FIELD EXPERIMENTS IN THE SCIENCE OF SCIENCE: LESSONS FROM PEER REVIEW AND THE EVALUATION OF NEW KNOWLEDGE

Revised January 2026

Kevin J. Boudreau, Northeastern & NBER[1]

Scientific evaluation and peer review govern the allocation of resources and certification of knowledge in science yet has been subjected to limited causal investigation. This chapter synthesizes randomized experiments embedded in live peer-evaluation systems at journals, conferences, and funding agencies, restricting attention to published studies. I organize this evidence using a Q–A–R–S framework that decomposes peer review into submission attributes (Q), author attributes (A), reviewer composition and expertise (R), and features of the evaluation system and institutional context (S), interpreting outcomes through a view of peer review as an information-processing institution in which reviewers form assessments using the existing stock of knowledge. The chapter treats experimental design choices as objects of analysis, assessing what existing interventions can—and cannot—identify about peer review under practical institutional constraints. Randomized experimentation is indeed feasible and currently spans the full Q–A–R–S space, albeit sparsely. I make several claims based on analytic synthesis: (1) interventions often affect reviewer or author behavior with little impact on core evaluative judgments; (2) evaluations are shaped primarily by reviewer identity and expertise; and (3) peer review functions more reliably as a filter of unacceptable submissions than as a fine-grained ranking mechanism among higher-quality contributions. Overall, the evidence and synthesis here does not support the view—often implied in the literature—that peer review is highly ineffective. Instead, it points to a functioning institution operating under binding constraints, with substantial scope for institutional innovation. The chapter concludes by identifying priorities for future high-yield experimentation, emphasizing system-level interventions, the role of automation in filtering tasks, and the distinct challenges of improving ranking reliability at the knowledge frontier.
<u>Keywords</u>: *Science of science; Research evaluation; Randomized controlled trials; Causal inference; Institutional design; Bounded rationality; Knowledge, Scientific institutions.*
JEL <u>Codes</u>: O31, O38, D83, C93, I23, D91, H43

---

1   INTRODUCTION

Modern scientific evaluation and peer review are relatively recent institutional arrangements. Many practices now treated as standard—routinized external evaluation, anonymized or double-blind assessment, standardized scoring rubrics, and formal decision timelines, and more—were adopted gradually and only became widespread in many fields in recent decades. Contemporary peer review should therefore not be treated as a settled benchmark, but rather as an evolving institutional response to considerable pressures.[2] At the same time, the institution is frequently criticized as slow, costly, unreliable, and perhaps biased (e.g., Smith, 1999; Smith, 2006; Godlee, 2012; Alberts et al., 2014). This raises questions regarding the scope to which the institution might continue to innovate an improve, apart from more basic questions regarding whether or to what extent to which these criticisms are fair. These questions might be addressed by study of workings and mechanisms shaping the performance of peer review. There indeed exists a vast empirical literature on scientific evaluation (National Academies of Sciences, Engineering, and Medicine, 2018; Bornmann, 2011; Squazzoni et al., 2021). However, the bulk of this work is descriptive or correlational, documenting outcomes without isolating the mechanisms that generate them. A much smaller experimental literature embeds randomized or quasi-randomized interventions directly within live peer-evaluation systems at journals, conferences, and funding agencies. To date, however, this experimental literature remains sparse and fragmented, spread across multiple disciplines and institutional contexts, and has yet devoted limited attention to generating cumulative or generalizable insights (Section 3). In this chapter, I draw together and provide an analytic synthesis of existing experiments in peer review with the goal of clarifying what they do—and do not—allow us to conclude about the limits of peer review, and of returning explicitly to the question of where there may be elasticity and scope for innovation in the performance of this institution.

This chapter draws together and interprets the experimental evidence on scientific evaluation, restricting attention to studies published in peer-reviewed academic journals (i.e., excluding unpublished working papers, conference proceedings, and internal evaluations, or studies not using embedded interventions within live systems.) Through 2025, I identify 52 such studies (Table 1). The study set was assembled using searches in Google Scholar, PubMed, Web of Science, supplemented by backward and forward citation tracing from prior syntheses (e.g., Bruce et al. 2016). Table 1 reports

---

[2] I do not enumerate these pressures here.

the final list reviewed here. The resulting corpus closely tracks experiments identified in prior literature reviews (Figure 1).

Nonetheless, the chapter is not intended as an exhaustive bibliometric exercise or a conventional review. The objective and approach instead is analytic synthesis: close reading of a subset of archetypal experimental designs to assess what different interventions do—and do not—identify about the operation of scientific evaluation. For each archetypal study, I examine not only the formal research design but also how the intervention was operationalized in practice—who partnered with the research site, where the randomization entered the evaluation workflow, and which institutional constraints shaped compliance, measurement, and implementation. I then assess what can—and cannot—be inferred from these studies, both individually and in the aggregate.

In doing so, the chapter treats experimental design choices themselves as objects of analysis—shaped by institutional context, access, and collaboration with research sites—and emphasizes the trade-offs these choices imply for identification, inference, and external validity. Because most studies in this literature were not designed with generalization as a primary objective, a central aim is to assess what can nonetheless be learned by interpreting their findings collectively, while remaining explicit about the fundamental limits such inference faces.

As a central organizing device, I introduce a Q–A–R–S framework (Section 2) that relates evaluation outcomes to submission attributes and true quality (Q), author attributes (A), reviewer attributes and expertise (R), and features of the peer-review evaluation system and institutional context (S). The framework serves both as an accounting device for sources of variation in observed outcomes and as a classification system for the space of feasible experimental interventions, given institutional and operational constraints. Crucially, it also provides a common language for interpreting a heterogeneous body of experimental evidence collectively rather than study by study.

I then connect this framework to a baseline characterization of peer review as an information-processing institution operating under uncertainty, in which evaluators form noisy, forward-looking judgments using imperfect signals drawn from an existing—and necessarily incomplete—stock of knowledge. This characterization disciplines cross-study interpretation by clarifying which empirical regularities should be expected even in well-functioning systems, and which patterns plausibly reflect the effects of specific interventions.

<Table 1> <Figure 1 > <Figure 2 >

Taken together, the experimental evidence examined in this chapter provides some indication of where peer review appears responsive to intervention. The studies rarely measure cost or delay

directly, and they do not observe true scientific quality against which reliability or bias could be assessed. What they do reveal is that many interventions generate behavioral responses—affecting participation, effort, or process—while evaluative judgments themselves often appear less responsive. At the same time, evaluations do change along certain margins, particularly with respect to who evaluates and what information is made salient. These patterns point to scope for innovation and improvement in peer review—a theme returned to and developed in the synthesis of the final section.

The chapter also helps situate where future experimental work may be most productively focused. By examining research designs and empirical results within a common analytical framework, the discussion highlights how choices about intervention margins, outcome measurement, and research design shape what can be learned, and why some directions for further research may ultimately prove more informative than others.

The chapter proceeds as follows. Section 2 develops the Q–A–R–S framework as a baseline characterization of the data-generating process in peer-review studies, providing a common structure for the discussion that follows. Section 3 reviews experimental evidence, organized into subsections by the primary margin of variation—submission characteristics (Q), author information (A), reviewer composition and expertise (R), and features of the evaluation system and institutional context (S). Section 4 synthesizes this evidence, develops the chapter's central claims, and identifies implications for future experimentation and institutional design, using the accumulated findings to highlight experimental margins where additional work is likely to yield higher inferential returns under real institutional constraints.

The chapter may be read from start to finish as a single, integrated analysis, providing an brief and economical path to understanding the disparate experimental advances in this area. Alternatively, it is written to permit modular reading, allowing individual sections to be read independently depending on the reader's interests.

2   BASELINE CHARACTERIZATION OF THE PEER REVIEW DATA GENERATING PROCESS

This section develops a baseline characterization of the peer-review data-generating process (DGP) that organizes and disciplines interpretation of the experiments that follow. Rather than offering a full theory or normative evaluation, the focus is deliberately narrow: a small set of structural features that shape outcomes under normal operating conditions. These features generate predictable

patterns in observed evaluations—even absent bias or institutional dysfunction—and place binding constraints on what empirical designs can identify.

2.1 A Simple Framework for Sources of Variation in Peer-Review: Q-A-R-S

Let reviewer *i* evaluate submission *j*. The observed evaluation—whether a score, ranking, recommendation, or decision—can be represented as being related to several broad components:

- **Q (Submission Attributes & Quality)**: attributes of the submission's underlying contribution and expected scientific merit.
- **A (Author Attributes)**: observable and latent characteristics of the author(s), such as training, affiliation, prior work, and reputation.
- **R (Reviewer Attributes)**: reviewer expertise, perspective, and epistemic position.
- **S (System Attributes)**: features of the review system and institutional context, including assignment rules, information structures, and aggregation procedures.

The observed evaluation can be represented as follows:

$$\text{evaluation}_{ij} = g(Q_j, A_j, R_i, S) + \varepsilon_{ij}$$

where *g(·)* denotes an unknown mapping from these components to observed evaluations and ε_{ij} captures idiosyncratic variation.

Here, "evaluation" is used broadly to denote any observed outcome of the review process, including individual reviewer scores, rankings, recommendations, error detection, or final acceptance decisions. The system component S bundles institutional features—such as assignment rules, anonymity, and aggregation—that determine how identical underlying inputs may map to different observed outcomes under different procedures.

The framework is agnostic with respect to accuracy, bias, efficiency, or dysfunction, and does not privilege any component as normatively relevant. Rather, it provides a common language for identifying sources of variation in review outcomes and for clarifying what empirical designs do—and do not—hold fixed. In practice, the Q–A–R–S framework functions as a diagnostic checklist for interpreting correlations and experimental contrasts in peer-review data.

It is also important to note that each component Q_j, A_j, R_i, and S is, by construction, a composite object rather than a scalar. Each bundles multiple, correlated attributes that are jointly produced and jointly interpreted in evaluation. Formally,

$$Q_j = Q(q_{1j}, q_{2j}, \dots), A_j = A(a_{1j}, a_{2j}, \dots), R_i = R(r_{1i}, r_{2i}, \dots), S = S(s_1, s_2, \dots),$$

2.2 Several Regularities of the Data Generating Process

While the Q–A–R–S framework is purely accounting, it is useful to recognize a small number of structural regularities that arise naturally from the task of evaluating new scientific knowledge. These regularities follow from basic epistemic and organizational conditions rather than from particular design choices or institutional failures. They therefore generate predictable patterns in observed evaluations and place fundamental constraints on inference.

The discussion here abstracts from incentive structures and governance arrangements that vary across settings and are taken up explicitly in later sections. Instead, the goal is to identify regularities that should be expected even in well-functioning systems and that discipline interpretation across otherwise heterogeneous experimental designs.

2.2.1 Structural Regularity 1: Evaluating New Knowledge under Uncertainty (Q)

A defining feature of peer review is that it evaluates claims whose true scientific quality is not fully knowable at the time of assessment. This is true not only for grant proposals, where outcomes are explicitly prospective, but also for completed journal manuscripts. Completion resolves uncertainty about execution, but it does not resolve uncertainty about scientific validity, generality, or value. Reviewers do not directly verify results through replication or reanalysis; instead, they assess plausibility, rigor, and coherence using indirect signals such as methodological choices, argumentation, and consistency with existing knowledge (Lamont, 2009; Hirschauer, 2010).

Several central dimensions of scientific quality—such as importance, generality, and long-run impact—are inherently forward-looking and cannot be observed at the moment of review (Bornmann, 2011). Even correctness is often established only through subsequent replication, critique, and cumulative use over time (Popper, 1959; Merton, 1973; Ioannidis, 2005). Peer review therefore does not measure realized quality, even for completed work; it forms predictions about expected scientific merit under conditions of persistent uncertainty (Kuhn, 1962).

This limitation is structural rather than incidental. Any genuine scientific contribution aims to extend existing knowledge and therefore lies, at least in part, beyond the current epistemic frontier (Popper, 1959; Kuhn, 1962; Polanyi, 1966; Merton, 1973). Reviewers must evaluate such contributions using the prevailing stock of methods, evidence, and theoretical understanding that the contribution itself seeks to advance (Hirschauer, 2010; Lamont, 2009; Csiszar, 2019). In informational terms, peer review evaluates new knowledge on the basis of existing knowledge.

Implication. Because true quality Q is unobservable at the time of review, substantial variance in assessments is to be expected even among well-qualified reviewers, along with attenuated and context-dependent relationships between ex ante evaluations and ex post outcomes. This uncertainty is greatest among submissions near the knowledge frontier, where evaluative cues are least informative.

2.2.2 Structural Regularity 2: Specialization and Structured Disagreement among Reviewers (R)

A second defining characteristic of peer review follows from the epistemic specialization of modern science. Expertise relevant to frontier problems is necessarily narrow, cumulative, and unevenly distributed. Even when research lies within a broadly defined field, understanding of particular problems reflects long, path-dependent investments in specific methods, datasets, conceptual frameworks, or empirical contexts (Jones, 2009).

Research on expert cognition shows that such expertise is highly domain-specific and shapes not only what individuals know, but also how they perceive, organize, and interpret information (Chi, Feltovich, & Glaser, 1981; Ericsson & Smith, 1991). As a result, experts—even when nominally similar in training—may attend to and treat as salient different, partially overlapping subsets of submission attributes, that is, different elements of

$$Q_j = Q(q_{1j}, q_{2j}, \dots).$$

In relatively mature or well-established areas of inquiry, shared training and evaluative conventions can generate substantial overlap in how submissions are read and assessed, leading to correlated judgments. As evaluation approaches the knowledge frontier, however, these shared anchors weaken. Small, accumulated differences in training and research trajectory—often invisible in coarse field classifications—can translate into systematic differences in perception, emphasis, and interpretation. These differences are amplified by the high-dimensionality of frontier research, where complex combinations of methods, data, and theory require selective attention and judgment.

Implication. Disagreement among reviewers should therefore be expected and need not reflect noise or bias. Rather, disagreement may be structured and systematic, reflecting heterogeneity in expertise—especially for submissions near the frontier.

2.2.3 Structural Regularity 3: Non-Separability of Attributes in Submissions, Authors, and Reviewers (Q, A, R)

A defining structural feature of peer review is that the core objects entering evaluation—submission attributes Q, author attributes A, and reviewer attributes R—are not composed of

separable elements that can be independently varied or interpreted in isolation. Instead, each consists of bundled, internally correlated attributes that function jointly as signals in evaluation.

For authors, observable characteristics such as institutional affiliation, disciplinary field, seniority, training environment, professional networks, prior publication experience, and demographic traits are tightly intertwined. No single author attribute is meaningfully observed or interpreted on its own; each is read as part of a broader bundle $A = (a_1, \ldots, a_k)$. An analogous structure holds for reviewers. Expertise, methodological orientation, epistemic position, prior exposure to related work, evaluative standards, and taste co-move, making it neither feasible nor conceptually coherent to vary a single reviewer attribute while holding others fixed.

Crucially, the same non-separability applies to submissions themselves. The attributes that constitute submission quality—such as novelty, importance, framing, methodological rigor, execution, data quality, theoretical contribution, and integration with prior literature—are jointly produced and jointly interpreted. In naturally occurring scientific work, these elements are inseparable aspects of a single contribution rather than independent dimensions that can be cleanly isolated. As a result, there is no meaningful sense in which one can experimentally vary a single component of Q while holding the remainder of the submission "constant."

Some experimental designs do manipulate perceived submission quality—for example, by injecting errors, fabricating results, resubmitting previously published work, or altering outcomes while holding text fixed. Such interventions can be informative, but they necessarily generate highly stylized or artificial versions of Q. These manipulations do not isolate individual attributes of scientific quality as they arise in frontier research; instead, they impose perturbations whose interpretation requires caution.

Implication. The non-separability of Q, A, and R places a fundamental constraint on causal interpretation in peer-review research. Observed associations between evaluations and particular attributes cannot, in general, be interpreted as the "effect" of that attribute alone, because the attribute cannot be experimentally assigned independent of other attributes.

2.2.4 Structural Regularity 4: Endogenous Matching across Submissions, Authors, Reviewers, and Review Systems (Q–A–R–S)

A distinct but equally fundamental regularity of peer review is that submissions, authors, reviewers, and review systems are not randomly paired, but are instead brought together through endogenous matching processes shaped by subject matter, expertise, reputation, institutional constraints, and capacity.

In practice, peer-review systems rely on coarse proxies—such as field classifications, prior publications, institutional affiliations, or keyword matching—to assign reviewers to submissions. Authors likewise sort submissions into venues with characteristic reviewer pools, editorial practices, and evaluative norms. Given the scarcity of relevant expertise and the importance of heterogeneity even among reviewers nominally within the same field, such matching is unavoidable. There is no meaningful counterfactual in which a generic population of evaluators assesses a generic population of submissions independent of context.

As a result, systematic covariation across the components of the peer-review data-generating process should be expected. Cross-component correlations such as

$$\text{Cov}(Q, A), \text{Cov}(Q, R), \text{Cov}(A, R), \text{ and } \text{Cov}(R, S)$$

may arise mechanically from assignment and sorting processes, even in the absence of bias, manipulation, or institutional dysfunction.

Endogenous matching also implies that the composition of reviewer pools, the distribution of expertise brought to bear on a submission, and the effective operation of review systems will differ systematically across venues, fields, and time periods. Consequently, empirical findings about peer review—especially those derived from particular journals, conferences, or funding agencies—may reflect properties of the local matching environment as much as general features of evaluation.

This regularity places a distinct constraint on inference. Even when individual attributes of Q, A, or R were conceptually separable, endogenous matching would still limit the extent to which observed correlations can be interpreted causally or generalized across settings.

Implication. Endogenous matching limits both causal attribution and external validity. Correlations between evaluations and author, reviewer, or system characteristics cannot be interpreted as causal effects without strong assumptions about the endogenous appearance of Q, A, R, and S.

2.2.5 Structural Regularity 5: Unobservability of True Quality (Q) and Use of Auxiliary Cues (A)

A derivative implication of unobservable quality and non-separable attributes is that reviewers may rationally draw on auxiliary informational cues beyond the content of the submission itself—such as author reputation, institutional affiliation, or prior work—when forming expectations of scientific merit. Prior to double blinded review, this was viewed as a feature rather than a problem with peer review (Merton, 1973; Shapin, 1994; Csiszar, 2019). This logic persists in modern institutions in settings where evaluation targets individuals rather than discrete projects. Programs such as the

Howard Hughes Medical Institute Investigator Program or the MacArthur Fellowships explicitly condition evaluation on investigator identity rather than particular submissions.

Implication. With unobservable true quality Q, reviewers may take cues beyond the submission itself to attempt to improve their predictions of submission quality.

3 WHAT'S BEEN TRIED: RANDOMIZED INTERVENTIONS, CONTROLLED COMPARISONS, AND RCTS

Across decades of research on peer review, only a small number of studies have moved beyond descriptive analysis to deliberate intervention embedded within live evaluative systems, with the aim of identifying causal effects. This section examines archetypal examples of these studies to illustrate what different experimental designs attempt, what they reveal, and what they leave unresolved.

The section proceeds by grouping studies according to the primary margin of variation. It begins with early re-evaluation and audit studies that attempt—imperfectly—to hold submission quality (Q) fixed, then turns to randomized manipulations of reviewer composition and expertise (R), author identity and blinding (A), reviewer training and feedback (R), and system-level features of review architecture and information flow (S). Incentive-based interventions are discussed separately. Figure 3 summarizes the distribution of studies across these categories.

Throughout, the Q–A–R–S framework is used to clarify what each design can—and cannot—identify causally. Not all interventions qualify as fully controlled experiments. Nonetheless, these studies collectively establish both the feasibility and limits of randomized experimentation in peer review and provide the empirical foundation for the synthesis that follows.

<Figure 3>

3.1 Proto-Experiments in Re-Evaluation: Attempts to Hold Submission Characteristics (Q) Fixed

The earliest empirical attempts to study peer review relied on a simple but revealing design: send the same work back through the evaluation system and observe what happens. Re-evaluation or resubmission studies treat a previously reviewed manuscript or proposal as a fixed reference point for quality, allowing researchers to examine the stability of evaluative judgments when nominally identical material is assessed again. Read through the Q–A–R–S framework, these studies attempt—necessarily imperfectly—to hold the underlying quality of the submission (Q) fixed, while observing how outcomes change as reviewers (R) and possibly other factors change. Their contribution is not causal

identification, but demonstration: they made evaluative instability visible and empirically undeniable. Several of these studies are among the most influential and well-cited in the literature.

Early Re-Evaluation of Psychology Journal Peer Review with Shocking Results: Peters and Ceci (1982)

One of the earliest and most influential examples was Peters and Ceci's (1982) "Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again," published in Behavioral and Brain Sciences, which remains one of the most widely cited and debated empirical probes of journal peer review, and which has since garnered over 1,400 citations. The authors resubmitted 12 previously published papers to the same 12 (undisclosed) psychology journals in which the papers had originally been accepted and published 18–32 months earlier. Each paper had originally included at least one author from a top-10 psychology department, and each target journal operated with non-blinded reviews. The resubmissions were identical in scientific content but featured fictitious, low-status author identities and institutional affiliations designed to remove reputational and institutional signals associated with the original authors (deliberately varying cues in (A) while holding the manuscript text fixed). As later clarified by Ceci and Peters, the published study represents only one half of an originally intended symmetric experimental design. The authors initially planned to resubmit manuscripts that had been rejected under low-status identities using the names of well-known authors at prestigious institutions, thereby permitting a two-sided test of author-status effects, but data collection for this second arm was curtailed following editorial detection and subsequent institutional intervention. The resulting design should therefore be understood as a truncated, one-sided experiment imposed by institutional constraints rather than as a complete realization of the authors' original experimental intent.[3]

Of the 12 resubmissions, three were detected as duplicates and the review process was discontinued. Of the nine that received reviews, eight were rejected, with 16 of 18 referees recommending rejection across the nine submissions. The article provoked immediate controversy and shock within psychology and beyond, as it appeared to suggest substantial unreliability in peer review and raised the possibility that outcomes were driven by author prestige rather than scientific

---

[3] Peters and Ceci first presented the study as a "manuscript masquerade" in *The Sciences* (1980), emphasizing its illustrative intent. In subsequent commentary responding to critics (Peters & Ceci 1982, *Behavioral and Brain Sciences*), they clarified the study's aims, ethical tradeoffs, and interpretive limits. A retrospective account by Ceci and Peters (2014, *The Winnower*) documents the originally intended symmetric design, the institutional resistance encountered, and how the study has been misinterpreted over time.

content. Over a dozen open commentaries and an authors' reply debated the ethics, interpretation, and generalizability of the design in the same journal, underscoring both the study's salience and the discomfort it generated within the scholarly community.

The reversal of 8 of 9 decisions was surprising and disturbing on its face. In historical context, the near-uniformity of rejection—particularly for work previously deemed publishable—was understandably galvanizing. However, it is nonetheless useful to distinguish the study's conceptual insight from features of the design that complicate interpretation. The most striking element of this research design was the unusual manipulation of institutional affiliation. This manipulation, however, does not break the usual principle that an individual author attribute (a_i) cannot be experimentally varied without simultaneously varying other correlated author attributes (Section XX). The study replaced known figures from top-10 psychology departments with unknown names from fictitious organizations such as "Tri-Valley Center for Human Potential" (the example given in the paper). Such affiliations do not resemble even marginal R1 or R2 research universities and would plausibly convey a strong negative bundled signal to reviewers. Even with this artificial manipulation and given the inherent uncertainty surrounding a submission's true quality (Q), reviewers would reasonably infer that an unknown identity and affiliation were correlated with other unobserved attributes—such as presumed ability, resources, or training environment. These inferred correlations would naturally inform expectations of true (Q) and thus affect scores. In this way, the Peters–Ceci design underscores a deeper principle: even when one attribute (a_i) is deliberately varied, reviewers interpret it as a bundled signal about broader author characteristics (A), meaning that correlations among attributes cannot be fully broken in practice.

The second notable feature of this design was the attempt to control for true (Q) by resubmitting the same papers. However, there are reasons to expect that the quality inferred from the submission itself could have diminished in the roughly 30–50 months between the original submission (prior to publication) and the resubmission. In that time, the perceived contribution could plausibly have drifted downward as the field advanced; reference lists and framing might also have appeared dated. Further, among the nine resubmitted papers that proceeded to review, reviewers who recognized or suspected prior publication may have handled the case by recommending rejection rather than explicitly raising plagiarism concerns—an ambiguity that cannot be ruled out given the available evidence.

