.057*** (.010)	.084*** (.005)	.118*** (.013)	.059*** (.016)	.058*** (.007)	.078*** (.015)	.060*** (.016)	.058*** (.007)	.079*** (.015)
Novelty (–)	.109*** (.012)	.175*** (.006)	.222*** (.014)	.125*** (.016)	.195*** (.007)	.280*** (.018)	.125*** (.016)	.195*** (.007)	.279*** (.018)
Replicability (–)	.010 (.010)	.016** (.005)	.013 (.012)	.011 (.016)	.009 (.006)	–.023 (.014)	.010 (.016)	.009 (.006)	–.021 (.014)
Thoroughness (–)	.019 (.011)	.028*** (.005)	.048*** (.012)	.011 (.020)	.021** (.007)	.039* (.015)	.010 (.020)	.021** (.007)	.040** (.015)
Accuracy (+)	–.058*** (.010)	–.074*** (.005)	–.109*** (.012)	–.147*** (.015)	–.120*** (.007)	–.120*** (.014)	–.148*** (.015)	–.120*** (.007)	–.120*** (.014)
Clarity (+)	–.057*** (.010)	–.080*** (.005)	–.112*** (.013)	–.072*** (.015)	–.069*** (.007)	–.078*** (.014)	–.073*** (.015)	–.069*** (.007)	–.079*** (.015)
Consistency (+)	–.023** (.009)	–.027*** (.005)	–.020 (.012)	–.064*** (.015)	–.034*** (.007)	–.032* (.015)	–.064*** (.015)	–.033*** (.007)	–.032* (.015)
Novelty (+)	–.203*** (.011)	–.193*** (.005)	–.160*** (.013)	–.200*** (.017)	–.174*** (.007)	–.164*** (.016)	–.200*** (.017)	–.173*** (.007)	–.163*** (.016)
Replicability (+)	–.002 (.009)	–.006 (.004)	–.001 (.011)	.005 (.019)	–.004 (.007)	–.012 (.015)	.005 (.019)	–.004 (.007)	–.012 (.015)
Thoroughness (+)	–.078*** (.010)	–.101*** (.005)	–.131*** (.012)	–.062*** (.016)	–.086*** (.006)	–.085*** (.014)	–.062*** (.016)	–.086*** (.006)	–.084*** (.014)
Social Biases									
Auth-Rev citation proximity							–.034 (.023)	–.020 (.010)	–.001 (.023)
Auth-Rev collaborative proximity							.006 (.019)	.005 (.008)	.004 (.018)
Ed-Rev citation proximity							.023 (.024)	.018 (.011)	–.019 (.026)
Ed-Rev collaborative proximity							–.006 (.021)	.000 (.009)	.011 (.021)
Non-Anglo-U.S. reviewer							.001 (.014)	.010 (.007)	.015 (.015)
High-status reviewer							–.004 (.017)	.005 (.008)	.028 (.019)
Woman reviewer							.003 (.014)	.001 (.006)	–.022 (.015)

Note: Models include manuscript (and thereby author) fixed effects; baseline (null) models include only these, which account for the variation explained. Manuscript-clustered standard errors appear next to standardized coefficients. Epistemic value judgments are first scaled by the total length of the review (*n* sentences) and are proportions. LSJ models include the politeness and readability of the review as controls.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$ (two-tailed tests).

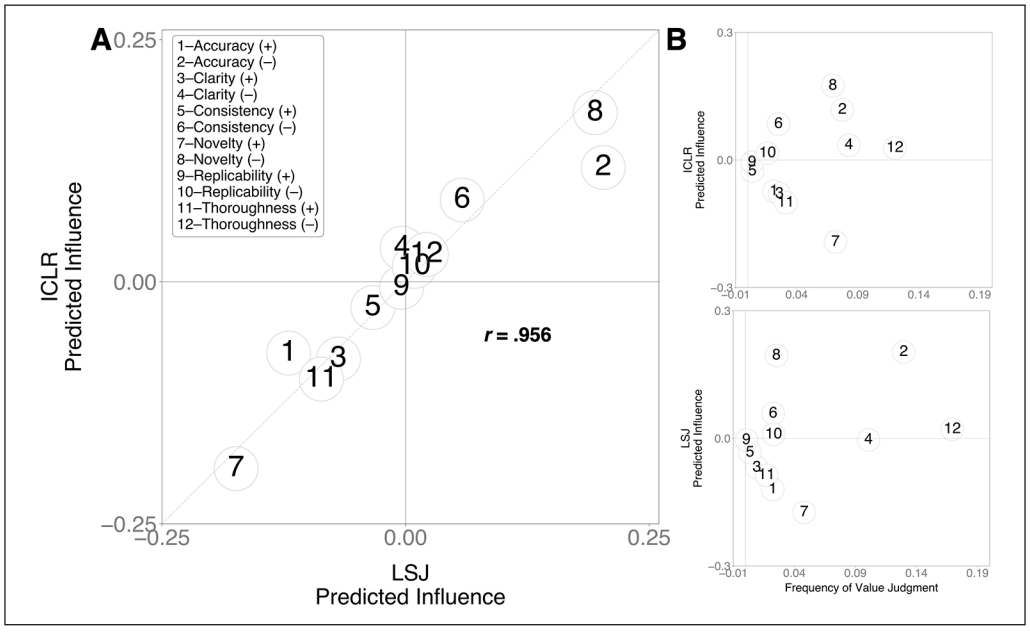

Figure 4. Explaining Influence: The Strength of Reviewers' Value Judgments in Predicting Their Influence on Editors' Decisions in ICLR versus LSJ (Panel A) and versus Their Respective Salience (Panel B)

Note: In Panel A, both axes represent the magnitudes of coefficients from Table 5. The diagonal line represents identical predictive strength. The off-diagonal line represents differential strength. So, replicability+ (marker 9) is roughly equally uninfluential for engineers (ICLR) and life scientists (LSJ), but accuracy– (marker 2) is slightly more influential for life scientists than for engineers. In Panel B, the *y*-axes measure coefficient sizes, and the *x*-axes measure the ratio of total sentences reviewers cite each criterion in their evaluations. For engineers (top plot), clarity– (marker 4) is cited more frequently but does not give reviewers all that more influence over editors' final decisions.

novelty matters less than thoroughness and clarity, as it is not as salient. This is not the case. When reviewers in LSJ critique a manuscript's clarity, they are *less* influential. Instead, accuracy and novelty again are most decisive, despite novelty's comparatively infrequent presence in reviewer evaluations. The epistemic culture of engineering and life sciences seen in these two venues is not simply salient, frequently held concerns among participants, but manifests as their relative authority in determining the outcome of peer review.