Apart from differences in multiple attributes of (A) and in perceived (Q), referee characteristics (R) necessarily differed as well. Even random differences in referee specialization (Property 3) can

generate outcome variation in such a small sample. Beyond random sampling variation, there may also have been systematic differences in referee assignment. The observed sample excludes cases in which original reviewers were reselected (as those would likely have triggered plagiarism detection). As a result, the resubmitted manuscripts were necessarily evaluated by a different—and potentially non-random—subset of referees. It is also plausible that manuscripts submitted several years after their original appearance—by unknown authors from obscure affiliations—were assigned to referees with different levels of seniority, patience, or stringency than timely submissions from top departments. Such endogenous matching and correlations are plausible and further limit clean attribution (Section 2).

Across the explanatory factors in expression (1)—(Q), (A), (R), and (S)—the dimension most successfully held constant is arguably the peer-review system (S). This was accomplished by resubmitting papers to the same journals in which they had originally appeared. The authors explicitly note, for example, that they discarded one resubmission because the journal's review policy had changed.[4]

There should also be considerable noise in the review of these psychology papers, apart from the systematic differences noted above. There may be substantial uncertainty in (Q) even under ideal conditions (Property 1). Consistent with the authors' predictions, the change in author identity (A) removed any positive auxiliary signal supporting favorable evaluation (Property 4). We should also expect variation generated by differences in reviewers (Property 3). If the original acceptances were, in part, the result of favorable draws, the second-round rejection rate of 89% (8 of 9) is not obviously inconsistent with baseline journal rejection rates of roughly 80%, as noted by the authors. This observation does not overturn the study's central insight but cautions against overinterpreting the magnitude of the observed reversal.

The experiment's simplicity—resubmitting the same work—remains elegant, intuitively graspable, and unmatched in rhetorical impact. Viewed through the Q–A–R–S framework, however, the study demonstrates the existence of instability in peer-review outcomes without identifying its source. By construction, variation in author cues, reviewer assignment, inferred quality, and evaluative context move together, sharply circumscribing what the design can be taken to show. At the same

---

[4] The selection procedure for the 13 originally chosen articles (the 12 mentioned as part of resubmission and this 13th that was ruled out) is not entirely clear; however, it appears the 12 papers were chosen given their appearance in 12 distinct journals and subfields of psychology.

time, the study foreshadows many of the methodological and institutional tensions that would shape later peer-review experiments.

Large-Scale Re-Evaluation that Raised Questions about NSF Review: Cole et al. (1981)

Cole, Cole, and Simon's (1981) *"Chance and Consensus in Peer Review"* reports an audit study conducted in the context of NSF grant review. As context, during the 1970s, NSF peer review became the object of heightened congressional attention, with critics questioning whether federal research funding was allocated fairly or instead reflected insular professional networks. These concerns—often articulated through highly visible critiques of NSF-funded projects and culminating in congressional hearings portraying peer review as an "old boy" system—created pressure on the National Academy of Sciences to assess the operation of peer review using empirical evidence rather than rhetoric.

In response, the National Academy of Sciences—through its Committee on Science and Public Policy (COSPUP)—sought investigators who could credibly examine peer review from within the scientific enterprise itself. Rather than commissioning external critics, COSPUP turned to scholars whose work was already recognized as empirically rigorous and theoretically grounded. A relationship was formed with Robert Merton's sociology of science program at Columbia University, which by the early-to-mid 1970s had become a focal point for empirical research on the norms and reward structures of science. Within that program, Jonathan Cole, Stephen Cole, and Gary Simon were part of a research group—spanning senior scholars and junior researchers—working closely with Merton and Harry Zuckerman on a series of studies examining how scientific institutions allocate recognition, resources, and authority.

The primary interest of this group of collaborators at the time was that engagement with NSF peer review offered an opportunity to empirically examine core theories in the sociology of science, rather than a belief that peer review represented a particularly dysfunctional institution. The collaborators became involved in a multi-year program of research and consultation with the National Academy of Sciences and NSF, including descriptive analyses of large samples of NSF proposals and interviews with NSF personnel, prior to eventually undertaking the re-evaluation exercise described here.[5] The 1981 study thus represents the culmination of a longer sequence of theory-driven empirical inquiry.

---

[5] The re-review experiment reported in *Chance and Consensus* was preceded by an earlier phase of COSPUP-sponsored research on NSF peer review. As described by Cole, Cole, and Simon (1981, p. 882), this initial phase involved descriptive analyses of approximately 1,200 NSF proposals drawn from ten NSF

For the re-review study, the NSF supplied COSPUP with 150 complete proposal files drawn from proposals that had already been submitted to, reviewed by, and decided upon within the NSF's standard grant review process. The sample was constructed to include three scientific programs—chemical dynamics, economics, and solid-state physics—with 50 proposals selected from each field, stratified to include equal numbers of funded and unfunded proposals. Importantly, the unfunded proposals were competitive submissions near the funding cutoff rather than low-quality rejections. The original NSF data consist of a single proposal-level numerical summary rating for each proposal, as recorded by NSF and used in the original funding decision; these ratings aggregate multiple referee evaluations.

To conduct the re-review, COSPUP assembled independent expert panels within each of the three scientific programs, drawing on a substantially larger pool of reviewers than was used in the original NSF evaluations. Each field-level panel consisted of approximately 10–18 experts, and proposals were assigned to two panel members, each of whom solicited multiple external reviews. As a result, each proposal in the COSPUP exercise received roughly a dozen independent numerical evaluations. Reviewers evaluated proposals independently, without interaction or deliberation, and their numerical scores were subsequently averaged to form proposal-level re-review ratings. The re-review was conducted outside the live allocation process, after funding decisions had been finalized.

Using these proposal-level summary ratings from the original NSF review and the proposal-level ratings generated in the COSPUP re-review—each reflecting multiple independent referee evaluations averaged at the proposal level—the study reports Pearson correlation coefficients computed across proposals. These correlations calculated ranged from approximately 0.60 to 0.66 across the different fields, corresponding to R^2 values of roughly 0.36–0.44. Using the data, they find that 25–30% of proposals would have moved above or below the funding line under the COSPUP ratings, illustrating how modest systematic agreement can nevertheless generate meaningfully different outcomes near a cutoff.

These results have often been interpreted within the scholarly debate as evidence that peer review is overly subject to chance. The interpretation and dissemination of these findings were

programs, along with extensive interviews with NSF program officers and reviewers. That work focused on documenting how NSF peer review operated in practice—including the relationship between reviewer ratings, applicant characteristics, and funding outcomes—and was not designed as an experimental test of reproducibility or bias. The re-review study represents a subsequent phase of this broader, multi-year research program, motivated in part by findings from the earlier descriptive analyses and by continuing interest in the reliability of expert judgment under peer review.

themselves the subject of considerable deliberation between the author team and NSF-affiliated stakeholders, reflecting the sensitivity of the results for prevailing views of peer review. To emphasize, however, Cole et al. do not claim that outcomes hinge primarily on chance. Instead, they emphasize a dual pattern: moderate consensus across independent reviewer groups (correlations of 0.60–0.66) together with substantial disagreement, particularly near the funding line. Their title—Chance and Consensus—underscores this central point.

Viewed through a Q–A–R–S perspective, the results invite a more careful parsing of which dimensions of the evaluation process are plausibly held fixed across review rounds and which are allowed to vary, and how those differences bear on the observed dispersion in proposal scores. The design holds submission characteristics (Q) fixed by re-evaluating identical proposal texts and holds stated author information (A) fixed insofar as author identities and affiliations are unchanged across reviews. However, again the passage of time between original and re-review means the submissions might not be perfectly identical relative to an advancing knowledge frontier – however, this passage of time was only a matter of months, much less than the time passed in the case of Peters and Ceci (1982) study, discussed above.

The main sources of systematic differences across studies are differences in reviewers (R) and the synthetic peer review system (S) created for study purposes. In the original NSF process, proposals were evaluated within a live funding system in which a program director solicited ratings from a small number of external referees (generally four to five) and used those inputs to make a consequential award decision. The COSPUP re-review, by contrast, was conducted post hoc, carried no funding consequences, and drew on a substantially larger and differently constituted reviewer pool. Within each field, COSPUP assembled panels of roughly 10–18 experts; proposals were routed through two panel members who each solicited multiple reviews, yielding roughly a dozen independent evaluations per proposal. As a result, the comparison bundles changes in reviewer composition, the number of

Given these systematic differences across reviewers (R) and systems of review (S), fundamental uncertainty in assessing true quality (Q), and some possible drift in this quality (Q), along with possible idiosyncratic error (ε), the reported correlations indeed provide evidence of consistent noise and uncertainty in review, but not an attribution.

Stepping back, the results might also be interpreted from the perspective of having sampled rather high quality proposals that made their way through one of the nation's most competitive submission processes and each was a competitive proposal. Given the uncertainty associated with evaluating ambitious research aiming to make a substantial contribution (Section 2.2.1), we might

regard the observed correlations as between committees as encouraging: one standard deviation increase in a proposal's score in one review exercise predicts roughly a one-third standard deviation increase in the other, a relationship that is statistically significant ($p < 0.001$). Although 25–30% of proposals might have been "reversed," this might be understood to reflect a sensitivity of rank order in a system with a fixed number of winners, rather than a sudden reversal of views of the merits of a given submission. Read in this light, the study is less damning of NSF peer review than it might have often been read, and it remains foundational in documenting both the existence of considerable chance—and consensus—in peer review.

Studying NIH Review Reliability Outside the Live Review System: Pier et al. (2018)

Pier et al. (2018) conducted another form of study to calibrate the degree of similarity and divergence in reviewers' assessments of the same grant proposal. Unlike centrally designed audits or program evaluations initiated by the National Institutes of Health (NIH), this research program was independently initiated by the principal investigator, Carnes, and grew out of her long-standing agenda on bias, evaluation, and equity in biomedical science (e.g., Carnes et al., 2012; Carnes et al., 2015).

Through professional networks, and as part of this externally initiated research program, Carnes obtained access to previously submitted NIH R01 proposals, with investigators voluntarily contributing applications that were subsequently de-identified under NIH confidentiality procedures. Although NIH has periodically sponsored or authorized internal evaluations of peer review, Pier et al.'s study was not designed or led by NIH and did not constitute an internal audit of the review process.[6] Rather, Carnes obtained permission to use de-identified proposals and to recruit experienced reviewers, while maintaining strict confidentiality over original NIH scores, written critiques, and panel outcomes.

Across this broader program of work—and in this particular study—Carnes assembled a highly multidisciplinary academic team, including Pier (learning sciences; gesture analysis and reliability analysis), Brauer (social psychology; experimental methods and statistical analysis), Filut (gender and bias in science), Kaatz (health equity and organizational analysis), Raclaw (discourse and interaction

---

[6] NIH has periodically conducted or commissioned evaluations of its peer-review system, largely in response to congressional oversight, budgetary pressures, and concerns about reliability or conservatism in funding decisions. These efforts have largely been descriptive or correlational analyses of administrative data housed within the Center for Scientific Review (CSR), as well as external oversight reports. A small number of peer-reviewed studies have analyzed NIH peer review using large-scale administrative data, including work by CSR or Office of Extramural Research staff (e.g., Eblen et al., 2016; Lindner and Nakamura, 2015; Erosheva et al., 2020) and externally led studies conducted with permissioned access to NIH data (e.g., Li, 2017).

analysis), Nathan (learning sciences; gesture analysis and embodied cognition), and Ford (linguistics; discourse and interaction analysis), among others, to study peer review across a series of related investigations.

The study reported in Proceedings of the National Academy of Sciences—"Low agreement among reviewers evaluating the same NIH grant applications" (Pier et al., 2018)—formed part of Pier's PhD dissertation and focused on quantifying inter-rater reliability in individual, pre-meeting reviewer evaluations. As Pier emphasized, the study's target estimand is operational reliability—whether the NIH review system, as actually implemented, yields consistent evaluations—not the existence of an underlying latent "true quality" of proposals. The focus on pre-meeting evaluations is deliberate: these preliminary scores determine which applications are discussed and serve as anchors for subsequent panel deliberation, making them a critical point at which noise or disagreement can shape downstream outcomes.

The study used 25 de-identified NIH R01 research proposals drawn from recent competitions within two National Cancer Institute (NCI) study sections (Oncology 1 Basic Translational and Oncology 2 Translational Clinical). Applications were voluntarily donated by investigators identified through NIH RePORTER and were selected and de-identified by the research team in consultation with an experienced former NIH Scientific Review Officer, under IRB approval. All scores analyzed in the study were generated through a constructed review process conducted by Carnes and her team. Identifying information about investigators, institutions, and budgets was removed, while the full scientific content of the proposals—including abstracts, specific aims, and detailed research plans—was preserved as originally submitted. The proposal sample included both funded applications and initial unfunded versions of applications that were ultimately funded: 16 were funded on first submission, and 9 were funded after resubmission, with the initial unfunded versions used to ensure variability in application quality.

The investigators recruited 43 reviewers with prior NIH experience, many of whom had served on NIH study sections within the preceding five years. Reviewers were informed that they were participating in a study of peer review practices and were asked to evaluate proposals as they would in a standard NIH panel. Reviewers were matched to proposals by expertise, with each assigned one or two proposals as a primary reviewer, yielding 83 evaluations across the 25 proposals (an average of 3.32 primary reviews per proposal). As Pier noted, the research team expected *some* degree of agreement ex ante, given the unusually favorable conditions for reliability: a pool of highly competitive

proposals, experienced reviewers, careful expertise matching, and deliberate efforts to replicate NIH review procedures with high fidelity.

The primary analysis relies on intraclass correlation coefficients (ICCs), which measure the proportion of score variance attributable to differences between proposals rather than differences among reviewers evaluating the same proposal. Across multiple outcome measures, the estimated ICCs are not merely statistically insignificant but are themselves equal to or near zero: for overall impact scores (ICC = 0.00; 95% CI: 0–0.14), strengths scores (ICC = 0.00; 95% CI: 0–0.15), and weaknesses scores (ICC = 0.02; 95% CI: 0–0.18). The authors conclude that "the outcome of the grant review depended more on the reviewer to whom the grant was assigned than the research proposed.

Viewed through a Q–A–R–S lens, submission characteristics Q are held constant across multiple reviewers by design. The contemporaneous review of proposals by multiple reviewers arguably maintains Q constant to a greater degree than the re-review studies, earlier. Author attributes A are intended to be neutralized through de-identification. The reviews were conducted through an equivalent review system and process S in each case. The study then examines variation associated with reviewers R.

Thus, this simpler design is instructive, as—in principle—it might achieve greater controls than those of re-evaluation studies, as above. As the authors will note, a first concern might relate external validity and comparability to the usual NIH review process. For example, the review system and process S used here is a synthetic one, essentially re-constructed in the lab. The authors should be lauded for going well beyond the usual hiring of undergraduates for lab studies; however, this point naturally opens questions of comparability. Further, the sampling of studies and reviewers is laudable for finding directly relevant participants. But, in doing so in a voluntaristic manner and with relatively small numbers of course, this also creates a possibility of deviation from usual distributions and patterns.

Of course, our interpretation of the variance decomposition in Pier et al. (2018) should also be understood as conditional on a sparse data structure, with limited reviewers per proposal and limited proposals per reviewer. Nonetheless, it is notable that the point estimates of ICC are not just statistically zero, but in fact zero. This might be relevant too to the sampling frame: the results are based on 25 proposals were ultimately funded by NIH. Thus, these are especially high quality proposals that would have overcome a series of hurdles and been deemed to be of exceedingly high quality by a panel of peers prior to entering into the Piers et al. (2018) analysis. As discussed in Section

2, evaluations might simply be inherently more compressed and difficult to rank order among acceptable, let alone highest quality submissions.

3.1.1 What Have We Learned from Re-Evaluation Studies so Far? Different Reviewers Give Different Evaluations

Re-evaluation and resubmission studies provide the earliest—and most rhetorically powerful—evidence that peer review outcomes are not uniquely determined by submission content alone. Across psychology journals (Peters & Ceci 1982), NSF grant review (Cole et al. 1981), and NIH-style proposal evaluation (Pier et al. 2018), nominally identical work receives meaningfully different scores, rankings, or decisions when assessed by different reviewers or panels. These studies establish, beyond dispute, that reviewer disagreement is a real and consequential feature of scientific evaluation.

Viewed through the structural regularities outlined in Section 2, this central finding is neither anomalous nor diagnostic of institutional failure. When true quality is uncertain, forward-looking, and multidimensional, evaluation is necessarily inferential. Reasonable experts draw on partially overlapping foundations of knowledge—shared disciplinary standards, common empirical benchmarks, and accumulated field-specific understanding—while also bringing distinct specializations, perspectives, and epistemic priors. The result is a predictable combination of overlap and divergence in judgment. Re-evaluation studies therefore document the existence of disagreement and variance across reviewers, not arbitrariness or the absence of shared evaluative foundations.Importantly, disagreement in these studies coexists with substantial agreement. Cole et al. (1981) report correlations of roughly 0.60–0.66 across independent NSF review exercises, implying meaningful shared signal alongside noise. Even in Pier et al. (2018), where inter-rater reliability is near zero at the individual-reviewer level, the proposals examined are uniformly high quality, suggesting compression of evaluative distinctions near the upper tail rather than wholesale evaluative collapse. Read correctly, these findings support the joint presence of chance *and* consensus, rather than the dominance of either.

At the same time, the designs themselves sharply limit attribution. In Peters and Ceci (1982), variation in inferred quality (Q), author cues (A), reviewer assignment (R), and evaluative context (S) move together. By contrast, Cole et al.'s NSF audit and Pier et al.'s NIH-style reliability study achieve cleaner control over submission content, but do so by constructing synthetic review systems that differ in consequential ways from live decision environments.

Another lesson concerns design efficiency. Re-evaluation and resubmission studies are institutionally costly, slow, and difficult to scale because they treat panels or entire review rounds as

the unit of analysis. Later work—including Cole et al.'s NSF audit and, more starkly, Pier et al.'s NIH-style reliability study—demonstrates that much of the relevant variation can be observed without re-running peer review at all. Simply having multiple qualified reviewers evaluate the same submission within a single review cycle is sufficient to generate re-evaluation in the relevant sense. This shift—from reconstituting full panels or review rounds to examining reviewer–submission level disagreement

3.2 Audit and Sting Studies: Stress-Testing with Varying Submission Characteristics (Q)

A second class of studies moves beyond passive observation and intervenes directly on what the review system sees. Audit and sting designs deliberately manipulate features of a submission—altering results, inserting errors, fabricating author identities, or re-submitting known work—to test whether peer review behaves as it should under controlled stress. In Q–A–R–S terms, these studies typically intervene on perceived quality (Q) or author attributes (A) while holding the broader review system (S) and reviewer pool (R) as constant as institutional realities allow. Rather than estimating average treatment effects, audit studies are diagnostic by construction: they embed normative benchmarks ("this error should be detected," "this duplicate should be rejected") and treat deviations from those benchmarks as evidence of failure modes in the review process. As a group, these studies reveal that reviewers often miss even serious flaws, that editorial screening can break down under certain incentive structures, and that identity cues can influence scrutiny. At the same time, because manipulations are usually narrow, outcomes are coarse, and reviewer heterogeneity is uncontrolled, audit studies are better read as structured probes of system vulnerability than as clean causal estimates. This section reviews the major audit and sting designs, emphasizing what they reveal about peer review's limits—and why they cannot, on their own, explain when or why those limits arise.

An Earliest Example of Randomized Referee Assignment: Mahoney (1977)

Mahoney (1977) conducted one of the earliest controlled attempts to probe bias in peer review by randomly assigning reviewers to evaluate versions of the *same* research article with systematically altered results. The introduction, theory, and methods were held constant across conditions. Only the empirical results (and, in two cases, the accompanying discussion) were rewritten to either support or contradict the stated hypothesis—that extrinsic rewards undermine intrinsic motivation. Thus the manipulation targeted a narrow component of perceived quality Q, while keeping author identity (A) fully blinded.

Mahoney assembled 75 behavioral-science referees (67 completed reviews) and informed referees that the manuscript was being evaluated "as though it had been submitted for publication"

(pp. 165–166). Unknown to them, Mahoney had created five versions of the same paper. All versions shared the same introduction and methods; only the results and, in some cases, the discussion were altered. Some reviewers saw results that supported the hypothesis, others saw results that contradicted it, others saw ambiguous results accompanied by either an affirming or skeptical discussion, and one group saw a version with no results at all. Each reviewer was randomly assigned one version and asked to evaluate it as they normally would.

Across rating dimensions—topic relevance, methodological adequacy, data presentation, discussion, and overall scientific contribution—reviewers assigned higher scores when the reported results confirmed the stated hypothesis. Mahoney interpreted this as evidence of confirmatory bias. The study also contained a deliberately inserted internal contradiction. This inconsistency was detected by only ~31% for positive-results manuscripts and ~71.4% for negative-results manuscripts.

Despite its conceptual clarity, the study's design introduces inferential limitations. The 67 completed reviews were divided across five experimental conditions, yielding small and uneven cell sizes that limit precision and complicate interpretation. In addition, Mahoney reports very low within-cell concordance, implying substantial heterogeneity in reviewer types within each treatment. Consistent with this interpretation, reported inter-referee agreement is low (intra-class correlation measures reported in Tables on pp. 170–171), indicating that reviewers evaluating the same manuscript version often disagreed substantially in their assessments—reinforcing reviewer heterogeneity R as a central limitation of inference in this design. With such small cell sizes, even random imbalances in reviewer expertise, orientation, or leniency could drive the observed mean differences. Because no balance tables were reported, systematic compositional differences across treatment groups cannot be ruled out.

A second interpretive complication lies in the broader theoretical landscape at the time. The debate over extrinsic rewards and intrinsic motivation was heated and far from unanimous. As Mahoney himself acknowledges, some reviewers would have found disconfirming evidence more theoretically appealing. Under such conditions, both possible patterns—higher ratings for confirmatory or disconfirmatory results—could be consistent with a form of confirmatory bias, depending on the reviewer's priors. With strong heterogeneity in reviewer beliefs and very small treatment cells, it is difficult to determine whether the observed differences reflect bias per se or the luck of assignment.

For these reasons, I view Mahoney's study as an important conceptual precursor rather than a precise causal estimate. It offers the first demonstration that randomized manipulation of results can

meaningfully alter referee judgments, and it innovated by imposing experimental control on what had previously been anecdotal observations. Yet, under the Q–A–R–S framework, key elements remain uncontrolled: reviewer attributes R, unobserved dimensions of perceived quality Q, and implicit system features S such as expectations about the publication outlet. The study therefore reveals the *possibility* of confirmatory bias, while falling short of isolating its magnitude.

Injecting Errors in a Fake Paper to Report How Many are Detected: Baxt et al. (1998)

In the study entitled *"Who Reviews the Reviewers?,"* Baxt, Waeckerle, Berlin, and Callaham (1998) conducted a kind of intervention to assess whether referees could detect flaws, at the Annals of Emergency Medicine.[7] The study was conducted and published in the same journal, where Baxt was a senior associate editor, Waeckerle was editor in chief, and Callaham was a senior associate editor. The idea was to test whether reviewers could identify deliberately seeded errors in a manuscript that was made-up by the authors. The fake study described a fictitious double-blind, placebo-controlled trial claiming that intravenous propranolol reduced pain from acute migraine headaches. In their own words, they inserted "10 major and 13 minor errors."[8] The authors goal was to fabricate a study appeared to be superficially sound but has such errors, with the view that a careful reviewer would detect these errors.

Two months before the fictitious manuscript was submitted (summer 1994), a letter was sent to all reviewers of the journal, stating the journal would evaluate their performance in the near future and that they could choose to opt out, without indicating what the assessment would be or when it would take place.[9] The manuscript was sent to 262 available reviewers (out of 309 on the roster); 203 reviews were returned (78%). The authors document roughly two-thirds of the major errors were not

---

[7] It's the official journal of the American College of Emergency Physicians (ACEP) and is considered one of the top emergency medicine journals globally. According to Journal Citation Reports, it had a 2021 impact factor of 6.762, ranking it first out of 32 journals in the "Emergency Medicine" category.

[8] This included such things as the absence of subject inclusion criteria, randomization based on "flipping a coin at midnight to determine which therapy would be used the next day," use of an unvalidated pain scale, and reported statistical results that did not support conclusions. Other problems included duplicated tables, missing figures, misspelled drug names, outdated references, and a few fabricated citations.