Returning to Table 5, we find many statistically significant differences between the two fields in the relative influence criteria have on editors' decisions (see right-hand side columns), but all are substantively quite small ($|\widehat{\beta}_{ICLR} - \widehat{\beta}_{LSJ}| = [0.007, 0.085]$). We interpret this as further evidence that the two fields are more similar than different, and that divergences are more stylistic than epistemological.

In Table 6 for the LSJ corpus, we again find relational biases have little influence on editors' final decisions. We interpret the non-significant, small magnitudes of the social bias terms, the consistency of the value judgment terms, and the consistency of *R*-squared across the "All" models in Table 6 as further evidence against Hypothesis 5. One counterargument here regarding reviewer influence, and above regarding reviewer ratings, is that our epistemic value judgments are colliders: including them in our models alongside social bias terms creates spurious associations. Our study of collider effects between biases and value judgments in Part R8 of the online supplement does not bear out this

counterargument and instead corroborates our finding that reviewer ratings and influence are relatively free of relational social biases in the review proceedings we analyze here.

Overall, these findings are consistent with our account of peer review in journal science: reviewers are most influential on editors' final decisions when they critically apply core epistemological concerns. Yet the tight linear distribution of the coefficients around the $y = x$ line in Figure 4A does not support our expectation that the two venues would be markedly different.

How Uncertainty Interacts with Valuation and Influence

We now explore how epistemic uncertainty alters reviewer influence on editors' decisions. In particular, we investigate whether the factors shaping editors' decisions shift when editors face pluralistic and conflicting comments from reviewers.

In Table 6, the center "All" columns represent all manuscripts, the left-side "Consensus" represents manuscripts where reviewers' written evaluations are the most similar, and the right-side "Dissensus" represents manuscripts with the most dissimilar evaluations. Read left to right, the reviews included in the analytic subsample are progressively more pluralistic and therefore more uncertain for the editor. The column-wise differences in coefficients tell us how the influence of a reviewer's evaluation on the final editorial decision *depends* on the uncertainty arising from the epistemic content of the other reviewers' evaluations.[14]

What we find is striking. The general pattern is that reviewers have even more influence when they make negative epistemic value judgments in contexts of greater editorial uncertainty. Sometimes the difference between consensus and dissensus is a factor of two, as is the case for consistency– in ICLR (0.057 versus 0.118) and novelty– in LSJ (0.125 versus 0.280). But other times the difference is smaller, as is the case for consistency– in LSJ (–0.059 versus 0.078). This reveals that although these concerns are normally decisive for editors (see "All" and "Consensus" columns), they become even more so when reviewers bring in a sundry mix of other concerns into their evaluations ("Dissensus"). In these settings, editors tend not to factor in these other, pluralistic judgments (especially the positive ones), but instead double down on the negative critiques, and especially accuracy, novelty, and consistency (for ICLR). Unexpectedly, social biases remain non-significant and small. Editors do not fall back on social cues when the decision is underdetermined by the epistemic value reviewers attribute to a manuscript.

Together, these findings support Hypothesis 4: as reviewers bring more and more diverse value judgments to bear in their evaluations of manuscripts, this presents a context of great ambiguity for editors. The decision to accept or reject is on a par, underdetermined by the mix of reviewer comments, and this increases the perceived risk of accepting a weakly supported and in some measure questionable paper. In this context, editors are most influenced by critical reviewers. In the face of uncertainty, editors rely more on reviewers' critical appraisals and not on social cues. Contrary to Hypothesis 6, bias matters even less.

Sensitivity Analyses

In Part R of the online supplement, we include the results of a battery of sensitivity analyses and summarize the results of investigating the most consequential threats to our study's internal validity.

Measurement error in ratings. Because we did not fine-tune our reviewer ratings classifier on human-labeled LSJ data due to issues of endogeneity, its performance in that corpus was not as good as its performance in the ICLR data, where it was trained (see Part A of the online supplement), as expected. To test for robustness, we computed the average per-label error, adjusted the predictions accordingly, and then re-ran all inferential models. Our results are consistent, suggesting noise introduced into LSJ ratings by an

untuned classifier does not cause inconsistent or unexpected results (see R1 in the online supplement). Moreover, in LSJ, because we use sentences to infer value judgments and reviews to infer ratings, there may be issues of endogeneity. We do not observe systematically stronger coefficients in LSJ versus ICLR, which would be clear evidence of endogeneity. Furthermore, our results after using a dictionary method to measure values (R2 in the online supplement) also suggest endogeneity plays no role.

Measurement error in epistemic values. Machine learning classifiers, which we use to infer the epistemic value judgments of reviewers in ICLR and LSJ, may introduce noise (see Part B of the online supplement). To assess the robustness of our findings to this noise, we substituted the classifiers with dictionary methods to measure the epistemic value judgments for both corpora. To do so, we developed reliable lists of keywords indicating each epistemic criterion. We then labeled review sentences based on the frequencies of these keywords and summed them to describe the review level. Finally, we reran all the inferential modeling. The results reported in Part R2 of the online supplement are substantively identical. We interpret this to be strong, conclusive evidence that our main findings are robust to measurement error generally introduced by our classifiers. Next, we ran 1,000 simulations on bootstrapped samples and estimated the variance in the resulting regression coefficients (see Part R7 of the online supplement). Again, we find corroboratory evidence that our main results are not sensitive to natural sampling variability. Finally, we ran 10,000 simulations on our performance metrics reported in Table B4 to assess their reliability; as reported in Part R6 of the online supplement, the benchmarks of our RoBERTa classifier in classifying epistemic value judgments are reliable and not sensitive to sampling error.

Reviewer selection into epistemic values. A core insight in the philosophical and sociological literatures of peer review is that reviewers are partial in their evaluations: based on their unique and idiosyncratic experiences and expertise, they selectively engage in epistemic criteria. This means that whatever concerns they raise and whatever ratings they give may be confounded by reviewer-specific unobserved variable bias. To address this, we reran our inferential models with both manuscript and reviewer fixed effects. Our results are consistent and suggest that reviewer selection into the epistemic concerns they draw on does not confound our main results (see Part R4 of the online supplement). We also explored the extent of collider bias in our estimates: namely, that reviewers raise different concerns based on who they are evaluating. To do so, we interacted all bias terms with our value judgment terms. We find little evidence that the main results are affected by collider bias (see Part R8 of the online supplement).

DISCUSSION

Critics and Gatekeepers

Peer review is a hierarchical decision-making arena, where scientists' different roles impinge on the relative meaning and weight they place on various criteria when deciding the outcome of review. Reviewers are critics who make a full range of judgments to attribute epistemic value to works. Reviewers also use these value judgments to persuade editors to take their recommendation, thereby seeking to exert influence on editors' decisions. We find both positive and negative judgments drive reviewers' valuation. In contrast, editors are gatekeepers; they receive multiple reviewer reports and must decide whether to advance or discontinue the work based on recommendations. We find editors overwhelmingly err on the side of skepticism and conform with the most critically negative of reviewers' comments when making their decisions, even when there is partial consensus among reviewers to accept. Thus, we show that peer review in our two STEM journals tends to be largely *unbalanced* and

partial, with a particular bent for eliminating works whose merit and value are questioned.

Salience, Valuation, Commensuration, and Influence

For reviewers, criteria of epistemic value are neither equally *salient* in evaluations nor equally valued. Indeed, a central and unexpected finding from our analyses is that the most salient concerns are not the most decisive for recommendations. If this were the case, we would have found that accuracy, clarity, and thoroughness were most decisive in reviewers' recommendations. While this is true for accuracy, it is certainly not true for clarity and thoroughness, which, although most frequently raised, are *least* consequential in reviewers' valuations. Above all other criteria, accuracy and novelty are the most powerful indicators of the value reviewers find in works. This contrast reflects a general social process, where reviewers differentially weigh criteria (seen in the variation of coefficients in Table 4). Through this process of commensurated valuation, we observe an important aspect of epistemic culture in action: it is expressed both as commonly held standards and as a discriminating act of ascribing value. Without this link to valuation, and without observing how reviewers commensurate criteria, we would be misguided in our understanding of scientific judgment and selection.

Valuation—how reviewers independently and privately arrive at their recommendations to the editor—is one thing. Reviewers' influence on editors' final decisions is altogether a different process. Editors must assess a reviewer's valuation and their judgments in relation to the other reviewers'. Yet, these recommendations are often divergent because they are adduced by differentially commensurated value judgments. We found that if a reviewer is ever to pull an editor away from the partial consensus of others in instances of divergent recommendations, it is by making negative value judgments about novelty and accuracy and recommending rejection, not the reverse. Epistemic culture in peer review manifests as a range of expressed values and discrete acts of valuation. But it is still more than that. It patterns the relative power differential among reviewers as they seek to influence the direction of the editor's final decision. We see how such values are not only a common frame of reference in evaluation but a resource that empowers actors to instantiate their own judgments as objective facts: as the value works have and as the recognition of knowledge claims as valid science.

Social Bias

Unbalanced though it may be, we still find peer review in our two venues seems fair. Nepotism, homophily, sexism, and ethnocentrism play a small role in shaping the advance of new works. Yet, the coefficients on these indicators of relational biases are small and rarely significant, and there is virtually no change in the explained variation in the review outcome when we add them to inferential models. Even in our robustness checks (see Parts R4 and R8 of the online supplement), we find little role of relational social biases. This finding corroborates recent research that increasingly finds bias plays a marginal role in peer review (Dahlander et al. 2023; Fox et al. 2023). Still, we must also note that editorial bias against authors based on their status is absorbed by manuscript fixed effects and is part of the variance explained in these models; our research design does not allow us to fully observe biases based on status characteristics of authors independently. Future work that develops robust independent variables jointly measuring manuscript qualities as well as author status and attributes would be a large advance in the field.

Plurality and Uncertainty

Contrary to past work that underscores the "unreliability" of divergent reviewers and emphasizes consensus among evaluators, we foreground that such divergence is reasonable but elevates uncertainty for editors.

Editorial uncertainty stems not only from opposing reviewer recommendations. It gets additionally compounded, not resolved, by the argumentative backings and justifications of these divergent recommendations (Toulmin 2003). This heighted uncertainty arises because reviewers independently draw on a plurality of standards of epistemic merit and value, depending on their own interpretations of the manuscripts, as well as their own diverse training and epistemological positions. Opposing recommendations, reject and accept, are warranted by a different constellation of commensurated value judgments across a diverse array of standards. Such a possibility of parity (Chang 2002) in the final outcome compounds the sense of ambiguity and uncertainty for editors. Past work suggests this underdetermination elevates the risk of "external" factors or bias playing a larger role in determining the science on hand.

A surprise from our work is that when editors face uncertainty, the authority of epistemic standards expands and the role of biases is effectively extinguished. We reckon that the uncertainty for editors induced by reviewers' epistemic plurality raises the risk of accepting false or erroneous scientific work, which in turn cues editors as gatekeepers to fall back even more on epistemic standards, in particular the negative evaluations of accuracy and novelty. Thus epistemic culture, as practiced in our venues, is heavily inflected by normative ideals of merit as well as an ethos of mitigating risk. This relationship may play out differently in other institutional settings of peer review. Editors of less prestigious and less exclusive outlets, for example, may give the benefit of the doubt to authors when there is partial consensus on acceptance (Bakanic et al. 1987; Zuckerman and Merton 1971).

For our two venues, we nonetheless speculate the presence of a distinctive customary rule (Lamont and Huutoniemi 2011) or convention (Becker 2008:29–30) that aids reviewers and editors in responding to uncertainty. The rule is to exercise an abundance of critical skepticism. The candidate truth claim is *neither valid nor new until proven otherwise and beyond a reasonable doubt* (Zuckerman and Merton 1971). This de facto presumption that works are not new and not true means reviewers are far more likely to doubt, question, and condemn than they are to confirm, affirm, or extol. It also means reviewers are more likely to weigh these critical concerns much more heavily in their discrete acts of valuation. And it means, at least for ICLR and LSJ, that specific values are epistemically elevated much more than others.

This focus on selecting the most original and accurate works may be reflective of peer review in elite STEM venues tasked with publishing only the most excellent and significant works. But we are less convinced this is the case. There is great variation in the relative influence of certain criticisms, and if selectivity were the principal explanation here, then we would expect *any* criticism as equal cause for doubt and rejection (Hypothesis 4). Editors' focus on these two master values above all others suggests to us more cultural alignment than mere exigencies of peer review.