[9] In 1994, when Baxt et al. ran this experiment, editorial peer-review studies generally weren't treated as "human-subjects research" under U.S. federal policy (the Common Rule). Reviewers were acting in their professional capacity, not as patients or research subjects, and the activity was framed as internal quality improvement by the journal rather than generalizable behavioral research. The authors did, however, obtain IRB approval from the lead author's institution, as they note explicitly in the paper ("IRB approval was obtained for the study from the institution of the lead author."). So, at the time, it met prevailing standards. A key issue from today's perspective might be some arguable degree of deception or lack of informed consent.

mentioned in the referee reports; which they interpret as, "Overall, reviewers failed to detect 68% of the major errors and 75% of the minor errors intentionally inserted into the manuscript" (p. 315).

What is perhaps most notable about this exercise is the idea that rather than model overall evaluations or scores, it might be possible to measure a specific component of an evaluation—here error detection. Relatedly, this point raises a question of exactly what is expected of human expert reviewers, and what part of this is error detection and else might be expected of a reviewer. Many of the designed errors in this study might, for example, today, be detected by machine intelligence. This is particularly so given the detection of objective errors well within the existing frontier of knowledge may be relatively straightforward, or at least more straightforward than say assessing a novel contribution (see Section 2.3.1). It is attractive, nonetheless, that the Baxt et al. design provided a truly observable *component* of quality (objective errors) Q that could be used to score reviewers against. Of course, in measuring this dependent variable, they are not in fact measuring error detection. Rather, they measured whether an error was explicitly reported in the referee report, so we might regard this as perhaps a lower bound of what they detected.

An alternative view of the data is that of the 203 returned reviews, only 15 (7 percent) recommended acceptance. For example in a randomly drawn 3 person review team (say, two reviewers and one editor), with the odds of a majority wishing to accept the paper, the probability of a majority choosing to accept is less than 1.5 percent, even in this wide sample of reviewers. We might expect the editor and more carefully chosen reviewers to be more critical and discerning (see discussion in Section 2.2.3 and Property 3), and less likely to miss these most basic objective errors in their specialty areas (see discussion in Section 2.3.1 and Property 1), leading to an even lower probability. Any layer of intelligent aggregation and synthesis of review inputs or mechanical review for errors should lead the chances of publication to drop further. Therefore, for a fake paper with 23 errors, it seems fair to expect the system will regularly reject. Perhaps the most feasible and easiest manipulation could have been to vary the number of errors introduced, or any errors at all, to determine how this corresponded with rejection rates. This could have been a more conclusive way to detect the effect of errors than relying on reporting of individual errors in the report.

This descriptive audit was built without controlled comparisons or experimental manipulations or assignments of any kind. Author identity and attributes A were blinded, so as to focus the reviewers here on the manuscript (Section XX). The quality was kept constant and even somewhat well-known. The system of peer review S was held constant.

The variation in reviewers attributes *R* was broader than would have otherwise been the case, and any effects of this extraordinary variation was not studied or documented. One might expect that the inclusion of over 200 reviewers to consider a paper on migraines strained usual empirically relevant degrees of variation in relevant knowledge, which could have also reduced error detection or could have offered useful variation to study effects of (ir)relevance reviewer knowledge and error detection (Section 2.3.3 and Property 3). This variation might still be used in a post hoc analysis of the data to detect the role of relevant expertise in mentioning errors in the report, or at least to assess probability of rejection.

This study helpfully shone the light on the idea that more tangible and intermediate variables (here, error detection) could create opportunities beyond just studying scores and decisions. However, the opportunity could have supported a more discerning experimental research design, such as randomly assigning reviewers different error bundles (manipulating *Q*); or say varying features of the review system to treat propensity to detect errors such as checklists, incentives, prompts, time guidance (manipulating *S*). It is not possible to experimentally vary the attributes of reviewers R (Section XX), but it would have been informative to perhaps map how reviewer types map to outcomes and responsiveness to treatments, just as examples. Exposing author identity and attributes A is the sole dimension that would have violated journal policy and therefore not a feasible dimension in which experimentation could have been conducted in this opportunity.

Injecting Errors at a Medical Journal Submission: Godlee et al. (1998)[10]

In "Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports: A Randomized Controlled Trial" (JAMA, 1998), Godlee, Gale, and Martyn conducted a field experiment with the *BMJ* to test whether modest changes to the review system **S**—author blinding and requiring signed reviews—influence error detection. Instead of a fabricated manuscript, they used a recently accepted *BMJ* article on cognitive impairment and mortality in older adults, with the original authors' consent. The editors inserted eight deliberate errors spanning inconsistent statistical results, incomplete description of randomization, and misinterpretation of findings, and then coded referee reports according to whether each error was mentioned.

The experiment implemented a 2×2 design crossing (i) blinded versus unblinded author identity and (ii) a requirement to sign the review versus the usual anonymous report, plus a fifth

---

[10] Godlee et al. (1998) straddles 3.2 and 3.4.

"business-as-usual" group in which reviewers were not told they were part of a study. From *BMJ*'s pool of 834 active referees, the editors selected approximately 420 for random assignment to the four experimental cells, oversampling the signature conditions in anticipation of lower participation, and added roughly 60 reviewers to the fifth, uninformed cell. In total, about 221 reviewers returned reports. Each invitation stated the treatment condition (for example, whether author names would be shown and whether the reviewer would be expected to sign) and, in four of the five cells, indicated that the review was "part of a study into ways of improving peer review."

On the authors' primary outcome—number of seeded errors detected—there was no statistically significant difference across the five conditions: reviewers in each cell identified, on average, between 1.7 and 2.1 of the eight introduced errors, and the authors cautiously concluded that neither blinding nor signing "had any effect on rate of detection of errors." They also reported no significant differences across cells in willingness to review. However, they did note suggestive differences in recommendation behavior: blinded reviewers were somewhat less likely to recommend rejection than unblinded reviewers, and the pooled comparison of all blinded versus all unblinded conditions yielded a borderline effect (odds ratio 0.5; 95% CI, 0.3–1.0) on recommendations to reject.

Viewed through the Q–A–R–S lens, the design tightly controls Q by using a single paper with eight known errors and randomizes S by varying blinding and signing policies. Author identity A is fixed in the unblinded conditions and removed in the blinded ones; reviewer attributes R are distributed across treatments in expectation by random assignment. In principle, if evaluation depends only on (Q), we should see the same mean error counts and recommendation distributions across all cells, with only within-cell variance reflecting idiosyncratic noise. Empirically, the mean error counts do not differ appreciably across treatments—consistent with the authors' "no effect" conclusion— but the borderline differences in rejection rates and participation hint that S may affect other dimensions of behavior not captured by the narrow error-count metric.

At the same time, the study illustrates how hard it is to prove a zero effect. With only a few dozen reviewers per cell, sparse reporting of variance measures, and a dependent variable that undercounts true error detection (only errors explicitly mentioned in the report are scored), the absence of statistically significant differences in error counts could reflect genuine robustness of reviewers to these policies—or simply limited power and measurement. The experiment is thus best read as an early, relatively clean randomized trial showing that, for this particular paper at *BMJ*, blinding and signing did not meaningfully change mean error detection, while leaving open the

possibility that these same interventions could affect other outcomes (tone, effort, selection into reviewing) or behave differently for other manuscripts and contexts.

Other Audit and Sting Studies: Large-Scale Stress Tests, with Limited Interpretation

Bohannon's "Who's Afraid of Peer Review?" (Science, 2013) emerged from investigative science journalism rather than from academic social-scientific research. Bohannon fabricated a biomedical manuscript reporting an implausible anticancer effect of a lichen-derived compound and intentionally seeded it with elementary methodological and statistical flaws—such as missing controls, implausible dose–response claims, and internal inconsistencies—that would ordinarily trigger rejection at first-pass review. Variants of this manuscript were submitted to 304 open-access journals; among those completing the review process, 157 accepted the paper. Bohannon interprets results as indication "little or no scrutiny at many open-access journals."

Cobey et al. (2020) resubmitted (with permission) a previously published *Nature* article to a large sample of biomedical journals spanning presumed predatory outlets, legitimate open-access venues, and subscription journals. Of the 308 journals that responded within the study window, four accepted the duplicated manuscript, 94.5% rejected it, and 4.2% requested revision. As the authors interpretation is that "our findings suggest that all three types of journals may not have adequate safeguards in place to recognize and act on plagiarism or duplicate submissions." I should be emphasized nonetheless that the vast majority chose to reject a paper that was publishable, but for its having been previously published.

3.2.1 What Have We Learned from Audit Studies So Far? Existence of Vulnerabilities

From a Q–A–R–S perspective, the audit literature establishes **that failures in evaluation can occur**, but it does not identify **when, why, or through which channel**. In most designs, multiple components move together: flaw salience and severity (Q), reviewer expertise and attention (R), editorial screening and workflow (S), and sometimes author cues (A). As a result, observed acceptance or rejection outcomes are composites of several mechanisms. The natural next step—and the main opportunity—is to turn these stress tests into designs with greater inferential leverage: experiments that vary specific components of (S) (and where feasible, information in (A)) while holding other margins fixed, or that instrument reviewer expertise (R) in a way that makes heterogeneous detection interpretable.

On the one hand, read against the structural regularities in Section 2 we might not be surprised there are often errors in assessment. However, many of these errors would sometimes appear to be

regarding relatively basic items—such as repeated or erred manuscripts. These problems are not the sort of problems that might be expected of high quality contribution advancing the frontier at venues with a high stake and interest in seeking best possible research and maintaining a reputation for doing so. These results surely should raise concerns about how feature of a peer review system (S), including its governance and business model may shape the process.

Finally, it deserve emphasis that a distinguishing feature of these audit and re-evaluation studies is their disproportionate visibility within the literature. Many of the most rhetorically striking contributions—such as Peters and Ceci (1982), Mahoney (1977), Godlee et al. (1998), and Bohannon (2013)—are also among the most heavily cited papers in the entire peer-review evidence base, despite their limited capacity for causal attribution.

3.3   Manipulations of Reviewer Composition and Expertise (R)

Peer review depends not only on what is being evaluated, but also on who is doing the evaluating. A focused set of randomized experiments directly manipulates reviewer composition—altering the expertise, methodological orientation, or intellectual proximity of reviewers assigned to the same submission—while holding the submission content fixed. These designs isolate the reviewer component (R) of the Q–A–R–S framework more cleanly than earlier audit or resubmission studies, which allowed reviewer variation only incidentally. Some interventions add specialized reviewers, such as statisticians or information specialists, to augment existing review teams; others systematically vary which reviewers are assigned, creating exogenous differences in intellectual distance between evaluator and work. Together, these studies provide the strongest causal evidence to date that who evaluates a submission meaningfully shapes how it is judged. At the same time, they reveal important design trade-offs: additive reviewer interventions often bundle expertise with reviewer count, while relational designs require large reviewer pools and careful measurement. This section reviews the core reviewer-composition experiments and shows how they clarify the role of expertise, proximity, and perspective in shaping evaluative outcomes.

Does Adding a Statistical Reviewer Improve Manuscripts? Early Evidence from a Randomized Trial: Arnau et al. (2003)

Arnau et al. (2003) report one of the earliest randomized experiments testing whether adding a dedicated statistical reviewer to a peer-review team improves the quality of submitted manuscripts. The trial was conducted at *Medicina Clínica*, a Spanish weekly journal of internal medicine, and focused on original research articles that had passed initial editorial screening. The intervention directly targets

reviewer composition (R) by introducing methodological expertise into the review process while holding the submission itself fixed at the point of randomization.

A total of 43 manuscripts entering peer review during the study period were randomly assigned either to standard clinical peer review or to clinical review plus an additional statistical reviewer. The study does not describe the size or composition of the statistical-reviewer pool, nor how statistical reviewers were selected or assigned beyond their inclusion in the treatment arm. As implemented, the intervention is additive: manuscripts in the treatment group received more total reviewer input as well as more specialized expertise.

Manuscript quality was assessed using a 36-item reporting-quality instrument applied to the initial submission and again to the revised manuscript after peer review. Two evaluators, blinded to treatment assignment, independently scored each manuscript. The outcome of interest is the change in the aggregated quality score between the initial and revised versions, collapsing across dimensions of reporting, design, analysis, and presentation. This aggregation bundles multiple aspects of quality into a single scalar measure, limiting insight into which dimensions—if any—are most responsive to statistical review.

Under the original intention-to-treat assignment, manuscripts in the statistical-review condition exhibited slightly greater improvement than those under standard review, with an estimated difference of approximately +1.35 points on the 36-item scale. This difference was not statistically distinguishable from zero. The article does not report group-specific mean levels or changes in sufficient detail to support more granular interpretation of effect size or variability.

During analysis, the authors discovered noncompliance with the randomized protocol: four manuscripts assigned to the control group had in fact received statistical review. When outcomes were reclassified according to the review actually received, the estimated improvement associated with statistical review increased and became statistically significant, with a reported confidence interval of approximately +0.3 to +3.7 points. While suggestive, this "as-treated" analysis sacrifices the clean causal interpretation afforded by randomization and is vulnerable to selection bias.

Several design features limit interpretation of the findings. First, the unit of observation is the manuscript, yielding only 43 observations and thus limited statistical power. Second, because the intervention adds a reviewer rather than holding reviewer count constant, the study cannot distinguish the effect of statistical expertise from the mechanical effect of increased reviewer attention. Third, the reliance on a single aggregated quality score obscures whether statistical review improved methodologically relevant dimensions in particular. Only a subset of the 36 items directly concern

statistical design or analysis, raising the possibility that meaningful improvements on those dimensions are diluted in the aggregate outcome.

Nevertheless, the study is important as a proof of concept. Unlike audit or resubmission studies that manipulate manuscript content or presentation, Arnau et al. embed a randomized intervention directly within a functioning editorial workflow. In Q–A–R–S terms, underlying manuscript quality (Q) is fixed at randomization; author attributes (A) are not manipulated; system features (S) remain unchanged apart from the added reviewer; and reviewer composition (R) is experimentally varied. The modest estimated effects therefore speak less to the irrelevance of expertise than to the difficulty of isolating its impact with limited power, coarse outcome measures, and bundled interventions.

A More Elaborate Follow-On Trial: Cobo et al. (2007)

Cobo et al. (2007) report a larger and more elaborate randomized trial conducted within the same journal, *Medicina Clínica*, building directly on the Arnau et al. experiment. While retaining the core question of whether statistical review improves manuscript quality, the study expands the design to evaluate the separate and joint effects of adding a statistical reviewer and providing structured reporting checklists. The trial therefore offers both increased statistical power and a more explicit test of complementary interventions.

The sampling frame again consists of original research manuscripts that passed initial editorial triage and entered peer review. The study enrolled 129 manuscripts over an approximately eleven-month period (May 2004–March 2005), reflecting increased submission volume relative to the earlier trial. Fourteen manuscripts did not submit a revision, leaving 115 manuscripts available for analysis.

The experiment implemented a 2×2 factorial design embedded in the journal's routine review process. Manuscripts were randomly assigned to one of four conditions: (i) standard clinical peer review; (ii) clinical review plus a statistical reviewer; (iii) clinical review plus a suggested reporting checklist; or (iv) both a statistical reviewer and a checklist. Statistical reviewers were individuals with training in epidemiology or biostatistics who used structured forms to assess study design, statistical methods, and reporting. As in Arnau et al., the intervention adds reviewer input rather than holding constant the number of reviewers, bundling reviewer count and reviewer type.

Manuscript quality was again measured using the same 36-item reporting-quality scale, scored for the initial submission and the revised manuscript. Unlike the earlier study, Cobo et al. report descriptive statistics for baseline quality (mean initial score 84.5; SD 19.1) and estimate treatment

effects both on the aggregated score and, secondarily, on individual reporting items. The primary outcome remains the change in the summed 36-item score across review.

Under standard clinical review, manuscripts improved by approximately +4.5 points on the scale. Providing reviewers with a checklist produced a nearly identical improvement (+4.7 points), with no statistically significant difference between the two. By contrast, adding a statistical reviewer increased quality by an average of +9.1 points. When averaged across checklist conditions, the estimated effect of statistical review is +5.5 points relative to no statistical review (95% CI: 4.3 to 6.7), and is statistically significant. The interaction between statistical review and checklist provision is positive but small (+1.1 points) and not statistically significant.

Item-level analyses reveal statistically significant improvements on 10 of the 36 reporting items, including several directly related to methodology and statistical reporting. However, the summed point estimates for the method-related items account for only a fraction of the overall +9.1-point improvement, underscoring how aggregation across heterogeneous dimensions complicates substantive interpretation of effect magnitude.

The authors conclude that adding a statistical reviewer has a positive but modest effect on reporting quality, describing the effect as "significant and positive" yet "very small relative to the scale range." Several interpretive challenges remain. With only 115 usable manuscripts spread across four experimental cells, the study is powered to detect only relatively large effects and is ill-suited to explore heterogeneity—for example, whether weaker manuscripts benefit disproportionately from statistical review or whether effects differ by study design.

As in Arnau et al., the design bundles two mechanisms: increased reviewer count and increased methodological expertise. The study therefore cannot distinguish whether the observed gains arise because statistical reviewers identify errors missed by clinical reviewers, because authors respond differently to methodological critique, or because editors weight recommendations differently when statistical concerns are raised. Similarly, the null effect of checklist provision is difficult to interpret absent a behavioral model: reviewers may have ignored the checklist, found it redundant, or lacked incentives to enforce its guidance.

The institutional context further conditions interpretation. Only 16 of the 115 manuscripts were rejected after peer review, indicating a highly permissive acceptance environment in which reviewer input primarily guides revision rather than screens submissions. Whether adding methodological expertise would have larger, smaller, or qualitatively different effects in higher-

rejection journals, in settings with stronger pre-review screening, or where editorial decisions hinge more directly on methodological critique remains an open question.

Viewed through the Q–A–R–S framework, Cobo et al. provide a well-executed randomized intervention on reviewer composition (R) within a stable editorial system (S). Manuscript quality (Q) is proxied by a detailed but highly aggregated reporting scale; author attributes (A) are not manipulated. The study delivers stronger causal leverage than its predecessor but remains constrained by aggregation of outcomes, limited power for interaction effects, and the inability to isolate reviewer expertise from reviewer quantity. Together with Arnau et al., it establishes that reviewer composition can causally shape manuscript development, while also illustrating the methodological demands required to move from detecting effects to explaining mechanisms.

Adding Librarians as Methodological Peer Reviewers: Rethlefsen et al. (2025)

Rethlefsen et al. (2025) examine whether adding a librarian or information specialist (LIS) to the peer-review team improves the quality of evidence-synthesis manuscripts in biomedicine. Their focus is on review articles—systematic reviews, meta-analyses, scoping reviews, and related formats—for which the construction, transparency, and reproducibility of literature-search strategies are central components of methodological quality. The study asks whether introducing specialized expertise in search methodology at the review stage improves reporting quality or reduces risk of bias, relative to standard peer review alone.

The experiment was embedded within the active editorial workflows of three BMJ journals—*The BMJ*, *BMJ Open*, and *BMJ Medicine*—all of which regularly handle high volumes of evidence-synthesis research. The author team includes external methodologists and information scientists, as well as a long-standing internal BMJ researcher who focused on researching the editorial process over multiple years of employment at the journal, reflecting the journal group's multi-decade program of conducting randomized evaluations of editorial practices.[11]

---

[11] BMJ has maintained an editorial-research function in varying forms and levels of formality since the late 1980s under editor-in-chief Richard Smith, who wrote a series of editorials arguing that journals should "study themselves" (e.g., Smith 1990; Smith 1997; Smith 1999). Early published studies associated with this ethos included the 1990 randomized trial of author blinding by McNutt et al. (JAMA 1990) and Godlee, Gale, and Martyn's 1998 "seeded-errors" experiment (JAMA 1998). By the early 2000s, BMJ's research transitioned into more sustained experimental investigation using real submissions and reviewers. Around this time, BMJ institutionalized this capacity by hiring PhD-trained researcher Sara Schroter, who for more than 15 years designed or co-designed trials and audits on reviewer training, incentives, reporting-guideline adherence, editorial decision-making, reviewer disagreement, open peer review, statistical-review processes, and reviewer burden. The program became less centralized after the mid-2010s; however, BMJ remained active through collaborations with Ottawa, Oxford, Maastricht, and other meta-research groups. The Rethlefsen et al. (2025) study,

Eligible manuscripts were newly submitted evidence-synthesis papers that passed initial editorial triage and were sent out for peer review between January and September 2023 for *The BMJ* and *BMJ Open*, and between May and September 2023 for *BMJ Medicine*. Across these journals and time windows, 2,670 submissions entered peer review; 400 met the study's inclusion criteria and were randomized 1:1 to either standard peer review or standard peer review plus an invited LIS reviewer. As in earlier statistical-reviewer trials, the intervention was additive: it increased reviewer input rather than holding constant the number of reviews and substituting reviewer type.

The study documents that LIS reviewers were drawn from the "Librarian Peer Reviewer Database," an international volunteer registry of unpaid volunteer librarians and information specialists with specific expertise in literature-search methodology. LIS reviewers were invited "using the same standard email as all other reviewers" and drawn "sequentially from the Librarian Peer Reviewer Database list.[12]

For manuscripts randomized to the intervention arm, editors contacted LIS reviewers sequentially following the registry order until one accepted the invitation or until ten invitations had been issued. This procedure produced substantial heterogeneity in reviewer assignment: some LIS reviewers handled multiple manuscripts while many reviewed only one; several manuscripts required 8–11 invitations to secure an LIS reviewer, whereas others obtained acceptance after a single invitation. These dynamics reveal the scarcity of specialized search-method expertise and the operational difficulty of scaling such an intervention in a high-volume editorial environment.

Of the 400 randomized manuscripts, 166 (76 in the intervention arm and 90 in the control arm) submitted a first revision by the prespecified follow-up date and were included in the primary analysis. Among manuscripts assigned to the LIS condition, 81.6% ultimately received an LIS review; the remainder did not, despite randomization, due to non-response within the invitation limits. Editors

---

conducted through a BMJ–Maastricht PhD collaboration, sits squarely within this continuing tradition. BMJ's long-running editorial-research function has produced numerous peer-review studies. Experimental studies include: Godlee et al. 1998; van Rooyen et al. 1999; van Rooyen et al. 2010; Schroter et al. 2004; Schroter et al. 2005; Callaham & Tercier 2007; Schroter et al. 2008; Hopewell et al. 2010; Schroter et al. 2010; Schroter et al. 2012; Wager et al. 2009–2011; Goldacre et al. 2014; Rethlefsen et al. 2025. The experimental studies were outnumbered by non-experimental studies, internal editorial reports, and working papers.

[12] It is worth noting that it would have been costless to randomize the order of LIS specialists when sending invitations, rather than proceeding top to bottom, to avoid possible correlation in reviewer attributes and the use of LIS specialists over time. With observations pooled over time in the analysis, the correlation of these and other possible variables over time might not have affected coefficient estimates. It is also plausible that recruiting this type of reviewer through sequential invitations systematically added time to the review cycle in the treatment group.

were not blinded to reviewer type and retained full discretion over editorial decisions, as under standard practice.

Two independent, blinded assessors scored the first revised manuscripts using a standardized eight-item instrument designed to evaluate the clarity, completeness, and potential bias of literature-search reporting. Importantly, the study did not score initial submissions, only revised manuscripts. As a result, the design cannot measure within-manuscript improvements over review; it instead estimates differences in post-review outcomes across treatment arms, conditional on survival to revision.

The intervention produced an unanticipated but—I believe—substantively important effect: manuscripts in the LIS treatment arm were 13.8 percentage points more likely to be rejected at the first post-review decision (95% CI: 3.9 to 23.8). On the study's primary reporting-quality outcome— full compliance with all required reporting criteria—the estimated difference between groups was small and statistically insignificant (6.6% in the intervention arm versus 2.2% in the control arm; difference 4.4 percentage points; SE ≈ 3.3). Individual reporting items likewise showed no statistically significant treatment effects. The authors interpret these results as evidence that adding an LIS reviewer did not measurably improve reporting quality or reduce risk of bias under the conditions of the trial.