Engineering and Life Sciences

That judgment in peer review is partial and reflective of individuals and distinctive epistemic subcultures across STEM fields motivated our initial expectation that we would find meaningful differences between engineers and life scientists in our two venues. While engineering fields have a general epistemological orientation toward advancing the state of the art in scalable and practical ways (Hansson 2007), life scientists tend to be oriented instead toward causal accounts that are thoroughly controlled (Snow [1959] 2012). We reasoned that this difference would manifest as different values or criteria being salient in each field. It was therefore surprising that peer review between the two venues unfolds so similarly. For example, we do not see issues of scalability (replicability) and benchmarking (consistency) being the most decisive priorities for AI engineers—they are not even *relatively* more decisive for them compared to life scientists. The converse is equally true for life scientists'

concern for rigorous controls and falsification (thoroughness).

Beyond the exigencies of elite publishing, we speculate a shared cognitive style could drive the robust overlaps in how values shape scientific advance. This shared style is steeped in positivism, which is anchored on quantitative and experimental techniques to falsify theoretically deduced hypotheses (Kuhn 1996; Mallard et al. 2009). This culture is likely further integrated by shared implicit assumptions about nature: namely, it exists, its existence is independent of the observing scientist, and scientific methodologies are optimal tools to understand it (Chakravartty 2017). We do not claim these commitments are true; that they are accounts our subjects consciously deploy; nor that they make STEM wholly unique (Snow [1959] 2012). Instead, we reckon but do not conclusively demonstrate that they constitute an a priori conventionalized frame that drives alignment and coherence across the evaluative proceedings of our two venues. Future work could take our unexpected findings as a point of departure to explore whether scientists across STEM fields do, in fact, exhibit such a shared cognitive style.

Functionalism and Constructivism

If science's expressed aim is the production of new and valid knowledge, then the reviewers and editors of ICLR and LSJ appear to engage works consistent with that aim. We do not find evidence of relational social bias. We do not find evidence that reviewers' valuations and editors' final decisions are loosely coupled from the content of evaluations or that the content of evaluations depends on the social identities of the participating scientists. Instead, we find editors and reviewers alike are chiefly focused on vetting works that are new and right.

Our findings support the view that the review proceedings of these two venues appear to work, but they do not warrant claims that it is natural; that it reduces uncertainty and fosters consensus; or that social and external factors do not matter. Our findings also do not warrant claims that scientific advance can be achieved only when "external" or "social" elements are removed. And they do not warrant claims that scientific facts produced by the evaluative proceedings we study are pure, unmediated expressions of reality. Contrary to these claims, which are stylized positions in the internalist/externalist debate in the sociology of science (Shapin 1992, 1995), our findings show the functionalism we observe in peer review depends on the contingency and uncertainty that arises from the social dynamics among scientists.

Peer review solicits independent and diverse reviewer interpretations of the same works. Reviewers commensurate their interpretations to justify often divergent recommendations to the editor. Neither the report under review nor the peer-review process itself is "objective" in this sense. On the contrary, the report cues deeply *subjective* and *pluralistic* judgments. In turn, this "unreliability" elevates the *uncertainty* about whether works are novel and right and results in the paradox of peer review. First, peer review is a largely subjective process because of the variably interpreted and applied "objective" criteria of merit. Second, this subjective character elevates and generalizes the risk that the work under consideration may, in fact, be neither true nor valid, the acceptance of which would result in a failure of peer review. As a result, critical reviewers are the most *influential*, and even more so when editors are uncertain.

Our results emphasize the plurality of meaning in standards and works; the role dissensus plays in elevating uncertainty; and the relative power that some actors have over others in determining editors' decisions. Ours is an account of social construction; yet our findings suggest the constructedness of peer review is not cause to categorically doubt the process or the works that gain recognition by successfully passing through it. This constructedness is the very mechanism by which our two venues maintain the legitimacy of peer review by being "impartial" and selecting works whose novelty and accuracy are beyond reasonable doubt.

Limitations and Possibilities

Our work is correlational and based on rigorously controlled analyses of observational data from the comprehensive peer review of two selective and prestigious publication venues (80,000 reviews of 28,000 manuscripts). These data do not represent peer review generally, or STEM fields, or even their respective fields of engineering and life science. Our claims are best read and understood as being reflective of these two venues, not others. This narrow representativeness was an insurmountable limitation in the current study, chiefly because access to private reviews and identifying information of all scientists for rejected and accepted works is still extremely limited. Should the current trend of opening up scientific peer review continue (Wolfram et al. 2020), then future systematic work could integrate more venues. Doing so would offer real advances. For example, venues could be sampled by their prestige and selectivity as well as their field. This would allow more thorough field comparisons while controlling for differences in the publishing pressures of different venues.

Another limitation is that our measurement of epistemic values remains at an abstract level. "Accuracy +" of what? The conclusions drawn? The characterization of past literature? "Novelty –" of what? The dataset? The instrument of observation? In other words, the two venues might emphasize similar criteria at this level of abstraction but diverge in their specific targets. Future work could improve on ours by developing richer criteria of evaluation that may help identify more inter-field heterogeneity than we are able to do with our coding scheme. We might conjecture that when an ICLR reviewer praises novelty, they are focusing on the instrument proposed. By contrast, an LSJ reviewer may be praising the discovery of a new biological mechanism for a treatment's efficacy. Criteria measured at the level of targets or claims would be a real advantage over our approach, but would likely be even noisier.

A feature of our design is our use of machine learning and GPT models to infer the epistemic content of reviews, which, to our knowledge has never been done before in sociological analysis of scientific knowledge-making. But this feature is not without limitations. The performance of our fine-tuned RoBERTa classifier is acceptable but not perfect. This concern is more relevant to our classification of ratings in LSJ because, to avoid issues of statistically induced endogeneity between our left- and right-sided variables, we did not fine-tune the classifier on LSJ texts. Analyzing variation in a variable with a high noise-to-signal ratio using the variation in another variable with a high noise-to-signal ratio leads to depressed estimates of any true relationships—the "iron law of econometrics." We have little cause to be concerned about noise, particularly compared to the inter-rater "reliability" among expert human evaluators, which meta-analyses have found ranges from 0.19 to 0.37 and 0.05 to 0.54 for ratings, and 0.12 to 0.28 for specific criteria (Cicchetti 1991; Jayasinghe et al. 2003; Lee 2012; Miller 2006). In our measurements, we observed a 0.84 correlation between predicted and true ratings (see Table A1 in the online supplement); and we saw an inter-rater reliability among human annotators of 0.53 to 1 and a machine-to-human F1 ranging from 0.6 to 0.84 for criteria (see Table B4 in the online supplement).