I disagree with this interpretation. While the intervention did not generate statistically detectable gains on the narrow reporting metrics used as outcomes, it produced a clear and consequential shift in actual editorial decisions. It is also the case that interpreting the reporting-quality estimates is complicated by post-randomization selection. Because manuscripts in the treatment arm were more likely to be rejected at the first decision, the set of revised manuscripts scored for quality is endogenously selected in a treatment-dependent way.[13]

Viewed through the Q–A–R–S framework, the experiment constitutes a relatively clean intervention on reviewer composition (R): a specialized reviewer with distinct domain expertise is added while underlying manuscript quality (Q) and author attributes (A) are fixed at randomization. However, system-level features (S) strongly condition how reviewer input translates into outcomes.

---

[13] If LIS reviewers disproportionately surfaced weaknesses in search methodology, leading editors to reject weaker submissions outright, then the remaining manuscripts in the intervention arm would be positively selected on precisely the dimensions the reporting instrument seeks to measure. This selection process could mechanically attenuate observed differences in reporting quality, even if the intervention improved evaluative rigor. Scoring initial submissions— an analysis the authors note was not conducted but remains feasible—would have allowed direct estimation of changes across review and clearer diagnosis of these selection effects.

Editors were not required to privilege LIS input; authors were not required to implement LIS suggestions; and reviewer uptake was imperfect due to sequential invitations and non-response. Under such institutional constraints, it is unsurprising that added expertise affected editorial triage more strongly than it affected revised manuscript presentation.

Intellectual Distance, Novelty and Reviewer Evaluation: Boudreau et al. (2016)

Boudreau et al. (2016) also vary reviewer composition, but in a very different way. Rather than using "add-a-reviewer" interventions (as in Arnau, Cobo, and Rethlefsen), they ask how incremental variation in intellectual distance between a reviewer and a submission shapes evaluation—directly varying R in relation to a fixed submission. The experiment was embedded in a large internal seed-grant competition on endocrine-related disease at a major research university, spanning multiple schools and departments. The coauthor team brought complementary capabilities to execute the large-scale design: Guinan contributed experience as a medical researcher and expertise in translational science, along with deep embeddedness in medical research administration that secured institutional cooperation; Boudreau brought training in the economics of knowledge and decision analysis, experimental design, and statistical analysis; Lakhani drew on expertise in field-experimental design and innovation management; and Riedl's information-systems background supported the extraction and construction of high-dimensional, text-based measures from large-scale publication and proposal datasets.

The research context is the *first-stage evaluation* of early-stage scientific ideas rather than full, multi-year grant proposals. The competition solicited short research proposals intended to articulate new hypotheses, research directions, and exploratory approaches related to endocrine-related disease, prior to the development of detailed budgets, extensive preliminary data, or multi-investigator teams. By design, proposals were standardized in format, single-authored, and focused on conceptual contribution and potential impact, making this stage of review particularly sensitive to how evaluators interpret ideas under uncertainty. Reviewers were asked to score each proposal on a common numerical scale based on its anticipated impact on disease understanding, treatment, or research progress, closely mirroring the evaluative task faced by scientific gatekeepers when allocating scarce exploratory funding.

As discussed in Section 2, most real-world peer-review systems involve substantial endogenous selection in reviewer assignment, often resulting in relatively narrow disciplinary matching between evaluators and submissions. In contrast, this experiment deliberately recruited a large pool of

qualified scholars and randomized assignments so as to generate incrementally greater heterogeneity in evaluators' intellectual backgrounds, while still restricting the pool to researchers with sufficient relevant training to serve as credible reviewers. This design choice expands the support of intellectual distance relative to conventional review settings, allowing the study to observe how evaluations change as reviewer–proposal distance increases along a continuous margin rather than across a small set of tightly matched specialties.

Instead of having review committees assign the "best-matched" reviewers (minimum intellectual distance), reviewers were randomly drawn from a broad pool of relevant medical researchers. This created deliberate, exogenous variation in intellectual distance while avoiding any systematic assignment based on unobserved characteristics. The pool was large and diverse enough to generate wide distance variation but restricted to researchers with at least some relevant training so that all reviewers remained credible evaluators.

Intellectual distance was coded as the cosine distance between Medical Subject Headings (MeSH terms)—keywords assigned by a professionally trained librarian—characterizing each proposal and the MeSH-term vector representing each reviewer's publication history. Although the inherent true quality of any one proposal cannot be observed, the design circumvents this limitation through multiple random assignments: each proposal receives many reviewers, and each reviewer evaluates many proposals. This enables the use of both reviewer and submission fixed effects, allowing the analysis to identify whether an evaluator scored a particular submission systematically differently *because of intellectual distance*.

The study finds a coefficient of 0.86 (s.e. = 0.33) in the fixed-effects model, implying nearly a one-point difference between the closest and most distant evaluators—an effect that explains a substantial share of the within-proposal variation in scores (s.d. = 1.7). Simulations show that relying only on the closest experts would shift proposal rankings by more than 30 positions, illustrating the magnitude of the effect. Variance does not increase at low distance, suggesting a homogeneous directional effect and offering no support for private-interest mechanisms. Instead, the authors argue that closer experts perceive more informational cues and detect more potential flaws, leading them to score more critically. Intellectual proximity thus systematically depresses evaluations, even though closer experts may also be more discerning. The finding implies that simple score aggregation is inherently problematic when reviewers differ systematically in intellectual distance from what they evaluate.

The study also tests whether these effects differ for more novel proposals "beyond the frontier." Novelty is coded as the share of MeSH-term pairs in a submission that had never appeared in prior biomedical literature. There is no interaction between novelty and distance—consistent with expertise primarily sharpening recognition of established knowledge rather than altering responses to novelty. However, the direct relationship between novelty and scores follows an inverted-U pattern: scores rise with modest novelty but fall sharply for the most novel proposals. Although the authors emphasize that novelty cannot be causally identified while holding quality constant, the relationship is highly robust to saturated controls and diagnostic checks.

Viewed through the Q–A–R–S lens, the study is unusually strong. Q (Quality) cannot be observed, but the design effectively handles this through multi–random assignment and evaluator/proposal fixed effects, enabling a clean causal estimate of the effect of distance. A (Author attributes) are held constant and blinded. R (Reviewer attributes)—the focus—are experimentally varied via randomized assignment and measured continuously using MeSH-based distance. S (System attributes) are tightly controlled through standardized proposals and a triple-blind process. As the authors note, novelty cannot be causally identified because it is inseparable from proposal content, though its relationship is still robustly estimated. Overall, this study provides one of the clearest causal demonstrations of how reviewer expertise shapes evaluation outcomes.

A further design distinction is the study's use of the **reviewer–submission pair as the unit of analysis**, rather than reviewer teams or committees, as in several earlier studies. This choice substantially increases statistical power by exploiting repeated observations on both reviewers and submissions, while also avoiding the confounding problem of whether observed effects reflect differences in reviewer knowledge per se or the mechanical addition of a team member. At the same time, the resulting estimates pertain to individual evaluative judgments, not to collective outcomes that emerge from aggregation, deliberation, or synthesis across reviewers. This is not a limitation of the design so much as a conceptual distinction: the study isolates how expertise shapes *individual inference under uncertainty*, leaving open how such judgments are subsequently combined—or potentially overridden—within committees or editorial processes. Notably, the aggregation and synthesis of evaluations remain relatively underexplored in the peer-review literature, despite their central role in real-world decision making.

It is also important to underscore that novelty cannot be experimentally varied. As discussed in Section 2, novelty is inseparable from proposal content and therefore cannot be independently manipulated while holding quality constant. The study accordingly treats novelty as an observed

attribute and relies on a wide range of specifications, fixed effects, and saturated controls to assess the robustness of its relationship with evaluations. While these analyses are unusually thorough, they do not convert novelty into a causal treatment. The findings on novelty should therefore be interpreted as descriptive of evaluative patterns rather than as causal effects—an important distinction that reflects a fundamental constraint shared by nearly all studies of frontier innovation.

3.3.1 What Have We Learned from Reviewer Manipulation Studies So Far? Reviewer Composition Matters Most When It Shapes Core Evaluative Judgment—Not Checklists

Across reviewer-manipulation experiments, the clearest and most consequential effects arise when interventions alter core evaluative judgment rather than when they add auxiliary forms of expertise or procedural guidance. Incremental differences in who evaluates—particularly in reviewers' intellectual proximity to the work—systematically shape scores and rankings, while add-on interventions such as statistical reviewers, librarians, or checklists tend to generate smaller, more conditional, or indirect effects.

The strongest evidence comes from designs that vary core expertise along a continuous margin. In Boudreau et al. (2016), small, exogenously induced differences in intellectual distance between reviewers and submissions produce large, directional shifts in evaluation. Closer experts evaluate more critically, depressing scores by nearly a full point and reshuffling rankings substantially. These effects are homogeneous, robust, and consistent with bounded-rationality models of inference at the knowledge frontier: evaluators' epistemic position shapes what they perceive as plausible, flawed, or promising.

By contrast, additive-expertise interventions—adding statistical reviewers (Arnau et al. 2003; Cobo et al. 2007) or information specialists (Rethlefsen et al. 2025)—yield more modest effects and are harder to interpret. Such designs bundle reviewer count and reviewer type, rely on coarse or highly aggregated outcome measures, and operate within institutional settings that limit how auxiliary input translates into author revisions or editorial decisions. Even when specialized reviewers influence outcomes, their effects often appear indirectly—through increased rejection or triage—rather than through measurable improvements in standardized reporting metrics.

These findings also clarify why checklist-based interventions frequently underperform expectations. Scientific evaluation requires synthesis across theory, evidence, and feasibility; it cannot be reduced to mechanical compliance with reporting standards. Auxiliary tools may surface issues or

reduce variance at the margins, but they do not substitute for variation in the evaluator's core interpretive frame.

Viewed through the structural regularities outlined in Section 2, these patterns are unsurprising. When true quality is uncertain and forward-looking, evaluation hinges on expert inference rather than rule application. Small differences in reviewers' knowledge and epistemic position therefore generate systematic differences in judgment, and such differences cannot be replicated by adding procedural scaffolding or auxiliary reviewers. Reviewer-manipulation studies thus identify one of the few experimental margins that reliably shifts substantive evaluative outcomes: changes in who evaluates, not in how evaluation is formatted or supplemented.

3.4   Manipulations of Author Identity (A) and Blinding

Far more randomized studies examine blinding than any other intervention margin reviewed in Section 3. Although blinding manipulates a comparatively narrow feature of the evaluation environment, the literature is disproportionately concentrated here: roughly 18 blinding experiments, versus about six studies on average in each of the other intervention categories. This skew likely reflects a combination of substantive salience—long-standing speculation regarding bias, prejudice, and the role of prestige—and practical tractability. Relative to interventions that require redesigning reviewer assignment, incentives, or editorial architecture, blinding is a discrete, easily implemented treatment that many journals and conferences could vary without disrupting core workflow. This is especially so if a context begin with a policy of single blinding in a broader context where double blinding was becoming more normalized.

Across dozens of randomized trials, the most robust conclusion is simple: author-identity information causally changes peer-review behavior and outcomes. When names, affiliations, and reputational cues are visible, evaluations shift—sometimes in scores, sometimes in recommendations, sometimes in willingness to review—relative to conditions in which identity is masked. This is established in early journal trials that randomize blinding within manuscripts (McNutt et al. 1990), in large-scale journal field experiments (Blank 1991), in objective "error-seeding" designs (Godlee, Gale & Martyn 1998), and in more recent high-powered settings ranging from conferences to grants and invitation experiments (Pleskac et al. 2025; Nakamura et al. 2021; Huber et al. 2022). The existence of identity-linked effects is not in serious doubt.

Two distinct questions should then be separated. The first concerns the first-order effect of blinding on evaluative stringency or level—whether concealing author identity systematically raises or lowers scores, recommendations, or acceptance probabilities on average. The second

concerns heterogeneous effects—how the visibility of author identity interacts with particular author characteristics.

On the first question, the literature does not point in a uniform direction. In some settings, masking identity appears to increase evaluative stringency or criticality, as reflected in higher editor-rated rigor (McNutt et al. 1990) or lower acceptance rates under double-blind review (Blank 1991). In other settings, blinding reduces rejection (Godlee et al. 1998) or leaves average evaluations largely unchanged while altering other margins of behavior.

On the second question—interactions between blinding and author characteristics—interpretation is necessarily more limited. Although within-submission designs identify the causal effect of identity visibility itself, interactions between blinding and specific author attributes cannot generally be interpreted causally. This is because any personal attributes tend to be correlated with other personal attributes (A), and personal attributes will tend to be correlated with features of the submission (Q). In any case, the direction and magnitude of identity interactions often differ across studies: effects are concentrated among mid-ranked institutions in Blank (1991), among senior coauthors in Pleskac et al. (2025), in race-linked comparisons in Nakamura et al. (2021), and are extreme when prominence is deliberately manipulated, as in Huber et al. (2022), and sometimes positive or negative effects of gender differences.

Evidence on "performance" or welfare is thin—but has been attempted. Some studies introduce benchmarks—most cleanly, the seeded-error design of Godlee et al. (1998)—but such benchmarks capture only one facet of review, and null effects on error detection can coexist with meaningful shifts in recommendations. More recent work attempts to link review scores to downstream indicators (Pleskac et al. 2025), but those outcomes are themselves noisy and only partially observed. The latter seem an especially promising route for establishing efficiency effects. This is especially important, as this literature has yet to link its clean causal designs to mechanisms or clear theoretical interpretation.

Blinding Reviewers to Author Identity at the Journal of General Internal Medicine: McNutt, Evans, Fletcher & Fletcher (1990)

The question of whether concealing author identity from reviewers improves the quality of peer review sits squarely within the Q-A-R-S framework: if evaluation should track the quality of the work (Q) rather than the identity of who produced it (A), then removing A from the reviewer's information set should—if A otherwise biases judgment—improve the correspondence between

review quality and manuscript merit. McNutt, Evans, Fletcher, and Fletcher (1990) conducted one of the earliest randomized controlled trials to test this hypothesis directly.

The study was embedded within the Journal of General Internal Medicine (JGIM), the official journal of the Society of General Internal Medicine—a small, academic medical journal with a rejection rate of 62% at the time. The journal's standard policy was single-blind review: reviewers saw author and institution names, but authors did not know reviewer identities. The editorial team—R. Fletcher and S. Fletcher served as editors, McNutt and Evans as physician-investigators—designed the trial as an editor-led collaboration, giving them full control over manuscript routing, blinding procedures, and outcome measurement. Reviewers were drawn from a pool of over 600, mostly academic physicians and Society members, selected for content expertise or demonstrated reviewing skill. These were experienced insiders to the general internal medicine community, not methodological specialists or outsiders.

The design was a randomized, controlled, double-blind trial using blocked randomization. Each of 127 consecutive original research manuscripts defined a block of two reviewers. Within each block, one reviewer was randomly assigned to receive the manuscript with authors' and institutions' names removed (blinded), while the other received it as submitted (unblinded). Both reviewers knew they were participating in a study—both groups received questionnaires about their attitudes toward blinding—with the idea that Hawthorne effects should operate similarly across arms. Reviewers were encouraged, but not required, to sign their reviews. The primary outcome was review quality as rated by editors, who were themselves blinded to both the manuscript's authors and the reviewer's treatment assignment. Editors graded each review on a 5-point scale across multiple dimensions: importance of the question, targeting key issues, methods assessment, and presentation (from the editor's perspective), plus thoroughness, constructiveness, fairness, courteousness, and knowledgeability (from the author's perspective), culminating in a summary grade. Authors also graded both reviews without knowing which reviewer had been blinded.

The blinding procedure in this case required an editorial assistant retyped title pages, applied opaque tape to remove identifiers throughout headers, footers, text, tables, and figures. No attempt was made to remove self-citations. Blinding was successful for 73% of reviewers (76% blinded to authors specifically). Unblinding occurred because reviewers recognized the author's prior work (27%), authors cited their own publications (31%), or editorial error (23%, mostly early in the study before tables and figures were systematically checked).

The results showed a modest but statistically significant effect. Editors rated blinded reviews higher on three of four dimensions from their own perspective: importance of the question, targeting key issues, and methods (all $p < 0.02$), with the largest difference on methods assessment. The summary grade—incorporating both editor and author perspectives—was 3.5 for blinded reviewers versus 3.1 for unblinded reviewers, a mean paired difference of 0.41 on a 5-point scale ($p = 0.007$). The distribution shift was most visible at the top: 17% of blinded reviews received the highest grade (5, "excellent") compared with only 6% of unblinded reviews. Notably, blinded reviewers gave lower scores to manuscripts on originality, methods, and presentation—significantly so for methods ($p \leq 0.05$)—suggesting that blinding increased critical scrutiny. Yet blinded and unblinded reviewers did not differ in their final publication recommendations.

Authors, by contrast, detected no difference in review quality between the blinded and unblinded conditions across any dimension or in summary grades. This divergence—editors perceived improvement, authors did not—suggests that blinding affected the evaluative rigor visible to editorial gatekeepers more than the tone or helpfulness perceived by manuscript submitters.

A secondary analysis examined signing, although this was not an experimental assignment and therefore should be interpreted as reflecting endogenous choices and as a correlation in that light. Forty-three percent of reviewers signed their reviews, with no difference between blinded and unblinded groups. Signing was not associated with overall review quality, but signers were rated as more constructive and courteous by editors and fairer by authors. Signers also gave higher manuscript scores and recommended acceptance more often—suggesting that signing may mark a more personalized, lenient reviewing style rather than a quality-enhancing intervention.

The study's strengths are notable for its era. The within-manuscript paired design controls for manuscript quality (Q): each paper serves as its own control, with variation arising only from reviewer assignment. Random assignment balances reviewer characteristics (R) across conditions in expectation. The editorial team's direct involvement enabled tight procedural control and credible outcome measurement. The study was adequately powered (estimated 98 manuscripts needed; 127 enrolled) and used appropriate nonparametric statistics given the ordinal outcome scale.

Several limitations temper interpretation. First, the dependent variable—editor-rated review quality—is subjective and reflects editorial preferences rather than any external benchmark of review accuracy or predictive validity. The study cannot determine whether blinded reviews were "better" in any ultimate sense (e.g., more accurate in identifying flawed work, more predictive of downstream citation or replication). Second, blinding was incomplete: 27% of reviewers in the blinded arm

successfully identified the authors, though there was no quality difference between those who unblinded and those who remained blinded. Third, the effect size (0.41 on a 5-point scale, or ~8% improvement) is modest and concentrated in editorial perception rather than author experience or publication recommendations. Fourth, the setting—a small, specialized medical journal with an academically homogeneous reviewer pool—may limit generalizability to larger journals, other fields, or venues with different reviewer cultures.

Viewed through the Q-A-R-S lens, the design holds (Q) constant within manuscript pairs and randomizes the information about (A) available to reviewers, while (R) is balanced across arms by randomization and (S) (the review system) is held constant except for the blinding manipulation itself. The study thus achieves relatively clean identification of the effect of author-identity information on review quality as perceived by editors. What it cannot identify is whether this effect operates through reduced bias (reviewers no longer favor prestigious authors), increased scrutiny (reviewers work harder when they cannot rely on reputational shortcuts), or some other mechanism. The finding that blinded reviewers were more critical of manuscripts—without changing acceptance recommendations— suggests that blinding may sharpen evaluative attention rather than simply level a biased playing field. But without variation in author prestige or an external quality benchmark, the study cannot distinguish these channels.

McNutt et al. (1990) established that blinding is feasible, that it modestly improves review quality from the editor's perspective, and that randomized experimentation on peer review is possible within working editorial systems. It set a methodological template for later blinding trials and demonstrated the value of editor-researcher collaboration. Its limitations—subjective outcomes, incomplete blinding, modest effects, narrow generalizability—would recur across the subsequent literature, leaving open the question of whether blinding improves the accuracy or fairness of peer review in any deeper sense.

Double-Blind Versus Single-Blind Reviewing at the American Economic Review: Blank (1991)

Blank (1991), in *"The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review,"* conducted one of the earliest and largest-scale field experiments on peer review, testing whether double-blind review—in which author identities are concealed from referees—produces different acceptance patterns than the conventional single-blind process. The experiment was editor-initiated and implemented in close collaboration with the editorial office of the American Economic Review, the flagship journal of the American Economic Association and one

of the most prestigious outlets in the discipline. At the time, the journal operated under single-blind review: referees saw author names and institutional affiliations, while authors did not know who reviewed their work. The study was subsequently published in the same journal in which the experiment was conducted and addressed a broad set of questions: whether concealing author identity changes referee behavior and editorial outcomes, whether effects vary across author characteristics, and whether double-blind review is operationally feasible in a field where identity cues may leak from content.

The experiment was motivated in part by concerns raised in the mid-1980s by the American Economic Association's Committee on the Status of Women in the Economics Profession about potential disadvantages under single-blind review, and more broadly by growing attention within economics to issues of credibility and empirical reliability (e.g., Leamer, 1983; Dewald, Thursby, and Anderson, 1986). In response, the AER editor and Board of Editors asked Blank—then at Northwestern—to design a randomized field experiment. The editorial office was physically located at Princeton, and paper-based manuscript handling made centralized randomization and masking operationally feasible in the pre-internet era.

In the study, over an 18-month period, approximately 1,500 manuscripts submitted to the AER were randomized into one of two conditions based on submission number: odd-numbered submissions were processed under double-blind review (author names and affiliations removed before being sent to referees), while even-numbered submissions proceeded under the journal's standard single-blind process. This alternating assignment ensured that, in expectation, the two groups were balanced on observable author characteristics—gender, institutional rank, domestic versus foreign affiliation—and, by extension, on the underlying distribution of manuscript quality (Q). Referees in the double-blind condition were informed that author identities had been masked; those in the single-blind condition reviewed manuscripts as usual.

The study found that acceptance rates were significantly lower under double-blind review. Among double-blind submissions, 10.6% were accepted, compared with 14.1% under single-blind review—a difference of 3.5 percentage points, or roughly 25% fewer acceptances. When Blank disaggregated by institutional prestige, the pattern became more nuanced: authors at the very top departments (ranked 1–5) showed little difference in acceptance rates across conditions, as did authors from small colleges and low-ranked institutions. The largest differences appeared for authors from near-top and mid-tier universities (ranked 6–50), U.S. nonacademic institutions (such as the Federal Reserve or policy think tanks), and foreign institutions—groups whose acceptance rates were

noticeably lower under double-blind review, though most subgroup estimates were statistically imprecise. Gender differences were modest and statistically fragile. Women's acceptance rates were slightly higher under double-blind review than under single-blind review—but these differences were small and not statistically significant once controlling for institutional rank. These estimates were also underpowered given the relatively low share of female-authored submissions at the time. Equally important, the study found that referees in the double-blind condition correctly guessed author identity in approximately 46% of cases. When referees asserted they could identify the author, they were correct most of the time, and particularly so for well-known authors.

Blank interpreted the results as showing that "referees were more critical," broadly, under double-blind review and that manuscripts evaluated without author identifiers faced "more stringent standards," as reflected in lower acceptance rates. She emphasized that concealing author identity "does affect the review process," but cautioned against strong conclusions about specific sources of bias.

Viewed through the Q–A–R–S framework, the design randomizes exposure to author identity and attributes (A) while balancing submission characteristics and quality (Q) across conditions; reviewer characteristics (R) are distributed by the natural flow of assignments, and the review system (S) is held constant by the venue. Imperfect blinding therefore reflects limits of institutional anonymization rather than a flaw in the experimental design and should shape interpretation of the estimated effects rather than serve as a criticism of the study.

A range of explanations could account for the observed patterns. The concentration of lower acceptance rates under double-blind review among near-top and mid-tier authors may reflect uneven leakage of identity cues, non-linear relationships between quality and reputation, changes in the use of reputational signals, or shifts in referee scrutiny under uncertainty. Blinding could also reduce evaluative efficiency if reviewers rely on author identity as an informational signal when submission quality is ambiguous. Absent an external benchmark of "true" quality, the design cannot discriminate among these accounts.

Although the American Economic Review adopted double-blind reviewing in the years following Blank's experiment, contemporaneous editorial records indicate that this shift was not a direct or dispositive response to the study's findings, given modest and statistically fragile gender

effects and unresolved mechanisms. The AER (and other AEA journals) again reversed its double-blind policy by editorial board vote years later, around 2011.[14]

Blank (1991) remains historically significant as one of the first relatively high-powered, editor-initiated field experiments on peer review and the first to study blinding using randomized assignment in a major economics journal. Beyond documenting statistically detectable changes in acceptance rates and referee behavior, the study cleanly established that concealing author identity alters evaluative outcomes. Although often retrospectively cited as a test of gender bias, its most robust contributions concern the role of author prestige, reviewer scrutiny, and the limits of institutional design in shaping evaluation. The experiment demonstrated the capacity of double-blind review to change acceptance patterns but left unresolved the reasons why or whether these changes were efficient and productivity-enhancing or not.

Blinding and Error Detection at the British Medical Journal: Godlee, Gale & Martyn (1998)

Godlee, Gale, and Martyn (1998) conducted one of the earliest fully embedded randomized controlled trials designed to test two commonly proposed reforms to peer review: blinding reviewers to author identity and requiring reviewers to sign their reports. The study was implemented in collaboration with editors at the BMJ (British Medical Journal), one of the world's leading general medical journals, which at the time operated under conventional single-blind review with anonymous referee reports. The experiment addressed a pair of linked concerns that had long animated debates about peer review: that knowledge of author identity might bias evaluation, and that anonymity might reduce reviewer accountability and effort.

The institutional setting was significant. The BMJ had a large, international readership and drew reviewers from across the biomedical research community—predominantly academic clinicians and researchers in the UK and Commonwealth, but also including substantial numbers from Europe and North America. The journal's editors had shown sustained interest in the empirical study of peer review, having previously collaborated on research examining reviewer characteristics and report quality. This culture of editorial self-scrutiny made the BMJ an unusually receptive site for embedded experimentation.

---

[14] This was not the first instance of double blinding at the AER. In 1973, editor George Borts initiated a first temporary "experiment" with double-blind reviewing (without control group or randomization), using it "for the remainder of my term as editor" after concluding that it was perceived as fairer and not excessively costly (Borts 1974, p. 476). That policy was subsequently reversed when "the editor appointed in 1979 returned the journal to a single-blind system" (Blank 1991, footnote 9).

The study employed a novel methodological approach: rather than relying on naturally submitted manuscripts with unknown true quality, the researchers used a single, recently accepted BMJ article—a study on cognitive impairment and mortality in older adults—into which the editors deliberately inserted eight errors spanning design, analysis, and interpretation. These included a tweaked statistical result, an incomplete description of randomization procedures, and several interpretive overstatements. The use of a fixed manuscript with known, researcher-inserted flaws allowed the dependent variable to be precisely defined: the number of the eight seeded errors that each reviewer explicitly mentioned in their report. This "error-injection" design offered an objective benchmark for review quality that did not depend on subjective editorial judgment.

The experimental design was a 2×2 factorial with a fifth comparison group. A total of 420 reviewers were randomly drawn from the BMJ's reviewer database and assigned to one of five conditions: (1) blinded to author identity and asked to sign their report; (2) blinded and anonymous; (3) unblinded and asked to sign; (4) unblinded and anonymous; and (5) a "usual practice" control group that received the manuscript under standard BMJ procedures—unblinded, anonymous, and with no indication that a study was underway. The comparison between groups 4 and 5 allowed the authors to test whether mere awareness of participating in a study altered reviewer behavior. Randomization was weighted to oversample the "signed" conditions, since a pilot study suggested that many reviewers would decline to sign.

A total of 221 reviewers (53%) returned reports. The response rate did not differ significantly across conditions, though the authors noted that the rate was somewhat lower than typical for BMJ reviewers, possibly because some invitees felt the manuscript topic was outside their expertise. The primary outcome—number of seeded errors mentioned—was assessed independently by an editor and an epidemiologist, both blinded to treatment assignment, with disagreements resolved by consensus.

The headline result was a null finding on error detection. Reviewers across all five conditions identified, on average, approximately two of the eight inserted errors (means ranged from 1.7 to 2.1 across cells). Only 10% of reviewers detected four or more errors, and 16% detected none. There were no statistically significant differences across the blinded versus unblinded conditions or across the signed versus anonymous conditions. The authors concluded that "neither blinding reviewers to the authors and origin of the paper nor requiring them to sign their reports had any effect on rate of detection of errors."

Yet the study revealed meaningful treatment effects on other dimensions of reviewer behavior that the narrow error-count metric did not capture. Most notably, reviewers who were blinded to author identity were substantially less likely to recommend rejection than those who saw the authors' names. The odds ratio for recommending rejection (blinded versus unblinded) was approximately 0.5, and this effect strengthened (OR ≈ 0.3) when the analysis excluded reviewers who had successfully identified the authors despite blinding. Blinding thus shifted evaluative stance in a more favorable direction—reviewers were more lenient when they did not know who had written the manuscript— even though it did not change their detection of specific errors.

This divergence illustrates a potential limitation of error-injection designs, as in the audit studies of Section 3.2: they might capture only a narrow slice of what reviewers do. Reviewers assess coherence, novelty, clarity, plausibility, and evidentiary weight; they calibrate scrutiny, decide whether to participate, and make holistic recommendations that determine publication outcomes. Therefore, it remains plausible that a reviewer can produce a thoughtful report without mentioning a single injected error—and conversely, we might imagine a reviewer who catches several errors but still fails to calibrate the contribution.

This divergence between the null result on error detection and the significant result on recommendation behavior illustrates a central limitation of error-injection designs: they capture only a narrow slice of what reviewers actually do. Peer review involves far more than spotting pre-planted defects. Reviewers assess coherence, novelty, clarity, plausibility, and evidentiary weight; they calibrate their level of scrutiny, decide whether to participate at all, and make holistic recommendations that ultimately determine publication outcomes. A reviewer could produce a more rigorous, skeptical, or constructive report without ever mentioning a single injected error—and conversely, a reviewer who catches several errors might still recommend acceptance. By construction, error-count metrics cannot capture these broader evaluative behaviors.

The blinding procedure itself achieved only partial success. Among the 90 reviewers in the blinded conditions, 23 (26%) correctly named the authors in their reports, typically because they recognized the research from prior presentations, inferred identity from self-citations, or noticed residual identifying information that editorial masking had missed. This rate of "unblinding" is lower than in Blank's (1991) economics experiment (46%) but still substantial, and it underscores the difficulty of achieving complete anonymity in fields where intellectual communities are tightly networked and research programs are distinctive.

Viewed through the Q-A-R-S framework, the design achieves unusually tight control over Q by using a single manuscript with a fixed set of known errors—an approach that eliminates variation in true quality across observations. The randomization of blinding and signing manipulates both A (author identity, visible versus masked) and a feature of (S) (the review system's accountability structure). Reviewer characteristics (R) are distributed across conditions by random assignment. The within-paper design and objective outcome measure (error counts) provide strong internal validity for testing whether these system-level interventions affect error detection specifically.

However, the design's strength is also its limitation. By collapsing "review quality" into a single, narrow metric—errors mentioned—the study cannot detect treatment effects on the many other dimensions of review that matter for editorial decision-making. The significant effect on rejection recommendations demonstrates that blinding does change reviewer behavior, but this effect is invisible to the primary outcome measure. The study thus illustrates a broader methodological lesson: interventions may leave error counts unchanged while producing meaningful differences in participation, tone, scrutiny, and recommendation patterns. Error-injection designs are valuable stress tests of one facet of evaluation, but their inferential scope is inherently constrained.

Several additional limitations bear mention. First, the use of a single manuscript limits generalizability: the inserted errors may have been more or less detectable than typical flaws, and the paper's topic (cognitive impairment in elderly cohorts) may have been more or less familiar to the BMJ reviewer pool than an average submission. Second, the sample size per cell (approximately 40–60 reviewers) may have limited power to detect modest effects on error detection, particularly if true effect sizes were small. Third, the 53% response rate, while acceptable, raises the possibility of differential selection into participation across conditions—reviewers who declined may have differed systematically from those who responded.

Godlee et al. (1998) remains a landmark study as one of the first large-scale, editor-led randomized trials embedded in a working journal, and one of the few to use a researcher-controlled benchmark of error detection for review quality. The study's headline null result on error detection has been widely cited as evidence that blinding and signing do not improve peer review. (The broader experimental literature on "accuracy" outcomes, however, raises the question of whether codifiable error detection is the most consequential dimension of evaluation.)

Experimenting with Blinding in a Decision-Making Conference—Notably Reporting Subsequent Outcomes: Pleskac et al. (2025)

Pleskac et al. (2025) report a field experiment comparing single-blind and double-blind peer review embedded directly in the submission review process of the Society for Judgment and Decision Making (SJDM) annual conference. The author team—Pleskac (decision sciences/psychology; long-standing SJDM member and conference chair), Kyung (marketing and judgment/decision making; active in SJDM governance and equity initiatives), Chapman (psychology and medical decision making; former SJDM president and senior figure in the society), and Urminsky (marketing and behavioral decision research; senior editor)—were all deeply embedded in the society's organization and stewardship. The team was perhaps unusual in its especially balanced methodological capabilities, relevant theoretical depth, and strong institutional knowledge and access to the research site, rather than being highly specialized across its team members.

The experiment emerged from an explicit internal debate within SJDM in the mid-2010s. By that time, double-blind review had become close to a disciplinary norm, while SJDM continued to rely on single-blind review, raising questions about why the conference had not adopted prevailing evaluation practices and whether this choice might affect the representation of junior scholars, women, or other groups. Observers noted descriptive disparities—such as a historically low proportion of women among conference attendees, reviewers, and leadership—though these patterns lacked any counterfactual basis for inference about the role of blinding. The internal discussion also surfaced competing concerns: some argued that author identities conveyed valuable information about speaker quality or audience draw, while others worried that blinding might implicitly accuse reviewers of unfairness, potentially eroding trust within a tightly networked scholarly community. Although the research team had not previously collaborated, they were familiar with one another through the venue and shared an interest in bringing appropriate data to bear on these questions.

Operationalizing the experiment required more than modifying an existing system. The team worked with a developer to design and build a new information technology platform that accepted submissions, managed reviewer assignments, and embedded experimental treatments, while also defining the core workflow and interfaces used by participants. The platform enabled collection of participant characteristics relevant to the study. Beyond reproducing the historical submission and review process with high fidelity, the primary design goals were to embed treatments unobtrusively while complying with IRB requirements.

Each of the 530 submissions—600-word conference abstracts rather than full papers—was reviewed under both blinded and non-blinded conditions, enabling within-submission comparisons while ensuring that no paper was systematically advantaged or disadvantaged in the live review process. Reviewers were assigned to a single review format to minimize the salience of the treatment, so identification relied on within-submission, cross-reviewer variation rather than within-reviewer contrasts. Implementing this design required balancing reviewers across conditions, which led the authors to recruit more than twice the usual number of reviewers—112 in total. Consistent with standard conference practice, each reviewer evaluated approximately 30 abstracts (mean ≈ 31), providing a single overall quality rating on a nine-point scale, yielding nearly 3,500 individual review observations.

The scoring results revealed substantial instability in peer review outcomes, with only moderate agreement between review systems. When average scores were computed separately for single-blind and double-blind reviewers for each of the 530 submissions, the Pearson correlation between these averages was approximately 0.54. As a consequence, fewer than half of the submissions that would be accepted as talks under one system would also be accepted under the other. The authors documented similarly high levels of instability within each review regime itself, indicating that reviewer-specific variation was a dominant feature of the evaluation process and complicating efforts to attribute rank differences cleanly to treatment effects. Agreement was also asymmetric across the score distribution: convergence was substantially higher for lower-rated submissions (correlation ≈ 0.45) than for higher-rated submissions (correlation ≈ 0.19), implying that rank orderings diverged most in the upper portion of the distribution where acceptance decisions are made.

Turning to systematic differences across review conditions, the authors examined whether author attributes were weighted differently when identities were visible. They found that more senior coauthors received higher scores—on the order of 0.19 standard deviations per rank increase in average coauthor seniority—under single-blind review, an effect largely absent under double-blind review. Single-blind review was also associated with lower scores for Asian first authors by approximately 0.17 standard deviations. Gender effects were more ambiguous: estimates were small, heterogeneous, and did not align with common priors that blinding disproportionately benefits women.

For the 108 submissions accepted as talks, the authors collected detailed conference-outcome data during the meeting itself. Each talk was attended in person by four independent observers— drawn from a pool of 18 faculty members and 12 doctoral students or postdoctoral researchers—who

were randomly assigned to sessions to record attendance, count audience questions, and provide structured evaluations of talk quality. For the accepted talks, neither the average single-blind nor the average double-blind review scores predicted any of the collected outcome measures—judged talk quality, observed attendance, or the number of audience questions. For poster presentations ($N \approx 56$), both single-blind and double-blind average scores showed modest, positive relationships with poster-quality ratings, with no credible difference between systems. Extending the analysis to all 530 submissions, both types of review scores were also correlated with the likelihood of eventual publication several years after the conference, but not different from each other. There was no correlation with journal impact factor for those published.

Although blinding policy was the central thrust and motivation of the paper, the authors emphasized that the dominant feature of the data was noise rather than systematic treatment effects, stressing that "differences between review systems partly stem from noise in the review system," that "both review systems exhibit only moderate reliability," and that "the conference lineup could differ substantially depending on which reviewers or which review system is used." The authors suggested that certain attributes "function differently across single- and double-blind review," particularly seniority and race, while cautioning that effects were heterogeneous and that gender patterns were "small and do not support a simple narrative." In assessing downstream outcomes, they emphasized that "there is no evidence that single-blind review is more valid than double-blind review," concluding that "including author identities does not appear to enhance reliability or validity enough to justify the risks this information poses."

Viewed through the Q–A–R–S lens, the design systematically varies author-identity information (A). The within-submission variation is particularly attractive: each submission receives multiple reviews under both regimes (at least six total, three per arm), creating a natural setting for submission-level fixed effects that would fully absorb submission-specific characteristics and latent quality (Q). In practice, the authors implement submission-level random effects. Each reviewer (R) evaluates many submissions (approximately 31 on average) but appears in only one treatment arm. Because reviewer fixed effects are collinear with treatment assignment, identification relies on balance induced by random assignment across the 112 reviewers. Additional description of reviewer characteristics across treatment arms—and analysis of heterogeneity (e.g., contrasts between more experienced and newly recruited reviewers)—could therefore be informative. Aside from reviewer assignment, features of the peer-review system (S) are tightly controlled by the platform infrastructure and process design.

The two main limitations of the study are types of limitations that will be faced by any blinding study. First, interpretation of blinding effects and how they might differ across authors with different characteristics (e.g., seniority, institutional affiliation, race, and gender) are highly collinear among themselves and with other unobserved attributes. Thus, estimates of how blinding differentially affects evaluation across author attributes should always be understood as reduced-form associations that may be correlated with any number of other attributes (Section 2.2.3). Second, even where were associations to be interpreted as causally related to a particular attribute, this design and most others cannot distinguish among competing mechanisms or explanations.

Importantly, these limitations do not detract from the study's core contributions, given the particular findings. Because the estimated interactions between blinding and author characteristics are nuanced, often unintuitive, and in several cases at odds with prior findings, they serve as existence results that go against simplest interpretations. Second, the study pioneers a systematic attempt to assess the downstream consequences of blinding by linking review scores to observable outcomes. The design cannot directly answer how overall conference quality would have differed under alternative review regimes, but it the next best strategy by testing whether review scores predict outcomes at all. For the most part, the data do not reveal any relationships. It is unclear whether this is the result of underpowered estimates, or truly zero differences. Nonetheless, this is a quantum step forward in terms of research design in this area. Perhaps further study of these downstream effects can finally allow us to begin to understand efficiency implications of blinding.

3.4.1   *Other Randomized Studies Examining Blinding Effects on Evaluation Scoring*

In addition to the anchor studies reviewed above, there are a small number of randomized experiments that address the same evaluative question. These studies intervene on author identity (A) and measure outcomes that proxy for judgments about manuscript quality (Q), while relying on across-submission randomization rather than within-submission comparisons.

Fox et al. (2023) implement a randomized comparison of single- versus double-blind review at *Functional Ecology*, finding that evaluations under double-blind review, on average, lead to lower scores and lower probabilities of invitation to revise or acceptance. The negative effect is largest for authors from high-income and high–English-proficiency environments. This pattern contrasts with Blank (1991), where differences between single- and double-blind review are concentrated among authors from near-top and mid-tier institutions. Fox et al. also find that the gap in evaluations between male- and female-authored submissions is similar under single- and double-blind review, echoing Blank's conclusion that gender effects are modest and statistically fragile.

John et al. (2019) randomize the disclosure of authors' conflicts of interest to reviewers—holding author identity fixed—and find no detectable effect of disclosure on reviewers' quality ratings, even for manuscripts reporting financial conflicts. Despite being well powered, the study shows that providing conflict-of-interest information neither lowers evaluation scores nor increases rejection or major-revision recommendations. Reviewers report that conflicts of interest are important and believe they can correct for them when disclosed, but their quantitative assessments remain unchanged.

A methodologically distinct group of blinding experiments study how double blinding affects the referee report itself—rather than the scoring of the manuscript. Justice et al. (1998) and van Rooyen et al. (1998, JAMA) both examine whether blinding reviewers to author identity affects review quality, as proxied by editor's scoring of review quality. Both find differences that are small in magnitude and not editorially meaningful. Across these studies, estimated effects on global review-quality scores are typically on the order of one-tenth of a point on five-point scales, well below thresholds the authors define ex ante as substantively important. Similar null or near-null findings appear in Fisher et al. (1994), Alam et al. (2011), Vinther et al. (2012), and Okike et al. (2016), which examine review length, tone, time spent reviewing, and related features of referee reports in specialty medical journals.

A distinct experimental approach is used by Clarke et al. (2016), which shifts the unit of analysis from individual reviewers to entire review panels by duplicating the full evaluation process within Australia's National Health and Medical Research Council Early Career Fellowship competition. The study reruns the complete review pipeline for 60 identical applications, each evaluated independently by two separate review panels, holding submission content and author identity fixed while varying only panel composition. In this sense, the paper also aligns with earlier resubmission and duplication studies (Section 3.1), though it is embedded directly within a live funding system. Each panel consisted of four reviewers who scored applications independently, after which scores were aggregated, ranked, and discussed in panel teleconferences to determine final funding decisions. Rather than exploiting variation in individual reviewer assessments, the design therefore captures the combined effects of aggregation and within-panel deliberation on final outcomes. The study finds 83% agreement in funding decisions across panels, indicating substantially higher reproducibility than would be suggested by individual-level reviewer disagreement alone. This design deliberately trades statistical leverage at the reviewer level for institutional realism, providing rare evidence on the stability of evaluation outcomes at the point where collective judgments are translated into consequential decisions.

Nakamura et al. (2021) report a large-scale randomized experiment using 1,200 real NIH R01 grant proposals from the 2014–2015 review cycles, re-reviewed for research purposes only, with no funding stakes attached. The central comparison is between 400 applications from Black investigators and 400 applications from White investigators selected to match the Black applications on original NIH impact scores and other review-relevant characteristics. Reviewers were recruited from lists of prior NIH study-section members, and submissions were randomly assigned to be reviewed with applicant identity visible or administratively redacted. Under unblinded review, applications from White investigators scored higher on average; blinding reduced this advantage primarily by lowering scores for White applications. However, in this matched comparison, the difference in the effect of blinding between Black and White applications (the interaction) was not statistically significant.

Huber et al. (2022) report on a field experiment embedded within the Journal of Behavioral and Experimental Finance using a single manuscript co-authored by V. Smith—a Nobel laureate—and Inoua, an early-career researcher at the time, both affiliated with the Economic Science Institute at Chapman University. The experiment involved sending the manuscript "Re-tradable Assets, Speculation, and Economic Instability" to more than 3,300 potential journal reviewers, with random assignment to conditions in which reviewers saw either (i) Smith's namelisted as corresponding author, (ii) no author name (full anonymization), or (iii) Inoua's name listed as corresponding author. Conditional on submitting a review, the share of reviewers recommending accept or minor revision was 9.9%when Inoua's name was visible, 23.6% under full anonymization, and 58.8% when Smith's name was visible. The study also documents systematic differences in invitation acceptance rates across author-identity conditions. Given the unusually large number of observations devoted to a single manuscript, these differences are estimated with high precision and are statistically significant at p-values well below 0.01.The findings are intuitively consistent with the idea that author prominence shapes peer review outcomes, although a number of particular mechanisms could be involved, along with alternative explanations. The study illustrates inherent trade-offs between concentrating statistical power, achieving interpretability, external validity, and generalizability, and costs.

3.4.2 *What Have We Learned from Blinding Studies So Far? Causal Designs, but Normative Implications Unclear*

Blinding studies constitute the most mature and most numerous experimental approach to studying peer review at scale (Figure 3). Blinding is a relatively easy treatment to implement, particularly in periods when unblinded review provided a natural control condition. From a Q–A–R–S perspective, blinding treatments often achieve comparatively clean causal identification, as they

directly manipulate the availability of author identity information (A) while holding manuscript attributes (Q), reviewer characteristics (R), and system architecture (S) fixed in expectation. Blinding studies have also attracted sustained interest, in part because blinding emerged as a widely adopted editorial practice beginning in the late 1980s and 1990s, creating unusually favorable conditions for large-scale causal evaluation.

Areas of agreement are clear. First, across the studies reviewed above, author-identity information causally affects evaluation outcomes. When authorship is visible, scores, recommendations, or acceptance probabilities differ by author characteristics such as institutional prestige, seniority, or demographic group. These effects appear in economics, computer science, medicine, and psychology, suggesting broad external validity of the existence of identity-linked evaluation effects. Second, blinding attenuates—but does not eliminate—these effects. Within-submission designs establish causality, but perfect anonymization is difficult, as reviewers can sometimes infer identity from content. The generalizable conclusion is therefore limited but consistent: blinding reduces the influence of explicit identity cues, though the magnitude varies across settings. Third, blinding often increases reviewer stringency or criticality, though direction and magnitude depend on context and outcome measure. Fourth, across studies that examine accuracy benchmarks or downstream predictive validity, blinding has not been shown to yield systematic improvements. Neither planted-error detection nor longer-run indicators such as talk quality, audience engagement, or eventual publication differ meaningfully across blinding conditions. In these senses, review "performance," as operationalized in these designs, is not clearly systematically improved by blinding.

Where studies diverge is equally important. The direction of blinding effects on decisions varies across settings. Blank (1991) finds lower acceptance rates under double-blind review at the American Economic Review, while Godlee et al. (1998) find that blinded reviewers at the BMJ are less likely to recommend rejection. These differences could reflect variation in outcome measures (editorial decisions versus reviewer recommendations), field norms, editorial moderation, or the distinction between holistic judgment and error-detection tasks. Similarly, heterogeneity in who is most affected by identity visibility does not replicate cleanly. Effects are largest for mid-ranked institutions in Blank (1991), for top-tier or well-known authors in Tomkins et al. (2017), and for senior coauthors in Pleskac et al. (2025). More recent work also emphasizes that these shifts occur against a backdrop of substantial reviewer disagreement and rank-order instability, such that noise in evaluation may rival or exceed systematic blinding effects, particularly near decision thresholds. The broad

consensus of decisions in blinded and unblinded panels in Clarke et al. (2016), also suggests the possibility—for example—that aggregation and deliberation may partially compensate for the removal of author-level signals at the individual reviewer level.

On the basis of the existing experimental literature, however, we do not yet know whether these effects emerge from reviewers attempting to use author background to better infer true quality (Q) when quality cannot be directly observed from the submission itself, or from other channels such as changes in scrutiny, effort, or evaluative thresholds (Section 2). Further, we do not yet know how to reconcile the apparent divergences in results across settings. As a result, the literature is strong on what changes when identity information is removed, but weak on why those changes occur or how they should be interpreted normatively.

Viewed through the structural regularities outlined in Section 2, the interpretive limits of blinding studies are intrinsic rather than accidental. Structural Regularity 1 implies that true scientific quality is not directly observable at the time of review, making auxiliary information—such as author reputation or institutional affiliation—a potentially informative input to evaluative inference. Structural Regularity 5 further implies that removing author identity alters the informational environment in which reviewers form expectations about quality, not merely the presence or absence of bias. As a result, while blinding studies cleanly establish that author identity causally shapes evaluation, they cannot distinguish whether this effect reflects inappropriate favoritism, rational statistical inference under uncertainty, or changes in reviewer scrutiny and effort. This ambiguity is not a weakness of the evidence, but a consequence of the structural role that auxiliary cues play in evaluation.

3.5 Training, Feedback and Guidance Interventions: Attempts to Augment Basic Reviewer Capabilities (R)

A distinct line of experimental work attempts to improve peer review by changing reviewers rather than submissions, assignments, or system rules. These interventions provide reviewers with training, feedback, mentoring, or performance evaluations, under the premise that reviewer skill can be upgraded through instruction or reflection. In Q–A–R–S terms, they intervene directly on the reviewer component (R), while attempting to hold constant manuscript quality and characteristics (Q), author attributes (A), and core features of system architecture (S) at the point of randomization.