We believe the best way forward is to apply a relational approach to observe and represent the arguments reviewers make in their evaluations. In this study, we relied on single-label classifiers to identify concerns within sentences and to aggregate these up to the review level as frequencies. Future work might parse these arguments in a way that honors the hierarchical relations among claims, warrants, and backings. In so doing, this work could investigate how variation in the relational structures of reviewers' arguments influences reviewers' valuation and editors' final decisions. Indeed, future work on peer review may even ask a different

research question altogether based on this relational representation: how does variation in reviewers' argumentative claims about works relate to whether and how authors address these issues in their revisions? Such a question moves beyond the comparatively narrower view of social construction we study here and gets at scientific *co*-construction, which drives the revision manuscripts undergo during peer review.

Authors' Note

No text, idea, argument, or research design choice in this article was the input or output of a generative AI model. We used such models for controlled data augmentation, inference, and debugging. Everything else, including errors and omissions, is ours alone. DSS and DAM designed the research, wrote, and revised the paper; DSS, DAM, NNK, and TD contributed measures, data, and methods; and DSS and DAM analyzed data.

Data Note

We developed a replication package to enable independent verification of our machine learning benchmarks and full replication of our main inferential analyses. This package can be found here: https://github.com/danielscottsmith/valuesInScience.

Acknowledgments

We thank Kyle Mahowald for his generous assistance with early machine learning models and exploratory analysis of the LSJ data. We also thank participants of the Stanford Computational Sociology Workshop, as well as Patricia J. Gumport and anthony lising antonio for helpful feedback on an early draft. Finally, we thank the anonymous reviewers and editors for their thorough, insightful, and constructive feedback that strengthened the science reported here.

Funding

Funding from NSF awards 2022435 and 2244804 supported this research.

Editors' Note

To avoid any possible conflict of interest, the *ASR* Editors were not involved in the evaluation of this paper. The entire review process was handled by a Deputy Editor who is unaffiliated with the University of Massachusetts-Amherst.

ORCID iDs

Daniel Scott Smith https://orcid.org/0000-0002-0979-4009
Tianyu Du https://orcid.org/0000-0003-1141-157X
Daniel A. McFarland https://orcid.org/0000-0002-6805-0798

Notes

1. https://www.scimagojr.com/journalrank.php?order=h&ord=desc&type=j
2. Following Kuhn's (1977:330–31) discussion of "values" versus other terms—"criteria," "rules," "maxims," and "norms"—we use epistemic values and epistemic criteria interchangeably. We emphasize, like Kuhn and philosophers of science dealing with the induction problem (McMullin 1982:9–18), that values *influence* choice but do not algorithmically determine it.
3. For example, Sage, which publishes more than 1,000 scholarly journals across the humanities, social, natural, and medical sciences, instructs its reviewers to "be careful not to make judgments about the paper based on personal, financial, intellectual biases, or any other considerations than the quality of the research and the written presentation of the paper." See "04 Ethics and Responsibility" (https://us.sagepub.com/sites/default/files/how_to_become_a_reviewer_new.pdf).
4. In 2018, ICLR switched from a single- to double-masked policy. Our empirical models account for this with manuscript fixed effects.
5. ICLR's peer review is hosted by OpenReview.net. Under the platform's standard workflow, based on ICLR guidelines, reviewers have access only to their own reviews before the discussion period. Once the discussion commences, access is opened to all (https://docs.openreview.net/workflows/conferences#:~:text=Guide%20based%20on%20ICLR%20venue).
6. These logged versions are freely available but must be requested from OpenReview with the permissions of each conference year's program chairs and board members.
7. https://huggingface.co/huggingface
8. RMSE is root mean squared error and is a standard metric typically used to evaluate regression model performance; it quantifies in the scale of the dependent variable (in this case, ratings) how "off" the model's predictions are, averaged across all values of the predictions. In Part A of the online supplement, we also compute the intraclass correlation and kappa statistics and clarify what these tell us when we compare the model to human labels (and to benchmarks reported on human raters).
9. We study the outcomes of the first round of peer review. This enables us to focus on the initial evaluations and on the editor's first, pivotal decision of

whether or not the work should advance. This first round occurs after the desk decision and before later revision rounds. The salience of various values is likely to vary by review round in complex ways: first, because different stages may cue different criteria (fit earlier on; perhaps clarity later on); second, because one stage depends on the exigencies of the previous one (reviewers' and editors' concerns in round $n + 1$ are influenced by those in round n).

10. We fit this model to each corpus separately using the PanelOLS method of the Python module linearmodels (version 6).
11. In this section, we use "engineers" and "life scientists" as shorthand for "engineers in ICLR" and "life scientists in LSJ."
12. Because these are panel OLS models with manuscript fixed effects, between-manuscript variation in ratings is less meaningful by design, as we do not theorize or test any manuscript-level explanations for differences in their *averaged* reviewer recommendations or influence (i.e., the between-effects estimator).
13. To convert standardized coefficients into the original Y scale, we first use $\beta = \beta^* \times \frac{SD_y}{SD_x}$:

    For ICLR:

    $$\beta = \beta^*_{ICLR} \times \frac{SD_{rating_{ICLR}}}{SD_{novelty(-)ICLR}}$$

    $$\beta = -.219 \times \frac{.660}{.068}$$

    $$\beta = -.219 \times 9.706$$

    $$\beta = -2.126$$

    For LSJ:

    $$\beta = \beta^*_{LSJ} \times \frac{SD_{rating_{LSJ}}}{SD_{novelty(-)LSJ}}$$

    $$\beta = -.217 \times \frac{.960}{.045}$$

    $$\beta = -.217 \times 21.333$$

    $$\beta = -4.629$$

    To compute the expected change in rating due to 15 sentences about novelty(–) added to an average-length review, we use: $\Delta rating = \beta \times \Delta novelty(-)$:

    For ICLR:

    $$\Delta x = 15 sents \div 29.147 sents$$

    $$\Delta x = .515$$

    $$\Delta y = -2.126 \times .515$$

    $$\Delta y = -1.095$$

    For LSJ:

    $$\Delta x = 15 sents \div 32.237 sents$$

    $$\Delta x = .465$$

    $$\Delta y = -4.629 \times .465$$

    $$\Delta y = -2.152$$

14. Note that our measures of consensus have nothing to do with their discrete recommendations (i.e., the dependent variable in Table 4), only their value judgments (i.e., the independent variables).

References

Bakanic, Von, Clark McPhail, and Rita J. Simon. 1987. "The Manuscript Review and Decision-Making Process." *American Sociological Review* 52(5):631–42 (https://doi.org/10.2307/2095599).

Bakanic, Von, Clark McPhail, and Rita J. Simon. 1989. "Mixed Messages: Referees' Comments on the Manuscripts They Review." *The Sociological Quarterly* 30(4):639–54 (https://doi.org/10.1111/j.1533-8525.1989.tb01540.x).

Barlösius, Eva. 2019. "Concepts of Originality in the Natural Science, Medical, and Engineering Disciplines: An Analysis of Research Proposals." *Science, Technology, & Human Values* 44(6):915–37 (https://doi.org/10.1177/0162243918808370).

Becker, Howard S. 2008. *Art Worlds: Updated and Expanded*. Berkeley: University of California Press.

Bordage, Georges. 2001. "Reasons Reviewers Reject and Accept Manuscripts: The Strengths and Weaknesses in Medical Education Reports." *Academic Medicine* 76(9):889–96.

Bornmann, Lutz, and Hans-Dieter Daniel. 2010. "The Manuscript Reviewing Process: Empirical Research on Review Requests, Review Sequences, and Decision Rules in Peer Review." *Library & Information Science Research* 32(1):5–12.

Bornmann, Lutz, Rüdiger Mutz, and Hans-Dieter Daniel. 2010. "A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants." *PLoS ONE* 5(12):e14331 (https://doi.org/10.1371/journal.pone.0014331).

Bourdieu, Pierre. 1974. "Cultural Reproduction and Social Reproduction." Pp. 56–68 in *Knowledge, Education and Cultural Change: Papers in the Sociology of Education*, edited by R. Brown. London, UK: Tavistock Publications.

Bourdieu, Pierre. 1988. *Homo Academicus*. Stanford, CA: Stanford University Press.

Brezis, Elise S., and Aliaksandr Birukou. 2020. "Arbitrariness in the Peer Review Process." *Scientometrics* 123(1):393–411 (https://doi.org/10.1007/s11192-020-03348-1).

Carson, Richard T., Joshua Graff Zivin, Jordan J. Louviere, Sally Sadoff, and Jeffrey G. Shrader. 2022. "The Risk of Caution: Evidence from an Experiment." *Management Science* 68(12):9042–60 (https://doi.org/10.1287/mnsc.2021.4292).

Cetina, Karin Knorr. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Chakravartty, Anjan. 2017. "Scientific Realism." *Stanford Encyclopedia of Philosophy*, edited by E. N.

Zalta (https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/).

Chang, Ruth. 2002. "The Possibility of Parity." *Ethics* 112(4):659–88 (https://doi.org/10.1086/339673).

Cheng, Mengjie, Daniel Scott Smith, Xiang Ren, Hancheng Cao, Sanne Smith, and Daniel A. McFarland. 2023. "How New Ideas Diffuse in Science." *American Sociological Review* 88(3):522–61 (https://doi.org/10.1177/00031224231166955).

Chubin, Daryl E., and Edward J. Hackett. 1990. *Peerless Science: Peer Review and U.S. Science Policy*. Albany: State University of New York Press

Cicchetti, Domenic V. 1991. "The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation." *Behavioral and Brain Sciences* 14(1):119–35 (https://doi.org/10.1017/S0140525X00065675).

Coronel, Ruben, and Tobias Opthof. 1999. "The Role of the Reviewer in Editorial Decision-Making." *Cardiovascular Research* 43(2):261–64 (https://doi.org/10.1016/S0008-6363(99)00177-7).

Crane, Diana. 1967. "The Gatekeepers of Science: Some Factors Affecting the Selection of Articles for Scientific Journals." *The American Sociologist* 2(4):195–201.

Dahlander, Linus, Arne Thomas, Martin W. Wallin, and Rebecka C. Ångström. 2023. "Blinded by the Person? Experimental Evidence from Idea Evaluation." *Strategic Management Journal* 44(10):2443–59 (https://doi.org/10.1002/smj.3501).

Demarest, Bradford, Guo Freeman, and Cassidy R. Sugimoto. 2014. "The Reviewer in the Mirror: Examining Gendered and Ethnicized Notions of Reciprocity in Peer Review." *Scientometrics* 101(1):717–35 (https://doi.org/10.1007/s11192-014-1354-z).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Pp. 4171–86 in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), edited by J. Burstein, C. Doran, and T. Solorio. Minneapolis, MN: Association for Computational Linguistics.

Druckman, James N. 2022. "Threats to Science: Politicization, Misinformation, and Inequalities." *ANNALS of the American Academy of Political and Social Science* 700(1):8–24 (https://doi.org/10.1177/00027162221095431).

Eagly, Alice H., and Steven J. Karau. 2002. "Role Congruity Theory of Prejudice toward Female Leaders." *Psychological Review* 109(3):573–98 (https://doi.org/10.1037/0033-295X.109.3.573).

Eagly, Alice H., Mona G. Makhijani, and Bruce G. Klonsky. 1992. "Gender and the Evaluation of Leaders: A Meta-Analysis." *Psychological Bulletin* 111(1):3–22 (https://doi.org/10.1037/0033-2909.111.1.3).

Espeland, Wendy Nelson, and Mitchell L. Stevens. 2008. "A Sociology of Quantification." *European Journal of Sociology* 49(3):401–36 (https://doi.org/10.1017/S0003975609000150).

Flaminio, Squazzoni, Bravo Giangiacomo, Farjam Mike, Marusic Ana, Mehmani Bahar, Willis Michael, Birukou Aliaksandr, Dondio Pierpaolo, and Grimaldo Francisco. 2022. "Peer Review and Gender Bias: A Study on 145 Scholarly Journals." *Science Advances* 7(2):eabd0299 (https://doi.org/10.1126/sciadv.abd0299).

Fleck, Ludwik. 1979. *Genesis and Development of a Scientific Fact*. Chicago: The University of Chicago Press.

Fox, Charles W., Jennifer Meyer, and Emilie Aimé. 2023. "Double-Blind Peer Review Affects Reviewer Ratings and Editor Decisions at an Ecology Journal." *Functional Ecology* 37(5):1144–57 (https://doi.org/10.1111/1365-2435.14259).

Gallo, Stephen A., Joanne H. Sullivan, and Scott R. Glisson. 2016. "The Influence of Peer Reviewer Expertise on the Evaluation of Research Funding Applications." *PLOS ONE* 11(10):e0165147.

Glonti, Ketevan, Daniel Cauchi, Erik Cobo, Isabelle Boutron, David Moher, and Darko Hren. 2019. "A Scoping Review on the Roles and Tasks of Peer Reviewers in the Manuscript Review Process in Biomedical Journals." *BMC Medicine* 17(1):118 (https://doi.org/10.1186/s12916-019-1347-0).