No Effect of Reviewer Training Reported at a Journal Testbed: Callaham et al. (2002) [15]

Between the mid-1990s and early 2010s, Annals of Emergency Medicine functioned as a rare, sustained institutional testbed for empirical research on peer review. Led by Callaham in his dual role as senior journal editor and investigator, the journal hosted a coordinated series of observational studies, audits, and embedded randomized interventions, particularly examining effects of reviewer training, feedback, and mentoring. This program spanned observational and measurement studies (e.g., Callaham, Wears & Waeckerle, 1998; Callaham & McCulloch, 2011; Green & Callaham, 2011), audit and stress-test studies (e.g., Baxt et al., 1998, see Section 3.2), embedded non-randomized intervention at the journal (Callaham & Schriger, 2002), and randomized interventions (Callaham, Knopp & Gallagher, 2002; , Houry, Green & Callaham, 2012).

Therefore, this study is best read as one experiment within a broader, unusually sustained effort at Annals of Emergency Medicine to determine effects of basic training. In "The Effect of Written Feedback by Editors on Quality of Reviews: Two Randomized Trials" (JAMA, 2002), Callaham, Knopp, and Gallagher report effects of providing coaching and feedback on refereeing. This included two interventions. Study 1 targeted a population of 51 low-volume, low-quality reviewers, with median editorial quality ratings less than or equal to 3 on a 5-point scale. These reviewers were randomized to receive feedback or not during the subsequent study period (approximately 24 months). Feedback consisted of the editor's numerical quality rating of the reviewer's report on a 1–5 scale, along with a short written description of the general components of a high-quality review (e.g., identifying methodological issues, commenting on interpretation and presentation, and providing context useful for editorial decisions)—and did not related to the

---

[15] Among general medical journals, *JAMA*, alongside the *BMJ*, has been the most sustained institutional contributor to empirical research on peer review. Beginning in the late 1980s, editors affiliated with *JAMA* published a dense sequence of studies on peer review and editorial decision-making, including: measurement and validation of editor rating instruments for review quality (Feurer et al., 1994; Callaham et al., 1998); descriptive analyses of reviewer reliability, variability, and correlates of perceived quality (Evans et al., 1993; Black et al., 1998; Weber et al., 2002); randomized and quasi-randomized trials of author blinding, reviewer anonymity, and signed reviews (McNutt et al., 1990; Justice et al., 1998; van Rooyen et al., 1999; Godlee et al., 1998); experiments on editorial procedures and reviewer recruitment (Pitkin & Burmeister, 2002); and randomized trials of reviewer feedback and training (Callaham, Knopp & Gallagher, 2002), with a minority of studies including randomized interventions. This work was closely linked to *JAMA*'s editorial leadership and to the founding of the **International Congress on Peer Review and Biomedical Publication** in 1989 (Rennie, 1990; Rennie & Flanagin, 2014), which provided a recurring venue for dissemination of peer-review experiments conducted across multiple journals and disciplines. Unlike the *BMJ*, which later developed a more explicit in-house methods function supporting peer-review experimentation (e.g., Schroter et al.), *JAMA*'s contribution appears to have taken the form of a **long-running, editor-led research and convening program**, rather than hiring permanent internal staff researchers.

reviewer's own report. The authors report that editors were somehow blinded to the purpose of the study and treatments.[16]

Of the 51 subjects, 16 were excluded due to attrition and failure to complete at least two post-randomization reviews, leaving 15 reviewers in the control group and 20 reviewers in the treatment group. The main comparison was of changes in editors' average ratings of reviews by each reviewer, before and after the intervention (essentially a difference in differences comparison). Control reviewers exhibited a small, statistically insignificant increase in editor-assigned ratings (+0.16 on a 5-point scale, 95% CI [–0.26, 0.58]), while reviewers assigned to receive written feedback exhibited a small, statistically insignificant decline (–0.13, 95% CI [–0.49, 0.23]). The authors conclude that "minimal feedback from editors on review quality had no effect on subsequent performance of poor-quality reviewers, and the trend was toward a negative effect." Given the small sample and nontrivial attrition, it is natural to ask whether the null result reflects limited power or bias. With so few observations, it is unlikely that adding control variables could have addressed these issues, unfortunately. The paper also reports average of 5.8 post-treatment reviews by those in the control group and 4.7 in the treatment; and attrition of 10 of 25 reviewers (40%) in the control group and 6 of 26 feedback reviewers (23%) in the treatment. My back-of-the-envelope calculations indicate no statistically significant differences can be detected in these differences.[17]

Study 2 expanded the subject pool to 127 reviewers (excluding those from Study 1) with editorial ratings of 4 or lower, largely repeating the Study 1 design; however, treated reviewers were also shown the editor's quality ratings of other reviewers on the same manuscript and an example of an "excellent" review from another submission. Thus, the idea was to provide a stronger treatment over a larger sample. Of the 127 reviewers, 95 completed at least two post-randomization reviews, contributing 324 rated reviews over roughly 22 months. Control group reviewers exhibited a small increase in average ratings (+0.12, 95% CI [–0.20, 0.26]), while treated reviewers exhibited a similarly small increase (+0.06, 95% CI [–0.19, 0.31]). The authors summarize both experiments by concluding that "simple written feedback to reviewers seems to be an ineffective educational tool." The reported attrition rates and number of reviews were also statistically similar across arms in Study 2.

---

[16] From the study's reporting, it is unclear how this statement applied to editor/coauthor Callaham.

[17] The paper does not report measures of variance. Treating the number of post-randomization reviews as Poisson yields no statistically detectable difference in review volume ($p \approx 0.16$). Treating attrition as a reviewer-level Bernoulli outcome does not reject equality of attrition rates ($p \approx 0.19$).

Viewed through a Q–A–R–S lens, across two randomized trials the intervention perturbs information and feedback provided to reviewers to assess a change in R, while aiming to leave manuscript quality (Q), author attributes (A), and core system architecture (S) unchanged. The isolation of variation in R is a strength of the design. Notwithstanding concerns related to small samples, subjective scoring, and attrition, the consistency of null findings is somewhat persuasive in demonstrating that this particular feedback, delivered to these particular reviewers, had little effect. The more difficult question is how to generalize this attempted shock to R. One interpretation is that the treatment is generic and heavily "checklist" oriented; yet references to editorial judgment and exemplar reviews that may contain richer discussions of contribution complicate that characterization. In the absence of a theory of reviewer learning, and without analyses of review-by-review learning, dose–response effects, or reviewer heterogeneity, any generalization requires substantial speculation.

In principle, randomization of treatments to R should leave Q, A, and S balanced in expectation. With small samples and substantial heterogeneity across reviewers, authors, and manuscripts, imbalance remains a concern—though it is muted here by the absence of detectable effects. Finally, the design induces attrition by construction. When the treatment has no effect, this poses little risk of endogenous selection; ironically, had the treatment worked, attrition itself might have become systematically related to treatment assignment, further complicating calibration and interpretation. Based on the description, I am also tempted not just to think of this as "checklist" feedback but that kind of feedback to "low quality" reviewers, based on the sampling. But, it similarly requires considerably more investigation, girded by disciplined theory to attempt to better comprehend what we are observing here and to generalize beyond this sample.

No Effect of Stronger Training and Mentoring at the Same Journal Testbed: Houry et al. (2012)

In "Does Mentoring New Peer Reviewers Improve Review Quality? A Randomized Trial", Houry, Green, and Callaham (2012) ask whether the earlier non-results of Callaham, et al. (2002), above, reflect the limited intensity and personalization of the prior intervention. Both Houry, et al. (2012) and the earlier Callaham, et al (2002) were conducted at the same journal (Annals of Emergency Medicine) and led by the same senior editor–researcher (Callaham). This follow-on study is notable for its ambition: it evaluates a four-year mentoring intervention—among the most intensive ever implemented in a live journal—explicitly designed to succeed where lighter-touch feedback had not.

In the four-year period 2006–2010, 50 newly recruited reviewers entered the reviewer pool at Annals of Emergency Medicine through the journal's standard recruitment channels, including

editor nomination and self-nomination. These new reviewers over the four-year period were randomized as to whether they would completed review assignments under the journal's typical workflow or whether their first three reviews would involve a "mentoring" condition. Those in the mentoring condition reviewed manuscripts that were also assigned a reviewer from a pool of volunteer mentors with expertise relevant to the manuscript. This pool of mentors was recruited by inviting volunteers to this mentoring program from the top five percent of the journal's reviewer pool, as determined by prior reviewer internal scores given editors.[18]

In each of the first three reviews by the new reviewer, both the reviewer and a mentor assigned to the same manuscript would complete regular reviews. After both reviews were written, mentees were prompted to discuss their review with their mentor by email or telephone. Mentors were asked to give feedback about how well the mentee addressed the review, what they might have done differently, and how the review would rate on the journal's 5-point quality scale. Beyond these general suggestions, the specifics and amount of communication were left to the mentor and mentee.

Similar to Callaham et al. (2002), this study uses as its main outcome variable the editor-assigned 5-point quality rating given to each review. However, whereas Callaham et al. use the reviewer as the unit of analysis—averaging scores across reviews for a given reviewer—Houry et al. treat the individual review as the unit of analysis. Using the 490 completed reviews written by treated and control reviewers over the four-year study period, the authors find no statistically meaningful difference in average review quality, with mean scores of approximately 3.5 in the mentoring group (95% CI ≈ 3.2–3.9) and 3.4 in the control group (95% CI ≈ 3.2–3.6) on the 5-point scale. The authors also compare just the first three completed reviews and find identical average quality scores in treatment and control at 3.5 (treatment CI: 3.2–3.9; control CI: 3.1–3.9). The authors also motivate the idea there could be learning dynamics over subsequent reviews, fitting the editor's 5-point review evaluation scores to the count of the review performed by each reviewer (i.e., 1st, 2nd, 3rd review, etc.). They find no evidence of a non-zero average linear slope across all reviews, nor differences in slope from treatment versus control groups. The authors conclude there are no effects of the intervention.

The evidence presented in the paper are consistent with the conclusions, of course. Given the focus of this chapter, I contribute several additional checks. Analyzing individual reviews nicely

---

[18] Given a total reviewer pool of roughly 1,000 reviewers, this implies on the order of 50 eligible top-ranked reviewers; if, for example, one-third accepted the invitation, the resulting mentor pool would consist of roughly 15–20 volunteer mentors, though the paper does not report the actual number or acceptance rate.

contributes more observations than comparing individual reviewers, but effectively reweights estimates to those reviewers conducting most reviews—who might presumably be the better reviewers, less affected by further training. The dispersion in numbers of reviews per reviewer is no reported in the study, nor the correlation of review scores with numbers of reviews, so it is difficult to approximate or recover how this unit of analysis could have altered estimates. Another challenge that authors faced was that 4 reviewers reported not engaging with mentorship and all their reviews were dropped from the treatment group, 14% of reviewers in the treatment, leaving 24 observations in the treatment. There is perhaps no great fix for this problem, as—on the one hand—they simply did not receive the treatment. It introduces a kind of subtle selection problem, as ideally we would like to drop the identical matched counterparts from the control, at least as a kind of robustness test. Failing this, I roughly re-estimated results presuming these four reviewers contributed relatively large numbers of reviews with relatively high or low scores and found the differences across treatment and control remain non-statistical. Another question we might raise is whether a data set of that has fewer than 50 observations, if we count individual reviewers the analytically most relevant unit of analysis, can deliver meaningful bounds of results. A simple comparison of means for the reviewers (akin to those reported in Callaham et al. (2002)) would be helpful here. However, it seems substantively persuasive that the editors simply find neither learning nor differences (one presumes that none is reported even in casual observation), with such a strong intervention. Examining all outcomes reported in the study, I noted too that reviewers assigned to the mentoring condition appeared to receive more invitations, on average, 23 invitations versus 15 for the control group; and they also declined more 5 versus 2 in the control group. From the study reporting, it is not clear what accounts for these differences. In my own simulations based on guesstimations of the effects in these data, I cannot generate significant differences between treatment and control, consistent with the authors' inference of no effects.

Another possible reason for failing to detect effects of this seeming strong intervention is the use of a subjective 5-point score given by the editor. The meaningfulness of this measure might be tested more explicitly. For example, the underlying assumption that one's ability to generate high quality scores is a reflection of underlying more or less stable skills (that this intervention is attempting to affect). If this score is meaningful in this regard and indeed we have no learning and not differences from the intervention, we should find stable and significant differences across reviewers, if tested. Indeed, in a dedicated measurement study, Callaham et al. (1998) reported that editor quality ratings exhibited moderate inter-rater reliability and substantial, stable differences across reviewers.

Viewed through a Q–A–R–S lens, in principle this experiment should share many attributes with Callaham et al. (2002), and again attempts to shock the quality of reviewer skill at reviewing (R), aiming to leave manuscript attributes (Q), author attributes (A), and the peer review system (S) unchanged. We might understand that treatment and control reviewers tended to follow a similar procedure in matching to manuscripts and authors. However, beyond the usual concerns about significant variance generated by heterogeneity of reviewers and those reviewed, there appears to be significance endogeneously chosen assignment by editors, with likely large differences in numbers of invitations and assignments and, plausibly, assignment of best reviewers to most demanding manuscripts or other such issues. Further, the system of allocating counterpart reviewers, the volunteer "expert" reviewers, departed from usual systems. In a reviewer pool of roughly 1000, from which volunteers were drawn from the top 5 percent, it could be as few as 15 volunteer reviewers (i.e. 30% volunteer rate). On the one hand, this could stretch these reviewers to manuscripts for which they are less well suited. On the other hand, always comparing the treatment reviewers with the very best reviewers could create downward pressure in comparative editorial scores of the new reviewers. All this said, it is clear that this study fails to find evidence of differences in reviewer performance caused by quite significant mentorship interventions.

3.5.1   Other Training, Feedback and Guidance Interventions: More Null Effects

Beyond the experiments conducted at Annals of Emergency Medicine reviewed above, a small number of other interventions have been reported in other published studies. Schroter et al. (2004) conducted a randomized controlled trial at the BMJ in which reviewers were invited to attend a face-to-face training workshop focused on how to conduct high-quality peer review. Review quality was subsequently assessed using blinded editor ratings. Despite positive participant feedback and evidence of improved self-confidence, the study found no statistically detectable improvement in the quality of reviews. This result might be seen as particularly consistent with Callaham et al's (2004) finding that general feedback on reviews appears to have negligible effects, but here in a different context.

A separate line of work examines whether reviewers can be guided more effectively by providing explicit evaluative criteria or prompts. Cobo et al. (2011) tested whether asking reviewers to apply formal reporting guidelines during peer review improved the quality of final manuscripts. The intervention produced modest improvements in compliance with narrow specific reporting items.

Speich et al. (2023) conducted randomized interventions across hundreds of submissions at multiple journals, including those published by BMJ and PLOS, to test whether explicitly prompting

peer reviewers to check for basic, objective reporting information would improve the quality of published articles. In these trials, some reviewers—at the moment they agreed to review a manuscript—were randomly asked to verify whether the paper clearly reported a small set of essential details, such as how participants were assigned to groups, what outcomes were measured, how sample sizes were determined, and whether the study followed a pre-specified protocol. Despite the narrow and concrete nature of the prompt—and the fact that it focused on information that reviewers could, in principle, easily verify—asking reviewers to check these basic reporting elements did not produce statistically or substantively meaningful improvements in reporting.

3.5.2 *What Have We Learned form Training, Feedback and Guidance Interventions so Far? Limited Effects.*

Across a wide range of experimental designs, interventions that aim to improve peer review by training, mentoring, or guiding reviewers yield little evidence of meaningful effects. This conclusion holds across delivery modes—including written feedback (Callaham et al. 2002), in-person workshops (Schroter et al. 2004), multi-year mentoring programs (Houry et al. 2012), and checklist- or prompt-based guidance (Cobo et al. 2011; Speich et al. 2023)—as well as across journals, reviewer populations, and institutional settings. The consistency of null or near-null findings is striking.

The pattern is robust across outcome measures. Studies find no systematic improvements in editor-rated review quality, no consistent changes in review length or thoroughness, and at best modest, narrow gains on highly specific reporting items. Even interventions that are unusually intensive by journal standards—such as multi-year mentoring arrangements involving top-tier reviewers—fail to generate detectable improvements in subsequent review performance. While individual studies face familiar limitations (small samples, subjective outcomes, attrition, or imperfect compliance), the convergence of results across heterogeneous contexts and designs suggests that these null effects are not idiosyncratic.

Several explanations have been offered by the original investigators. Some emphasize implementation constraints: feedback is often delayed, generic, or weakly coupled to reviewers' decision processes (Callaham et al. 2002), while the level of sustained, individualized instruction required to improve analytic judgment may exceed what journals can realistically provide (Houry et al. 2012). Others point to reviewer burden: additional prompts or checklists may impose cognitive costs without changing behavior, particularly when reviewer comments are advisory rather than binding (Speich et al. 2023).

I speculate that a more fundamental interpretation is that the skills targeted by these interventions are not the binding constraint in peer evaluation. The core task of peer review is not clerical verification or procedural compliance, but judgment under deep uncertainty—assessing novelty, importance, plausibility, and potential impact when signals are noisy and long-run outcomes cannot be observed. If evaluative performance is primarily limited by the intrinsic difficulty of inference at the knowledge frontier, rather than by deficits in instruction or effort, then marginal training, feedback, or guidance should be expected to yield limited returns. Viewed through the structural regularities outlined in Section 2, this pattern is unsurprising. Structural Regularity 1 implies that peer review is an inferential task under fundamental uncertainty, not a rule-based exercise.

3.6 Manipulations of the Peer Review System Architecture (S)

Beyond who submits work and who reviews it, peer review is shaped by how the evaluation system itself is organized. A distinct class of experiments intervenes on the architecture of the review process—altering workflow, triage rules, information presentation, or aggregation mechanisms—with the aim of improving efficiency, reliability, or decision quality. Read through the Q–A–R–S framework, these studies operate squarely on the system component (S), typically leaving manuscripts (Q), author attributes (A), and reviewer identities (R) unchanged at the point of randomization.

This section reviews system-level interventions using a tiered approach. A small number of representative experiments are examined in detail to illustrate distinct modes of system design—such as workflow structure, editorial triage, aggregation of judgments, and information revelation—and to clarify what kinds of causal inferences these designs support. Because the range of system-level interventions studied in the literature is broad, a larger set of related experiments is discussed more briefly, grouped by design logic and intervention type. This combination of in-depth review and clustered synthesis is intended to convey both the methodological diversity of system-level experiments and the limits of what can be learned from any single design.

An Early Attempt at System-Level (S) Experiment: Neuhauser & Koran (1989)

Neuhauser and Koran's (1989) "Calling Medical Care Reviewers First: A Randomized Trial" is one of the earliest explicit randomized interventions on the mechanics of reviewer recruitment and editorial workflow. The study emerged from a practical operational debate within *Medical Care* and

would be carried out and published in the same journal.[19] As the authors report, the journal processed roughly 600 manuscripts per year and relied on the services of approximately 300 external reviewers. In the late 1980s, reviewer invitations and manuscript circulation were handled primarily by conventional editorial correspondence, making reviewer non-response costly in both administrative effort and turnaround time. Influential figures in health-services research cited by the authors had urged *Medical Care* to adopt a "call-first" policy, whereby editors would telephone potential reviewers to confirm willingness before sending manuscripts. The conjecture was that advance confirmation would reduce wasted mailings and speed review completion. The question was taken up by Neuhauser, a health-services researcher at Case Western Reserve University with a background in healthcare quality and operations improvement, and Koran, who was affiliated with the *Medical Care* editorial office and thus positioned to implement and observe changes in editorial workflow.

The intervention was embedded in routine journal operations during a finite period in 1986. For 95 manuscripts being sent out to review, editors selected two external reviewers (and alternates) in the usual way. One of the two initial reviewer invitations was randomly assigned to the "call-first" protocol and the other to the journal's standard "no-call" protocol. In the call-first condition, *Medical Care* staff telephoned the initially selected reviewer to confirm willingness before sending the manuscript; in the control condition, the manuscript was mailed without prior contact. Where a reviewer did not accept, subsequent reassignments were excluded from the analysis.

Of the 95 initial reviewer slots in the call-first protocol, editorial staff reported 18 instances in which prospective reviewers could not be reached and 33 explicit declinations; after contacting alternates as needed, 90 of the 95 slots ultimately yielded completed reviews. Under the standard protocol, the authors report 21 explicit declinations, but do not separately enumerate other forms of non-response to the initial invitation, making acceptance behavior at the first invitation stage difficult to compare across protocols. In particular, it is unclear whether additional non-responses in the control group should be interpreted as implicit declinations or as delayed responses resolved through reassignment.[20] The authors then focus on eventual fill rates at the level of the original reviewer slots,

---

[19] Health Services studies how healthcare is organized, delivered, and financed. Medical Care is a leading specialty journal in health services research, widely indexed and well established within its field at the time of the study's publication in 1989. Medical Care has maintained consistent visibility and standing within health services research over subsequent decades and is currently ranked in the top quartile of Health Care Sciences & Services journals with a recent impact factor ≈2.8.

[20] It would be informative to know whether advance outreach increased reviewer declinations, but this is not identifiable from the study design. Moreover, even if such a pattern were observed, it would be difficult to disentangle whether it reflected the telephone modality or the identity of the person making contact (editorial staff versus editor).

emphasizing that completion was statistically indistinguishable across protocols (approximately 95% under call-first versus 92% under standard practice). This effectively shifts the unit of analysis from the initial reviewers to the "slot" for a reviewer that might be filled. Interpreting slot-level fill rates is further complicated by the fact that they reflect editorial effort and timing choices: with sufficient persistence, most slots could eventually be filled, making completion an endogenous outcome rather than a fixed constraint.

The authors report that referee reports in the call-first condition were returned somewhat faster on average (32.8 days versus 37.5 days). Because this comparison conditions on completed reviews, it is descriptive rather than causal; the cleaner system-level estimand is total time required to fill a reviewer slot. On this margin, overall turnaround time was in fact longer under the call-first protocol (44.2 days) than under standard practice (37.7 days), despite similar eventual completion rates. Interpreting these differences is complicated by limited reporting and the endogeneity of slot completion to editorial persistence and timing decisions.

Taken together, the results suggest that advance phone calls may have increased observable declinations and lengthened total review cycles, contrary to prevailing conjecture at the time. However, the study was not designed to establish these effects definitively, nor to distinguish among competing mechanisms. Small sample size, incomplete reporting of non-response in the control condition, and the absence of formal statistical tests limit causal inference.

In relation to the Q–A–R–S lens, it would have been possible to draw comparisons between initially invited reviewers in both arms, as differences between author attributes (A) and submission attributes (Q) would have been held constant through randomization—if the number of observations were sufficient to account for variance and the influence of heterogeneity (not reported). System attributes (S) other than the call-first might have been kept roughly constant. There was some possible scope for variability in (S) across treatments if the recipients of invitations perceived invitations to come from different people (staff vs. editors) in the different treatment arms.

Editorial Triage as System Design: Process Efficiency and the Allocation of Evaluative Effort (Johnston et al. 2007)

In *"Early Editorial Manuscript Screening Versus Obligate Peer Review: A Randomized Trial,"* Johnston, Lowenstein, and Ferriero (2007) examine peer review as a problem of process design and effort allocation within *Annals of Neurology*—the same journal in which the experiment was conducted and subsequently published, and where the coauthors themselves constituted the newly appointed editorial

leadership team. At the time, the journal operated under a near-universal "obligate peer review" model, in which essentially every manuscript meeting most basic evaluation was sent to external referees. Initial assessments were conducted by associate editors—senior clinician–researchers (MDs and/or PhDs)—sometimes in consultation with other editors or the full editorial board. The paper does not describe a formal scoring rubric or checklist governing this assessment. This inherited workflow imposed substantial delays on authors and absorbed large amounts of reviewer effort, including for manuscripts that editors believed had little chance of acceptance. The central question motivating the study was whether reallocating evaluative effort upstream—by empowering editors to reject clearly unsuitable manuscripts before external review—could improve the efficiency of the review process without compromising the quality of published work.

Over a three-month period, all newly submitted manuscripts entering the journal (**N = 351**) were randomly assigned at submission to one of two editorial regimes: traditional (obligate) peer review or early editorial screening, with an intentionally unbalanced allocation placing approximately one quarter of submissions in the obligate-review arm (n = 88) and the remainder in the screening arm (n = 263). In effect, the journal routed the bulk of its submission throughput through the new early-screening regime during the experiment, while deliberately retaining a smaller subset of manuscripts under obligate peer review as a counterfactual. The study therefore resembles an in situ implementation of a revised editorial system with an experimental holdout, rather than a parallel trial layered onto an otherwise unchanged workflow—suggesting a degree of institutional commitment to the policy change prior to the completion of formal evaluation.

In both arms, each manuscript first received the same initial editorial assessment conducted by an associate editor—often in consultation with other editors—using the journal's customary criteria. This initial assessment was identical across conditions and reflected standard pre-experiment practice. Under the traditional (obligate) peer review regime, the associate editor's initial judgment was advisory only; desk rejection—available under prior practice—was deliberately suspended in this arm, rendering it more permissive than the journal's historical system. Under the early editorial screening regime, the same initial judgment became binding. Associate editors were aware that an experiment was underway but made initial screening decisions without knowledge of a manuscript's randomization status, as much as possible.

Associate editors initially recommended rejection for 69% of manuscripts in the early-screening arm (182 of 263) and 64% in the obligate-review arm (56 of 88); these recommendations were binding only in the early-screening arm, whereas all manuscripts in the obligate-review arm were

sent to external review by design. The paper reports that the average number of reviewers per submitted manuscript was 2.3 under obligate review and 0.7 under early screening, which implies nearly identical reviewer intensity conditional on review (approximately 2.26 reviewers per reviewed manuscript) across regimes. Final acceptance rates were three percentage points lower under early screening, at 9.5% (25 of 263), compared with 12.5% under obligate peer review (11 of 88)—a difference of approximately 24% in relative terms, though not statistically significant (p = 0.41).[21]

The authors also report that mean time from submission to final decision fell from 48 days under obligate peer reviewto 18 days under early editorial screening. This measure is unconditional on whether a manuscript was sent to review. Back-of-the-envelope calculations suggest that conditional on being sent to external review, decision times were nearly identical across regimes. For example, assuming desk rejections were processed within approximately five days, the implied review-cycle length is roughly 47–48 days in both arms, indicating that early screening primarily affects selection into review rather than the speed or intensity of review itself.

We might expect some of the study's efficiency-related outcomes to have been at least partially knowable without an experiment—for example, the baseline rate at which editors would have recommended rejection prior to the intervention (not reported), which plausibly approximates the two-thirds rejection rate observed here. On the margin, early screening could also have altered the review process if the subset of manuscripts sent to external review were systematically easier or harder to evaluate, altering review duration—however, the evidence suggests there is no difference here.

The central remaining question is whether reallocating evaluative effort altered the quality of review or editorial judgment itself. On this dimension, the available evidence is limited but uniformly points toward no detectable difference. Final acceptance rates differ numerically but are statistically indistinguishable given the precision of the estimates. Among accepted manuscripts, reviewer ratings of scientific impact and clinical impact, measured on a 1–4 scale, are also similar across conditions: mean scientific impact scores of 3.2 (SD = 0.6) under obligate review versus 3.1 (SD = 0.7) under early screening, and mean clinical impact scores of 3.1 (SD = 0.6) and 3.1 (SD = 0.5), respectively. While additional post hoc analyses—such as rereviewing manuscripts or examining longer-run citation

---

[21] Despite the seemingly large relative difference, the difference in acceptance rates is not statistically significant (p = 0.41). Although a more balanced allocation than the 75/25 split would, all else equal, have increased statistical power, simple calculations indicate it would not have altered this conclusion. Holding the observed rates fixed (9.5% vs. 12.5%) and total sample size constant (N = 351), an even split yields a z-statistic of approximately 0.9 (p ≈ 0.37). Detecting an absolute difference of three percentage points with 80% power at α = 0.05 would require on the order of 1,700 manuscripts per arm, far exceeding the study's scale. Adding well-selected model controls would be another way capture variation within a much smaller study.

outcomes—could in principle shed further light on epistemic effects, such evidence is not available in the study.

Viewed through the Q–A–R–S framework, this experiment constitutes a clean manipulation of S (system design). The intervention alters a single, core rule governing when peer review is mandatory versus when editorial judgment is decisive. By randomizing manuscripts to binding editorial screening or mandatory external review, the design aims to hold fixed reviewer attributes (R), author characteristics (A), and manuscript quality (Q) in expectation. The principal limitation is therefore not conceptual targeting but statistical precision. Short of substantially larger sample sizes, incorporating covariates capturing observable dimensions of R, A, and Q—such as reviewer expertise match, author experience, or manuscript type—could have improved efficiency of estimation and enabled exploration of heterogeneity. More broadly, the experiment highlights how system-level rule changes may interact with heterogeneity in authors, reviewers, and submissions, suggesting avenues for theory development and refinement of editorial processes beyond average treatment effects.

Aggregation and Deliberation: Fogelholm et al. (2012)

Fogelholm and colleagues (2012) examine a central structural component of grant peer review: whether panel discussion improves the reliability of evaluation beyond what can be achieved through aggregation of independent reviewer judgments. Working with the Academy of Finland, the authors study panel-based grant review in clinical medicine and epidemiology, asking whether the substantial time and cost devoted to panel meetings yields measurable gains in evaluative consistency relative to simpler, less resource-intensive procedures. This question is consequential for funding agencies, many of which justify deliberative panels on the grounds that discussion corrects idiosyncratic judgments and improves reliability.

The study exploits a rare experimental opportunity. Thirty reviewers were pair-matched by expertise and randomly allocated to two parallel expert panels, each charged with evaluating the same 65 grant proposals. Within each panel, proposals were first scored independently by two reviewers using a six-point ordinal scale. Panel members then convened, discussed each proposal, and produced a final consensus score. This design enables three distinct reliability comparisons: (a) agreement among individual reviewers within and across panels; (b) agreement between the two panels based on their post-discussion consensus scores; and (c) a constructed counterfactual in which inter-panel reliability is computed by mechanically averaging the pre-discussion reviewer scores within each panel, abstracting from deliberation rather than arising from a distinct experimental arm.

70

The quantitative results are striking. Inter-reviewer reliability was low, with large score disagreements (defined as differences of at least two points on the six-point scale) occurring in 40% of cases in one panel and 36% in the other. Panel discussion reduced—but did not eliminate—this dispersion: large inter-panel differences occurred for 26% of proposals when consensus scores were used. However, simple averaging of reviewer scores performed substantially better on this margin, reducing large inter-panel differences to 14%. In other words, mechanical aggregation cut the frequency of substantively large disagreements nearly in half relative to panel deliberation.

Reliability statistics tell a similar story. Inter-panel agreement, measured using Cohen's weighted kappa, was 0.23 (95% CI: 0.08–0.39) when panel consensus scores were used. When panel scores were replaced by the mean of the two reviewer scores, inter-panel reliability was numerically identical: kappa again equaled 0.23 (95% CI: 0.00–0.46). The identity of these estimates undercuts the core rationale for deliberation as a reliability-enhancing mechanism: discussion did not add incremental consistency beyond aggregation.

The authors further assess reliability using dichotomized funding classifications (scores 1–4 versus 5–6), a metric closely aligned with real funding decisions. Agreement across panels was higher when mean reviewer scores were used (69.2%) than when consensus scores were used (64.6%). Mean reviewer scores also achieved the highest positive predictive value for identifying proposals judged fundable by the other panel. Across continuous scores, large-difference rates, and binary classifications, aggregation consistently matched or outperformed panel consensus.

From a Q–A–R–S perspective, this study examines variation in system architecture (S) primarily through aggregation rules rather than through randomized exposure to deliberation itself. All proposals were discussed in both panels; there is no experimental arm in which deliberation was removed. The relevant estimand is therefore the incremental reliability gain from deliberation over and above aggregation, not the effect of deliberation per se. Because two independent panels evaluate the same proposals, reviewer composition (R) differs across panels, and inter-panel reliability reflects both reviewer heterogeneity and system design. The central outcome—agreement across panels—captures the stability of evaluation under alternative aggregation regimes.

The authors' own interpretation is notably pragmatic. Given that averaging independent reviewer scores performs at least as well as panel discussion on all reported reliability metrics, they conclude that funding agencies could rely more heavily on remote evaluation by multiple reviewers without convening physical panels, thereby reducing administrative costs without sacrificing reliability.

They also observe that panel consensus scores were, on average, lower than reviewer means, raising the possibility that discussion induces conservatism rather than information gain.

Overall, Fogelholm et al. provide unusually clean quantitative evidence that deliberative panels do not improve the reliability of grant evaluation beyond what is achieved by simple aggregation of independent judgments. The study's contribution lies not in demonstrating high reliability—levels remain modest—but in showing that the costly and institutionally entrenched practice of panel discussion delivers little incremental epistemic benefit on the dimension it is most often claimed to improve.

Information Revelation Among Reviewers and Converging Scores: Lane et al. (2022)

Lane et al. (2022), in "Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation," embedded two field experiments within grant-funding competitions at Harvard Medical School (HMS) to investigate whether sharing of evaluations would affect evaluations. More narrowly, they hypothesized that sharing scores (unblinding scores of committee members) would favor incremental, low-risk projects over novel, high-risk proposals. The research was operationalized by Lakhani, with experience running large-scale field experiments, with a capable team of three researchers working as postdocs at the time—Lane (Business), Teplitskiy (Sociology/Information Science), and Menietti (Economics)—along with three HMS administrators who provided essential institutional access: Guinan (HMS and Dana-Farber Cancer Institute), Gray (HMS), and Ranu (HMS). The study is notable for the simplicity and cleanliness of its experimental manipulation, the scale at which it is implemented, and the clarity of the behavioral response it reveals: reviewers overwhelmingly adjust their evaluations toward peer-provided reference scores when shown that they are out of step with others.

Study 1 involved a nationwide "ideation competition" for computational health solutions with $10,000 in total awards distributed across 12 winners, which attracted 47 proposals evaluated by 277 reviewers, with most reviewers evaluating just one proposal (mean = 1.5). Study 2 involved a microbiome research competition with $250,000 in total funding distributed as five awards of up to $50,000 each, which attracted 50 proposals evaluated by 92 reviewers who each assessed between 1 and 8 proposals (mean = 3.7). In both studies, reviewers independently scored proposals on multiple dimensions (feasibility, impact, innovation, investigator expertise, and overall scientific merit), with overall merit recorded on a bounded ordinal scale (1–8 in Study 1 and 1–9 in Study 2). After submitting these initial scores, reviewers were randomly shown fabricated peer scores for the overall merit

dimension only—presented as a single range (e.g., "2–5" or "7–9") to suggest multiple anonymous reviewers, with peer identities or individual scores not disclosed. Reviewers were then given the opportunity to revise any of their scores.[22]

The main analysis pools data from both interventions and finds that revealing scores led to strong and nearly universal convergence, with reviewers updating in the direction of the fabricated signal in 99.7% of cases. A small control group (34 reviewer–proposal pairs in Study 1 and 4 pairs in Study 2) saw their own scores displayed again and none made changes. Reviewers exposed to a lower fabricated signal decreased their own scores by 0.75 points on average, while reviewers exposed to a higher signal increased their scores by 0.43 points on average.

The results are interpreted by the authors as suggesting a mechanism through which information sharing may favor more conservative research, as highlighted in the title of the paper. Specifically, their interpretation proceeds in several steps. First, they note that people often allocate greater weight to negative information than to positive information, drawing on evidence from prior psychology research. Second, they observe that reviewers exposed to lower fabricated peer scores adjusted their own scores downward by a larger amount than reviewers exposed to higher fabricated scores adjusted upward (-0.75 versus $+0.43$ points) and treat this asymmetry as evidence that such a negativity bias is operating in peer evaluation. Third, they argue that novel or high-risk proposals are more likely to present identifiable weaknesses alongside potential strengths. Finally, combining these premises, they conclude that information-sharing formats may systematically disadvantage novel proposals by amplifying attention to weaknesses.

This work raises several interesting avenues for further study. Given the simplicity and strength of the experimental manipulation, the design naturally invites additional analyses that could help distinguish among alternative interpretations of the observed convergence. The hypothesis that novel proposals are specifically disadvantaged could be tested directly by stratifying low- and high-novelty proposals, making matched comparisons on the basis of scores in the control group, and determining whether the effect of negative fabricated signals is systematically larger for more novel submissions (e.g., $\Delta\text{Score} = \beta_1\cdot\text{Lower} + \beta_2\cdot\text{Higher} + \beta_3\cdot\text{Lower}\times\text{Novel} + \beta_4\cdot\text{Higher}\times\text{Novel} + \text{controls}$). Relatedly, because the peer scores shown to reviewers were fabricated, the estimated

---

[22] Study 2 used fixed, absolute score ranges ("1-3," "4-6," "7-9") assigned independently of reviewers' actual scores, while Study 1 constructed ranges relative to each reviewer's initial assessment (e.g., a reviewer scoring 5 might see "2-5" or "7-9"). Whether the fabricated range was "lower," "higher," or "neutral" was determined post-hoc in Study 2 by comparing the fixed range to what the reviewer had scored, making treatment assignment fully exogenous.]

treatment effects could be applied to the actual distribution of first-round reviewer scores to simulate how rank-ordering would have changed under alternative information-revelation regimes. By using this approach, simulated post-sharing scores could help assess whether information revelation systematically reshuffles proposals in ways that plausibly track quality, novelty, or neither. More generally, the design would also permit heterogeneity analyses by proposal novelty and simulations of rank reshuffling under alternative information-revelation regimes; such analyses would help sharpen the interpretation of whether convergence disproportionately disadvantages novel submissions.

Beyond these extensions, the results speak directly to the practical question of whether score revelation is constructive in this context. Even without any ability to observe the true quality of proposals, the patterns strongly suggest that this particular form of revelation produces compression rather than information aggregation. Score revelation led to near-universal convergence without any new signal being introduced. Evaluators predictably shifted their scores up or down in response to content-free reference points, with no opportunity to interact with peers, interrogate the basis of disagreement, or even know the identities of other reviewers. The authors report that lower peer scores prompted greater evaluative effort, reflected in longer and more detailed comments identifying weaknesses. This pattern, however, is also consistent with motivated reasoning: additional effort may serve to rationalize movement toward a salient reference point rather than to support genuine reassessment. Only 2 of 761 cases showed updating in the opposite direction of the fabricated signal.

We might also further probe into the idea of asymmetric effects from lower or higher peer reviews. Reviewer scores here are bounded, ordinal, and plausibly non-linear in their substantive meaning. A first point here is they are limited from above to be a maximum of 10—a mechanical asymmetry separate from any behavioral asymmetry. In one their robustness tests, the authors re-estimate their model on the subset of reviews with initial scores of 4, 5, or 6—where both upward and downward adjustments are mechanically feasible—and find that the earlier full-sample estimates of -0.75 (s.e. ≈ 0.06) versus $+0.43$ (s.e. ≈ 0.05), a difference of about 0.32 (s.e. ≈ 0.08), become -0.62 (s.e. ≈ 0.08) and $+0.56$ (s.e. ≈ 0.06) in the restricted sample, a difference of about 0.06 (s.e. ≈ 0.10).[23]

Viewed through the Q–A–R–S lens, Lane et al. (2022) implement one of the cleanest system-level manipulations in the peer-review literature. Q (proposal quality), A (applicant attributes), and (R) (reviewer characteristics) are held constant through randomization, yielding a within-proposal, within-

_____

[23] I estimate the standard errors of differences of coefficients (the "asymmetry") in both cases based on reported standard errors for the separate coefficients, reported numbers of observations, and assuming independence of the coefficient estimates.

reviewer design in which only the system-provided information varies. The estimated coefficients therefore identify causal effects of information revelation on reviewer behavior, revealing an overwhelming tendency toward convergence. What remains less clearly identified is how this convergence should be interpreted: the design distinguishes responsiveness to reference information, but does not by itself distinguish conservatism, novelty aversion, or bias from more general social-influence and anchoring mechanisms. Absent independent measures of proposal quality or a model specifying how reviewers should optimally update on peer information, the results are best read as evidence of how powerfully system design can reshape evaluative judgments—and of how careful interpretation is required when strong behavioral responses are observed under minimal informational change.

3.6.1   Other System-Level Experiments

The four studies reviewed above illustrate study of workflow choices (who gets reviewed and when), aggregation rules (how multiple judgments are combined), and information architecture (what reviewers are shown about others' evaluations). A broader collection of experiments, which have typically been smaller in scale, narrower in outcomes, or more explicitly operational in intent—have tested additional (S)ystem interventions. Here I offer a quick glimpse of several of these other investigations.

Workflow and Recruitment Protocols

A first cluster targets the "front-end plumbing" of peer review: how editors recruit reviewers and move manuscripts through the pipeline. Pitkin and Burmeister (2002), implemented at JAMA, randomize a simple recruitment protocol—asking reviewers' permission before sending a manuscript ("ask-first") versus simply sending the manuscript with a request ("just-send"). Like Neuhauser and Koran (1989), the intervention is operational rather than epistemic: it is designed to reduce time-to-review and wasted solicitation effort. Both studies highlight the limited leverage of isolated recruitment tweaks: while protocols can shift refusal rates or completion times at the margin, they do not touch how reviewers form judgments, and thus cannot speak to accuracy or bias.

Herbert et al. (2015) push the workflow margin further by comparing simplified grant-review procedures to standard, multi-stage evaluation in a major funding setting. Although their design is prospective rather than a tightly controlled within-proposal randomized trial, the core question is similar to Johnston et al. (2007): can system-level simplification and triage reproduce roughly the same funding decisions at lower cost? The broad message across these workflow interventions is that S-

changes can plausibly reduce reviewer and administrative burden—sometimes substantially—yet they rarely generate clean evidence about the epistemic properties of evaluation, because they alter who gets evaluated, under what criteria, and with what attention, all at once.

Viewed through Q–A–R–S, these studies are "pure S" interventions in intent: Q and A are background conditions determined by the submission stream; R is not manipulated but may shift endogenously because recruitment protocols and simplified procedures select different subsets of reviewers and different subsets of submissions into later stages. Their dependent variables are accordingly operational (acceptance and response rates, turnaround times, reviewer burden, and occasionally decision concordance), which makes them valuable for administrative design but intrinsically limited for identifying mechanisms of judgment.

Reporting Tools and Procedural Scaffolding

A second set of studies intervenes on (S) by providing tools, prompts, or structured checklists intended to improve the completeness of reporting in manuscripts, often with the hope of indirectly improving review quality or downstream scientific usability. Cobo et al. (2011) represent a canonical early example: manuscripts are randomized to conventional peer review versus conventional review plus an additional review structured around reporting guidelines, and manuscript reporting quality is scored in revised versions following review. Hopewell et al. (2016) evaluate a more technology-mediated variant through WebCONSORT, an author-facing, editor-requested web-based tool used at the revision stage intended to improve CONSORT adherence. Blanco et al. (2020) implement an editorial intervention at BMJ Open that provides authors with a structured editorial report explicitly flagging deficiencies in a core set of CONSORT reporting items, while Speich et al. (2023) test lightweight reminders to peer reviewers about reporting guideline items and find no evidence of improved reporting completeness in two randomized trials. Struthers et al. (2025) extend this family with GoodReports, examining whether customized article templates improve reporting completeness, but do not find reliable evidence of improvement, in part due to limited uptake and statistical power.

These interventions tend to produce a mixed but broadly consistent pattern: structured scaffolding can improve some specific reporting items or checklist-linked dimensions of completeness, but effects are typically modest, narrow, and uneven across items, with several low-cost reminder or template interventions producing null results. Conceptually, this is not surprising. Reporting tools and prompts operate through attention allocation and compliance rather than through

evaluative discrimination: they can help authors and reviewers notice or document certain elements, but they do not directly address harder evaluative tasks such as assessing novelty, plausibility, causal inference, or the validity of claims. In many cases, the outcome measure is itself downstream of multiple unobserved decisions—editorial rejection, author responsiveness, and reviewer uptake of the tool—so even when improvements are detected, attribution to any single mechanism remains ambiguous.

In Q–A–R–S terms, these are again predominantly S-interventions, but they often interact strongly with Q and with author behavior. Q is not held fixed (different manuscripts have different reporting deficits), authors differ in capacity and willingness to revise, and editorial decisions induce selection into what is ultimately scored. The "epistemic" interpretation of these experiments should therefore be modest: they provide evidence that procedural scaffolding can shift what gets reported, but they offer limited leverage on whether peer review becomes more accurate, less biased, or more predictive of downstream scientific value.

Information Exchange and Anchoring Variants

A third set of studies alters S by changing the information reviewers receive about other evaluations or about the evaluative frame itself. Das Sinha, Sahni, and Nundy (1999) test whether exchanging comments between Indian and non-Indian reviewers improves review quality, effectively adding a structured cross-review information channel. Liu et al. (2024) test for reviewer anchoring by experimentally manipulating reference information that reviewers see—an intervention conceptually adjacent to Lane et al. (2022), though with a different implementation and empirical setting.

These designs underscore an important general lesson: information architecture can produce large behavioral responses even when it adds little or no new signal about (Q). Lane et al. (2022) show near-universal updating toward fabricated peer signals; anchoring and exchange designs test closely related mechanisms—social influence, conformity, and reference dependence—that can compress or shift scores without necessarily improving accuracy. This is precisely the kind of effect that makes S-interventions attractive to administrators (because they move outcomes) and simultaneously dangerous (because movement need not reflect improved inference). Without independent measures of true quality—rare in this literature—information-sharing interventions are best interpreted as identifying behavioral regularities about evaluation under social information, rather than as improvements to peer review.

3.6.2 What Have We Learned from System-Level Interventions So Far? We are Only Beginning to Scratch the Surface with Such Experiments

System-level interventions are categorically different from the other experimental margins reviewed in Section 3 in both scope and identification logic. Whereas manipulations of author identity (A), reviewer attributes (R), or submission characteristics (Q) typically alter a well-defined input into an otherwise fixed review pipeline—and often induce bundled changes in correlated attributes—interventions on the review system (S) operate on the architecture of the organization itself. These include the division of labor between editors and reviewers, the sequencing of judgments, the information reviewers see and when they see it, the rules by which individual evaluations are aggregated, and the workflow that determines which submissions are reviewed at all. This is a high-dimensional design space, but importantly, many system features can be modified "locally" without mechanically being correlated with other attributes of the system others (unlike the case of attributes of A, R, or Q, see Section 2). As a result, (S) interventions more naturally permit experimental variation.

The existing experimental literature on system architecture should therefore be read less as a mapped body of knowledge than as a set of initial, tentative steps into a much wider frontier. Even so, several conclusions emerge. First, some S-interventions generate substantial operational gains without obvious changes or costs to evaluations being detected. Editorial triage and early screening (Johnston et al. 2007) sharply reduce reviewer burden and accelerate decisions, largely by reallocating evaluative effort upstream rather than by speeding review conditional on external evaluation. Recruitment-protocol tweaks (Neuhauser & Koran 1989; Pitkin & Burmeister 2002) affect administrative waste and cycle time at the margin, though their effects depend critically on how "completion" is defined and on endogenous editorial persistence.[24] These studies can also present new puzzles, as in Fogelholm et al. (2012) which finds that deliberative panel discussion does not improve inter-panel reliability beyond simple aggregation of independent scores, at least in that research context and condition, despite the substantial time and cost devoted to meetings.

The main point here, however, is not that we now "know what works." It is that system architecture is experimentally tractable and behaviorally consequential, yet massively underexplored.

---

[24] Of course, several other subsections in Section 3 might be viewed too as part of shifts in the peer review system.

3.7   Incentives and Motivation Manipulations within the Peer Review System (S)

Peer review relies on voluntary labor, making reviewer participation a natural target for incentive-based intervention. A small set of field experiments manipulates the motivational environment surrounding review—through monetary payments, messaging, and timing nudges—to test whether reviewers can be induced to accept assignments more readily or complete them more quickly. In Q–A–R–S terms, these studies intervene on system-level incentives (S), influencing participation and effort while attempting to hold manuscript quality (Q), author attributes (A), and reviewer expertise (R) balanced in expectation through randomization.