Grover, Aditya, and Jure Leskovec. 2016. "Node2vec: Scalable Feature Learning for Networks." Pp. 855–64 in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (https://doi.org/10.1145/2939672.2939754).

Guetzkow, Joshua, Michèle Lamont, and Grégoire Mallard. 2004. "What Is Originality in the Humanities and the Social Sciences?" *American Sociological Review* 69(2):190–212 (https://doi.org/10.1177/000312240406900203).

Hansson, Sven Ove. 2007a. "Praxis Relevance in Science." *Foundations of Science* 12(2):139–54 (https://doi.org/10.1007/s10699-006-0003-2).

Herrera, Antonio J. 1999. "Language Bias Discredits the Peer-Review System." *Nature* 397(6719):467 (https://doi.org/10.1038/17194).

Horne, Christine, and Stefanie Mollborn. 2020. "Norms: An Integrated Framework." *Annual Review of Sociology* 46(1):467–87 (https://doi.org/10.1146/annurev-soc-121919-054658).

Hsieh, Nien-hê, and Henrik Andersson. 2021. "Incommensurable Values." *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta (https://plato.stanford.edu/archives/fall2021/entries/value-incommensurable/).

Hug, Sven E., and Michael Ochsner. 2022. "Do Peers Share the Same Criteria for Assessing Grant Applications?" *Research Evaluation* 31(1):104–17 (https://doi.org/10.1093/reseval/rvab034).

Iezzoni, Lisa I. 2018. "Explicit Disability Bias in Peer Review." *Medical Care* 56(4):277–78.

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: Series A (Statistics in Society)* 166(3):279–300 (https://doi.org/10.1111/1467-985X.00278).

Johnson, Cathryn, Timothy J. Dowd, and Cecilia L. Ridgeway. 2006. "Legitimacy as a Social Process." *Annual Review of Sociology* 32:53–78.

Kagan, Jerome. 2009. *The Three Cultures: Natural Sciences, Social Sciences, and the Humanities in the 21st Century*. Cambridge, UK: Cambridge University Press.

Kennard, Neha Nayak, Tim O'Gorman, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Rajarshi Das, Hamed Zamani, and Andrew McCallum. 2021. "DISAPERE: A Dataset for Discourse Structure in Peer Review Discussions" (https://doi.org/10.48550/arxiv.2110.08520).

Kern-Goldberger, Adina R., Richard James, Vincenzo Berghella, and Emily S. Miller. 2022. "The Impact of Double-Blind Peer Review on Gender Bias in Scientific Publishing: A Systematic Review." *American Journal of Obstetrics & Gynecology* 227(1):43–50.e4 (https://doi.org/10.1016/j.ajog.2022.01.030).

Krieger, Joshua, Kyle Myers, and Ariel Dora Stern. 2023. "How Important Is Editorial Gatekeeping? Evidence from Top Biomedical Journals." Harvard Business School Entrepreneurial Management, Working Paper No. 22-011 (https://dx.doi.org/10.2139/ssrn.3882949).

Kuhn, Thomas S. 1977. "Objectivity, Value Judgement, and Theory Choice." Pp. 320–39 in *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Kuhn, Thomas S. 1996. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Lamont, Michèle. 2015. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Lamont, Michèle. 2012. "Toward a Comparative Sociology of Valuation and Evaluation." *Annual Review of Sociology* 38(1):201–21 (https://doi.org/10.1146/annurev-soc-070308-120022).

Lamont, Michèle, and Katri Huutoniemi. 2011. "Comparing Customary Rules of Fairness: Evaluative Practices in Various Types of Peer Review Panels." Pp. 209–32 in *Social Knowledge in the Making*, edited by C. Camic, N. Gross, and M. Lamont. Chicago: The University of Chicago Press.

Lamont, Michèle, and Grégoire Mallard. 2005. "Peer Evaluation in the Social Sciences and the Humanities Compared: The United States, the United Kingdom, and France." Report prepared for the Social Sciences and Humanities Research Council of Canada.

Lane, Jacqueline N., Misha Teplitskiy, Gary Gray, Hardeep Ranu, Michael Menietti, Eva C. Guinan, and Karim R. Lakhani. 2022. "Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation." *Management Science* 68(6):4478–95 (https://doi.org/10.1287/mnsc.2021.4107).

Langfeldt, Liv. 2004. "Expert Panels Evaluating Research: Decision-Making and Sources of Bias." *Research Evaluation* 13(1):51–62.

Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.

Laudan, Larry. 1984. *Science and Values: The Aims of Science and Their Role in Scientific Debate*. Berkeley: University of California Press.

Lee, Carole J. 2012. "A Kuhnian Critique of Psychometric Research on Peer Review." *Philosophy of Science* 79(5):859–70.

Lee, Carole J. 2015. "Commensuration Bias in Peer Review." *Philosophy of Science* 82(5):1272–83 (https://doi.org/10.1086/683652).

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in Peer Review." *Journal of the American Society for Information Science and Technology* 64(1):2–17 (https://doi.org/10.1002/asi.22784).

Lin, Jialiang, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. "Automated Scholarly Paper Review: Concepts, Technologies, and Challenges." *Information Fusion* 98:101830 (https://doi.org/10.1016/j.inffus.2023.101830).

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach" (https://doi.org/10.48550/arXiv.1907.11692).

Longino, Helen. 1983. "Beyond 'Bad Science': Skeptical Reflections on the Value-Freedom of Scientific Inquiry." *Science, Technology, & Human Values* 8(1):7–17.

Luukkonen, Terttu. 2012. "Conservatism and Risk-Taking in Peer Review: Emerging ERC Practices." *Research Evaluation* 21(1):48–60 (https://doi.org/10.1093/reseval/rvs001).

Machamer, Peter, and Lisa Osbeck. 2004. "The Social in the Epistemic." Pp. 78–89 in *Science, Values, and Objectivity*, edited by P. Machamer and G. Wolters. Pittsburgh, PA: University of Pittsburgh Press.

Mallard, Grégoire, Michèle Lamont, and Joshua Guetzkow. 2009. "Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review." *Science, Technology, & Human Values* 34(5):573–606 (https://doi.org/10.1177/0162243908329381).

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(2):151–69.

McMullin, Ernan. 1982. "Values in Science." *PSA: Proceedings of the Biennial Meeting of the Philosophy of*

Science Association 1982:3–28 (http://www.jstor.org/stable/192409).

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27(1):415–44.

Merton, Robert. 1973. "The Normative Structure of Science." Pp. 267–78 in *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: The University of Chicago Press.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* 159(3810):56–63 (https://doi.org/10.1126/science.159.3810.56).

Miller, C. Chet. 2006. "From the Editors: Peer Review in the Organizational and Management Sciences: Prevalence and Effects of Reviewer Hostility, Bias, and Dissensus." *Academy of Management Journal* 49(3):425–31.

National Research Council. 1978. *Peer Review in the National Science Foundation: Phase One of a Study*. Washington, DC: The National Academies Press (https://doi.org/10.17226/20041).

Nielsen, Mathias Wullum, Christine Friis Baker, Emer Brady, Michael Bang Petersen, and Jens Peter Andersen. 2021. "Meta-Research: Weak Evidence of Country- and Institution-Related Status Bias in the Peer Review of Abstracts." *eLife* 10:e64561 (https://doi.org/10.7554/eLife.64561).

Oreskes, Naomi. 2019. *Why Trust Science?* Princeton, NJ: Princeton University Press.

Roumbanis, Lambros. 2022. "Disagreement and Agonistic Chance in Peer Review." *Science, Technology, & Human Values* 47(6):1302–33 (https://doi.org/10.1177/01622439211026016).

Sandström, Ulf, and Martin Hällsten. 2008. "Persistent Nepotism in Peer-Review." *Scientometrics* 74(2):175–89 (https://doi.org/10.1007/s11192-008-0211-3).

Scott, William A. 1974. "Interreferee Agreement on Some Characteristics of Manuscripts Submitted to the Journal of Personality and Social Psychology." *American Psychologist* 29(9):698–702 (https://doi.org/10.1037/h0037631).

Searle, John. 2018. "Constitutive Rules." *Argumenta* 1(7):51–54.

Shapin, Steven. 1992. "Discipline and Bounding: The History and Sociology of Science as Seen through the Externalism-Internalism Debate." *History of Science* 30(4):333–69 (https://doi.org/10.1177/007327539203000401).

Shapin, Steven. 1995. "Here and Everywhere: Sociology of Scientific Knowledge." *Annual Review of Sociology* 21(1):289–321 (https://doi.org/10.1146/annurev.so.21.080195.001445).

Siler, Kyle, Kirby Lee, and Lisa Bero. 2015. "Measuring the Effectiveness of Scientific Gatekeeping." *Proceedings of the National Academy of Sciences* 112(2):360–65 (https://doi.org/10.1073/pnas.1418218112).

Sizo, Amanda, Adriano Lino, Luis Paulo Reis, and Álvaro Rocha. 2019. "An Overview of Assessing the Quality of Peer Review Reports of Scientific Articles." *International Journal of Information Management* 46:286–93 (https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2018.07.002).

Smith, Olivia M., Kayla L. Davis, Riley B. Pizza, Robin Waterman, Kara C. Dobson, Brianna Foster, Julie C. Jarvey, et al. 2023. "Peer Review Perpetuates Barriers for Historically Excluded Groups." *Nature Ecology & Evolution* 7(4):512–23 (https://doi.org/10.1038/s41559-023-01999-w).

Snow, Charles Percy. [1959] 2012. *The Two Cultures*. Cambridge, UK: Cambridge University Press.

Squazzoni, Flaminio, Giangiacomo Bravo, Mike Farjam, Ana Marusic, Bahar Mehmani, Michael Willis, Aliaksandr Birukou, Pierpaolo Dondio, and Francisco Grimaldo. 2021. "Peer Review and Gender Bias: A Study on 145 Scholarly Journals." *Science Advances* 7(2) (https://doi.org/10.1126/sciadv.abd0299).

Steel, Daniel. 2010. "Epistemic Values and the Argument from Inductive Risk." *Philosophy of Science* 77(1):14–34 (https://doi.org/10.1086/650206).

Teplitskiy, Misha, Daniel Acuna, Aïda Elamrani-Raoult, Konrad Körding, and James Evans. 2018. "The Sociology of Scientific Validity: How Professional Networks Shape Judgement in Peer Review." *Research Policy* 47(9):1825–41 (https://doi.org/https://doi.org/10.1016/j.respol.2018.06.014).

Terama, Emma, Melanie Smallman, Simon J. Lock, Charlotte Johnson, and Martin Zaltz Austwick. 2017. "Correction: Beyond Academia – Interrogating Research Impact in the Research Excellence Framework." *PLOS ONE* 12(2):e0172817 (https://doi.org/10.1371/journal.pone.0172817).

Toulmin, Stephen E. 2003. *The Uses of Argument*. Cambridge, UK: Cambridge University Press.

Travis, G. D. L., and H. M. Collins. 1991. "New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System." *Science, Technology & Human Values* 16(3):322–41 (https://doi.org/10.1177/016224399101600303).

Wennerås, Christine, and Agnes Wold. 2010. "Nepotism and Sexism in Peer-Review." Pp. 64–70 in *Women, Science, and Technology*, edited by M. Wyer, M. Barbercheck, D. Cookmeyer, H. Ozturk, and M. Wayne. New York: Routledge.

Wolfram, Dietmar, Peiling Wang, Adam Hembree, and Hyoungjoo Park. 2020. "Open Peer Review: Promoting Transparency in Open Science." *Scientometrics* 125(2):1033–51 (https://doi.org/10.1007/s11192-020-03488-4).

Yuan, Weizhe, Pengfei Liu, and Graham Neubig. 2022. "Can We Automate Scientific Reviewing?" *Journal of Artificial Intelligence Research* 75:171–212 (https://doi.org/10.1613/jair.1.12862).

Zhuang, Zhenzhen, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. "Large Language Models for Automated Scholarly Paper Review: A Survey." *Information Fusion* 124:103332 (https://doi.org/10.1016/j.inffus.2025.103332).

Zuckerman, Harriet, and Robert K. Merton. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System." *Minerva* 9(1):66–100 (https://doi.org/10.1007/BF01553188).

Daniel Scott Smith is an Assistant Professor of Sociology at Duke University, where he studies the social foundations of science. He holds a PhD from Stanford University.

Neha Nayak Kennard is a PhD candidate in Computer Science at the University of Massachusetts Amherst. Her research is on the use of natural language processing techniques both to study and to support the peer review process.

Tianyu Du is a PhD student in the Institute of Computational and Mathematical Engineering (ICME) at Stanford University, whose research integrates machine learning, data science, economics, and sociological perspectives to explore the labor market.

Daniel A. McFarland is a Professor of Education, and by courtesy, Sociology and Organizational Behavior, at Stanford University. He completed his PhD at the University of Chicago. His current work focuses on topics in the sociology and philosophy of science, as well as relational sociology.