A Rare Study Motivated by Theory – Focused on "Nudges": Chetty et al. (2014)

Chetty, Saez, and Sándor (2014) embed a large-scale field experiment in the referee workflow of the Journal of Public Economics to study how alternative incentive and motivation mechanisms shape reviewer behavior. The study is largely motivated by the authors' pre-existing research agendas on incentives, attention, and prosocial effort—rather than by peer review per se—with peer review serving as a high-skill, repeated, and largely unpaid task in which competing motivation theories can be tested in the field. The authors held no editorial roles at the journal and exercised no discretion over manuscript handling; implementation instead relied on formal cooperation with the journal and Elsevier, with editors blinded to treatment assignment and standard editorial procedures preserved. Feasibility hinged on publisher-level integration of randomized incentives and the journal's reliance on a stable pool of repeatedly invited referees, enabling permanent treatment assignment and longitudinal observation. The study itself was published independently in the *Journal of Economic Perspectives*.

The experiment is explicitly theory-driven. The authors hypothesize that, in professional tasks governed largely by norms rather than contracts, delays reflect failures of attention and prioritization rather than fully rational responses to incentive magnitudes. They therefore design treatments that distinguish between effort shaped by salience and default timing (deadlines), explicit extrinsic incentives (monetary rewards), and social or reputational pressure (public observability). In this sense, the study is not merely an operational intervention in peer review, but a clean test of whether modest changes that shift salience and default timing can move behavior, and how such effects compare to price-based and reputational incentives.

The experiment included all referees invited by the *Journal of Public Economics* over approximately 20 months, encompassing 3,397 referee invitations sent to 2,061 distinct referees, of

which roughly two-thirds were accepted during the primary experimental period. Referees continued to be selected in the usual fashion by handling editors, following standard editorial procedures. Random assignment to one of four treatments occurred at the moment a referee was first invited during the experimental window, and that assignment persisted across subsequent review invitations.

Under the control condition, referees faced a six-week deadline (45 days) to submit a report. The authors' primary theoretical interest centered on a timing intervention—shortening the deadline to four weeks (28 days)—intended to isolate the role of attention and prioritization. Their logic is that peer review is a normatively expected, unpaid professional obligation: most referees are already willing to review, so variation in behavior should arise primarily along the timing margin rather than the participation margin. Accordingly, they predict that participation will be largely unaffected by changes in deadlines, while review timing responds to the greater salience of a shorter timeline.

The remaining treatments vary incentives asymmetrically across timelines. Among referees assigned to the four-week deadline, a subset were offered a $100 Amazon gift card conditional on meeting the deadline. Separately, among referees assigned to the original six-week deadline, a subset are exposed to a treatment in which the invitation email indicates that individual review turnaround times are publicly posted by name on the journal's website. While the resulting design is not fully factorial, it is deliberately structured to isolate three questions: whether deadlines alone affect behavior, whether monetary incentives add effects beyond shorter deadlines, and whether reputational pressure matters absent any change in deadlines.

The resulting design is not full factorial, but instead is structured to be able to answer three questions: Does changing the deadline itself alter behavior? Does adding monetary incentives generate effects beyond those induced by a shorter deadline? And does public observability of review times matter on its own, absent any change in the deadline? Further, the design allows causal estimation of effects on acceptance of review. and, conditional on participation, comparison of effects on review timing.

In the control group, 67.6% of invitations were accepted. Shortening the deadline from six weeks to four weeks leaves acceptance rates statistically unchanged (64.1%, p = 0.25), consistent with the authors' prediction. Adding a monetary incentive under the four-week deadline increases acceptance to 72.0% (p = 0.01 relative to the four-week deadline alone), while mentioning that review times would be posted reduces acceptance to 61.1% (p = 0.045).

Having established that changing deadlines does not affect acceptance rates, the authors turn to review timing. In the six-week control group, the median review is submitted after 47.8 days; this

falls to 35.5 days under the four-week deadline, a reduction of 12.3 days. Adding a $100 incentive further reduces the median to 27.2 days, while the treatment indicating that review times will be posted produces only modest acceleration. Because review times are observed only for accepted invitations, these estimates could in principle reflect selection as well as behavioral responses. The authors address this concern by reweighting observations using detailed pre-experiment review histories to equalize pre-experiment reviewer speed across treatment arms, leaving the estimated timing effects largely little changed.

The authors interpret the contrast between unchanged acceptance rates and sharply accelerated completion as evidence that deadlines operate primarily through salience and default timing rather than explicit incentive tradeoffs. In support of this interpretation, they show no association between treatment status and refereeing behavior at other Elsevier journals, which they emphasize as indicating that the intervention shifts attention and prioritization rather than reallocating effort across commitments.

From a Q–A–R–S perspective, the study intervenes cleanly on system design (S) at the point of the review decision, while balancing reviewer characteristics (R), manuscript attributes and quality (Q), and author attributes (A) through randomization. For outcomes conditional on acceptance—such as review timing—the authors appropriately supplement randomization with econometric adjustments to address selection. Relative to much of the peer-review literature, the study stands out for its scale, design discipline, and theoretical ambition.

Re-probing Monetary Payments for Review Acceptance and Completion: Cotton et al. (2025)

Akin to Chetty et al. (2014), Cotton et al. (2025) studied the effect of monetary incentives on reviewer recruitment and review times in an intervention embedded. They embedded an intervention at *Critical Care Medicine*, the journal where the study would publish, to investigate how cash payments affected reviewer recruiting and review-cycle times. Unlike Chetty, Saez, and Sándor (2014), which was motivated by theory testing of "nudges" where monetary incentives were used as a contrast, Cotton and coauthors were motivated an internal effort by journal leadership to address reviewer shortages following the COVID-era surge in submissions and heightened difficulty recruiting qualified reviewers. The author team included the journal's Managing Editor (Tosta) and senior editorial leadership (Buchman and Maslove), alongside economists with considerable experience studying incentives (Cotton and Alam, both at Queen's University, where Maslove was also based).

The experiment ran within the live editorial workflow over a six-month period from September 2023. All 715 invitations sent to 595 unique reviewers across 131 different manuscripts were included. Under the control condition, invitations proceeded as usual without cash incentives. In the cash incentive condition, reviewer invitation emails included an offer of $250 contingent on completion of the review—not tied to a specific deadline—paid by check mailed to the reviewer after the journal received their report. Therefore, the magnitude of the incentive was larger than in Chetty et al., but the payment mechanism differed fundamentally: Cotton et al. paid for completion regardless of timing, whereas Chetty et al. paid only for meeting a four-week deadline. The payment format also differed: checks mailed after review receipt versus Amazon gift card codes emailed immediately upon meeting the deadline.

Rather than randomizing at the level of individual reviewers or invitations, the study alternates treatment status across two-week time blocks, reflecting practical implementation constraints acknowledged by the authors. This approach provides less leverage for assessing balance and heterogeneity than invitation- or manuscript-level randomization, as the unit of assignment is calendar time rather than reviewers. As a result, treatment and control blocks may differ along unobserved temporal dimensions, and individual reviewers may be exposed to different conditions across invitations.

Of the 715 invitations sent, 414 were issued during incentive blocks and 301 during control blocks, reflecting unequal invitation volume across time blocks. Acceptance rates did not differ significantly across conditions (52.7% versus 47.8%, p = 0.20), nor did time from invitation to acceptance, which was approximately one day in both groups.

The authors emphasize the key outcome of interest as the rate of completion. Of incentivized invitations 206 of 414 resulted in a submitted review, compared to 127 of 301 for the control 2-week blocks—a difference of 7.6 percentage points (p = 0.04). Rather than correcting for endogenous acceptance, Cotton et al. redefine the estimand: their primary outcome is the probability that an invitation results in a completed review, with all invitations included in the analysis. This approach sidesteps selection into acceptance and does not separately identify behavioral responses conditional on participation. The treatment group had 42.0% on-time completions (i.e., within 14 days of acceptance), compared to 32.2% in the control group, with the difference significant at p = 0.008.) Because this outcome is conditional on acceptance, it may reflect a combination of treatment effects and selection. The authors could not find a discernible difference in quality of reviews based on ratings available from editors' scoring.

The authors interpret their findings as showing that monetary incentives operate primarily on reviewer supply and timeliness, while leaving evaluative judgment unchanged. As they summarize, "*there was no discernible difference in review quality, as rated by the handling editors*," even as cash incentives led to "*a modest but statistically significant increase in the share of invited reviewers who complete reviews*" and slightly faster turnaround times. The authors argue that their results suggest incentives may be useful for managing capacity constraints, writing that monetary payments "*may be one strategy to sustain robust peer review systems in conditions of increased demand, as was seen during the recent global COVID-19 pandemic*," while cautiously pointing to possible limits or unintended consequences outside the scope of the study.

It remains possible that the true magnitude of the effects exceeds the "modest" estimates the authors cautiously report. The higher volume of invitations issued during incentive blocks, while not problematic per se, could reflect periods in which editors drew more deeply into the reviewer pool, potentially inviting more marginal or time-constrained reviewers who may be less responsive to incentives. If so, the reported estimates could understate effects under more typical recruiting conditions. More generally, while the authors mitigate the most obvious source of imbalance by excluding holiday periods, additional sensitivity checks—such as reporting results with and without excluded blocks, incorporating time trends or block-pair fixed effects, or presenting balance tables for reviewer, author, or submission characteristics—would further clarify whether balance is achieved in implementation.

More broadly, the study highlights opportunities for extending what can be learned about incentive design in peer review. By design, it evaluates a single incentive type—cash compensation—at a single level ($250), applied uniformly to review completion within one institutional context. This yields a clear and informative benchmark, but leaves open how responses might vary with incentive magnitude, structure, or reviewer characteristics. Modest within-study variation—such as testing alternative payment levels, linking incentives to timeliness rather than completion, or examining heterogeneous responses across reviewer types—could substantially deepen understanding and improve generalizability, while remaining operationally feasible.

3.7.1 *What Have We Learned from Incentive Interventions So Far? Impacts on Timing, Little Impact on Judgment*

Given the central role that incentives play in organizational theory, it is notable how little experimental evidence exists on incentive design in peer review. Peer review operates in an unusually stark incentive environment—high-skill, repeated effort undertaken largely without pay and under

anonymity—yet, to my knowledge, only two published field experiment interventions to have directly intervened on incentives in live review systems.

Taken together, the two incentive experiments reviewed here—Chetty, Saez, and Sándor (2014) and Cotton et al. (2025)—deliver unusually clean causal evidence on how incentives operate in live peer-review systems. Despite substantial differences in field, institutional context, and design, both studies converge on a clear pattern: incentives reliably affect reviewer participation and follow-through, but do not measurably alter core evaluative judgments.

On the participation and throughput margins, effects are consistent and economically meaningful. In Chetty et al. (2014), offering a $100 payment conditional on meeting a four-week deadline increased acceptance rates by roughly eight percentage points relative to an otherwise identical deadline condition. In Cotton et al. (2025), offering a $250 payment for review completion—regardless of timing—increased the probability that an invitation yielded a completed report by 7.6 percentage points. Although the institutional bottlenecks differ across settings, both studies show that monetary incentives expand reviewer supply and increase system throughput at the margin.

Incentives also affect timing, though magnitudes depend on baseline conditions and incentive structure. Chetty et al. document large accelerations in review completion—over 12 days from shortening deadlines alone and an additional eight days from deadline-contingent payment—consistent with salience and default-timing mechanisms. Cotton et al., operating in a setting with shorter baseline review cycles and a payment tied to completion rather than speed, find more modest timing effects. These differences underscore that incentive design and institutional context jointly shape behavioral responses.

At the same time, the evidence base remains thin and there might be any number of alternative explanations for observed patterns. Nonetheless, the existing evidence establishes a robust baseline: incentives are effective tools for managing capacity and delay in peer review, but they do not appear to reorient expert judgment itself.

4 SYNTHESIS & CONCLUSION

4.1 Emerging Patterns across Studies

The analytic synthesis provided of the research clarifies that the foregoing published literature of interventions in live peer review platforms is not in most cases geared for external validity and generalizability, nor cumulative advance—and most studies remain somewhat isolated "data points."

Nonetheless, across the entire set of prior embedded intervention studies of Section 3 suggests a several recurring patterns.

Claim 1: Asymmetry of Behavioral Responsiveness vs. Inelasticity of Core Evaluations. Across the disparate body of isolated interventions one seemingly common pattern emerging thus far is that institutional interventions in scientific review systems and procedures consistently produce clear changes in reviewer behavior, while generating limited—and in many cases no discernible—changes in core evaluations, themselves. The same qualitative asymmetry between behavioral responsiveness and evaluative inelasticity recurs with notable consistency across studies, settings, and designs.

This pattern is evident, for example, in incentive and timing interventions. In Chetty, Saez, and Sándor's randomized experiment at the *Journal of Public Economics* (Section 3.7), shortening deadlines and introducing monetary incentives accelerates review completion by large margins—reducing median turnaround time by roughly 25–30 percent of the baseline review cycle—without measurably changing acceptance rates, referee recommendations, or editor-assessed review quality. Cotton et al.'s subsequent replication in a medical journal context (Section 3.7) similarly finds that cash payments increase the probability that an invitation yields a completed review by approximately 7–8 percentage points and modestly improve timeliness, while leaving evaluative content unchanged. In both settings, incentives generate economically meaningful changes in participation and throughput, but do not reorient substantive evaluative judgment.

A closely related pattern appears in workflow and system-architecture interventions. Early editorial screening at *Annals of Neurology* (Johnston et al. 2007; Section 3.6) reallocates evaluative effort upstream, reducing average time to decision from roughly seven weeks to under three and lowering the number of external reviews solicited per submission, yet produces no detectable differences in acceptance rates or reviewer-rated impact among accepted manuscripts. Recruitment-protocol experiments (Neuhauser & Koran 1989; Pitkin & Burmeister 2002; Section 3.6) similarly affect refusal rates and administrative delay at the margin, but do not alter how reviewers assess scientific merit once engaged. Information-architecture interventions reinforce this pattern: designs that alter the information reviewers see or the timing at which it is revealed often induce large behavioral responses—including near-universal score updating in response to salient reference information—without introducing new information about submission quality or improving evaluative discrimination.

Training, feedback, and guidance interventions reinforce the same conclusion. Across multiple randomized trials embedded at *Annals of Emergency Medicine* (Callaham et al. 2002; Houry et al. 2012;

85

Section 3.5), written feedback, exemplar reviews, and multi-year mentoring programs fail to produce statistically distinguishable differences in editor-rated review quality, despite substantial effort. Where effects can be quantified, changes in review-quality scores are typically on the order of a few tenths of a point on a five-point scale and statistically indistinguishable from zero. Checklist-based and prompt-based guidance (Cobo et al. 2011; Speich et al. 2023; Section 3.5) occasionally improves compliance, but these gains are narrow and item-specific and appear somewhat orthogonal to overall evaluation or decision outcomes.

Read in light of the characterization developed in Section 2, this pattern is unsurprising. When true scientific quality is difficult to discern and complex, interventions that alter effort, timing, or procedural scaffolding may change behavior, while leaving substantive assessments largely constrained by more inelastic limits in perception and inference.

Claim 2: Core Evaluations Respond to Reviewer Identity and Expertise. The instances in which evaluations were seen to be highly responsive and elastic involved depended on who does the review and the knowledge they. This pattern was of first-order importance and emerged consistently across heterogeneous designs and settings and stands in sharp contrast to the limited movement in evaluative outcomes induced by incentives, training, checklists, blinding, or workflow reforms discussed above.

The pattern is visible in the variance recorded in re-evaluation and duplication studies (Section 3.1), which provide the earliest demonstrations of evaluator-driven variation, such as Cole et al.'s NSF audit and Pier et al.'s NIH-style reliability study. These show that nominally identical work can receive materially different scores, rankings, or decisions when evaluated by different reviewers. Moreover, essentially most every study on peer review that transparently reports descriptive statistics on evaluations reveals considerable variation and dispersion, particularly in individual evaluations.

Several reviewer-composition experiments (Section 3.2) indirectly show this pattern by showing how the add-a-reviewer designs (Arnau et al. 2003; Cobo et al. 2007; Rethlefsen et al. 2025) lead authors to respond quite differently, given the different reviews they receive. More decisive evidence comes from designs that incrementally vary reviewer expertise along a continuous margin and show resulting differences in evaluations; in Boudreau et al. (2016), small, randomized differences in intellectual distance between reviewers and submissions generate nearly one-point shifts in evaluation scores, large relative to within-proposal variance, and produce substantial rank reshuffling. These effects are homogeneous and directional and exceed the magnitude of most procedural or incentive-based interventions reviewed elsewhere in Section 3.

Crucially, each of these differences do not require stark mismatches between qualified and unqualified evaluators or fundamental differences across fields. Even among comparably trained and credible reviewers, incremental differences in training, epistemic position, or intellectual proximity lead to differences in reviews across this literature. Read in light of the characterization developed in Section 2, the focused dependence of evaluative elasticity on reviewer identity and knowledge is perhaps not surprising. When scientific contributions lie close to, at, or beyond the knowledge frontier and expertise is specialized and unevenly distributed, evaluation necessarily proceeds through inference using individuals' partial and heterogeneous abilities to discern information cues and make sense of them according to their idiosyncratic interpretative frameworks.

Claim 3: Scientific Evaluation More Reliable as "Filter" of Poor Submissions than "Ranker" of Contributions. A common critique of peer review—repeated across disciplines and decades—is that it is unreliable, noisy, or arbitrary, often citing high variance in reviewer scores or instability in outcomes. Much of this critique, however, rests on evidence drawn from highly selected samples of already competitive submissions, or from sensitivity of rank orderings near decision thresholds. When evaluation is conditioned on manuscripts that have passed multiple screens and cluster near the upper tail of quality, even modest disagreement can generate large apparent reversals in rank or funding. Read in this light, many canonical demonstrations of "unreliability" reflect the intrinsic difficulty of fine-grained discrimination under uncertainty, rather than a breakdown of evaluative signal or an outright failure of peer review. Even in these settings, however, this literature consistently finds non-trivial convergence in assessments: correlations across independent review exercises are typically positive, often moderate in magnitude (e.g., $\approx$0.5–0.7 in NSF re-review studies), and highly statistically significant—a degree of shared signal that is frequently obscured by emphasis on rank reversals or variance measures taken in isolation.

By contrast, when experimental designs focus on identifying submissions that fail to meet basic standards—or that do not plausibly constitute a meaningful scientific contribution—the literature shows substantially greater convergence in evaluative judgment. Re-evaluation studies, audit designs, and large-scale duplication exercises consistently find that reviewers and panels agree far more on which submissions are clearly uncompetitive than on how to rank those that are broadly acceptable. In NSF and NIH re-review studies (Cole et al. 1981; Pier et al. 2018), disagreement is concentrated near funding cutoffs, while proposals well below threshold are reliably identified as such. Limited evidence that explicitly stratifies evaluations by quality level likewise shows lower dispersion for low-quality submissions (e.g., Pleskac et al. 2025). Although these patterns are suggestive rather than

definitive—and may in part reflect mechanical features of scoring systems—they are nonetheless consistent with the broader regularities documented here.

Further evidence comes from error-injection experiments, which directly separate fine-grained diagnostic performance from coarse screening outcomes. In Baxt et al. (1998), reviewers at Annals of Emergency Medicine identified, on average, only a minority of the injected errors explicitly in their reports; nonetheless, only 15 of 203 returned reviews (≈7%) recommended acceptance, implying that the paper would almost certainly have been rejected under any reasonable aggregation rule. Similarly, in the BMJ experiment by Godlee, Gale, and Martyn (1998), reviewers explicitly mentioned only 1.7–2.1 of eight seeded errors on average, with no statistically significant differences across blinding or signing conditions—yet rejection recommendations were common. In the systematic duplication study by Cobey et al. (2020), which resubmitted a previously published Nature paper to 308 biomedical journals, only 4 journals (≈1–2%) accepted, while over 94% rejected the submission. Read together, these studies point not to a failure of screening, but to substantial robustness in identifying submissions that should not proceed. (Exceptional cases of "stings" revealing larger problems tended to identify such failures in fringe institutions not known for quality.)

Taken together, these outcomes appear to reflect structural features of knowledge and evaluation rather than institutional malfunction. As noted in Section 2, to the extent that contributions lie close to, at, or beyond the knowledge frontier, uncertainty about their validity, generality, and long-run significance is inherently high, and the relevant knowledge required for evaluative convergence may be scarce or unevenly distributed. Under these conditions, bounded rationality and epistemic uncertainty should play a central role: evaluators must rely on coarse signals, selective attention, and inferential shortcuts, including extrapolation beyond established knowledge and beyond their own areas of expertise. This situation should generate systematic disagreement and instability in fine-grained evaluation of high-quality contributions. By contrast, flawed submissions or those that do not plausibly constitute a contribution lie further from the knowledge frontier, where evaluation becomes less uncertain and less dependent on highly specialized judgment. In this region, reviewers can draw on shared disciplinary foundations, well-established standards of rigor, and widely agreed-upon norms of plausibility.

The experimental evidence reviewed here therefore supports a reframing of peer review: it functions more robustly—and more reliably—as a filter for non-contributions than as a high-resolution ranking mechanism for frontier science.

4.2 Directions for Experimentation, Improvement, and Innovation

Popular critiques of peer review often characterize scientific evaluation as slow, costly, unreliable, and biased (Section 1). Here, we revisit these concerns by drawing on the experimental literature reviewed in this chapter. By design, these studies are not well suited to calibrating absolute performance levels—a task often better addressed through careful descriptive measurement of review times, workloads, and costs.

Much of the existing experimental work is best understood as **diagnostic**, probing behavioral responses and sensitivities within existing systems rather than evaluating fully specified counterfactual designs, efficiency benchmarks, or explicit models of institutional innovation and change. Nonetheless, when interpreted collectively and through a common analytical framework, these studies are highly informative about the channels through which institutional innovation can operate and the relative elasticity of different dimensions of peer review to intervention and redesign.

Innovating Speed & Costs. Regarding claims that peer review is slow or costly, baseline speed and cost can, in principle, be assessed directly from observable facts—such as review cycle times, reviewer workloads, and administrative budgets—often more straightforwardly than through experimental intervention. It remains plausible that future experiments, combined with more developed theory, could exploit clearer counterfactuals to assess performance relative to an efficiency frontier. However, even though most existing studies are not explicitly designed to measure costs or delays, the cumulative experimental patterns and analytic synthesis provide substantial insight into the margins along which speed and cost appear most amenable to improvement.

The existing literature and its synthesis nonetheless indicate several channels through which speed and cost show clear promise for improvement. Across incentive, workflow, and system-architecture interventions, modest changes in institutional design regularly produce shifts in behavior, with non-trivial magnitudes of effects on participation, responsiveness, and throughput (Sections 3.6–3.7; Claim 1). This pattern alone suggests considerable scope for redesigning the system to make more effective use of scarce expert labor. Especially relevant here is that any institution or organization can be thought of as an extremely long list of decisions or practices that make up its functioning and procedures. Only a small subset of these dimensions has yet been experimentally probed (see especially Section 3.6).

Another encouraging finding pointing to scope for innovation in reducing cost and time to review is that changes in system design can occur without deterioration—and often with little change at all—in core evaluative judgments. Many interventions operate on system-level features (S)—such

as deadlines, assignment rules, and workflow structure—that can be varied largely independently of review quality, at least as a first-order approximation supported by the existing evidence and the interpretation of expertise and assessment as the result of largely inelastic and difficult-to-accumulate knowledge (Sections 2 and 4.1).

Another channel for seeking cost and time efficiencies relates to scientific evaluation as a "filter" of low-quality submissions (Section 4.1). The filtering function appears to be far more predictive and convergent than the careful discernment and ranking among high-quality contributions. Inasmuch as filtering relies to a greater degree on knowledge not close to the knowledge frontier, there may be ways of codifying established knowledge bases to carry out many of these tasks, including through automation and machine-based intelligence.

Innovating "Filtering" Reliability. Reliability can be understood as the extent to which evaluative judgments are stable, reproducible, and informative about true scientific quality. Because true quality is unobserved at the time of evaluation (Section 2), reliability cannot be directly measured, as cost or timing can. Consequently, the literature has approached reliability primarily by documenting disagreement or variance across evaluations while holding submissions fixed. As discussed in Sections 3.1 and 4.1, although such variance is pervasive, interpretations that treat it as evidence of an absence of evaluative signal are often overstated. Substantial scope remains to improve how reliability is conceptualized and calibrated, including by distinguishing among its sources (Q, A, R, S, and random error; Section 2) and by interpreting reliability under different evaluative tasks, beginning with filtering versus ranking (Section 4.1).

More importantly, the foregoing discussion provides clues about the responsiveness of reliability to interventions and innovation, beginning with peer review's role as a filter for errors and clearly unacceptable submissions (Section 4.1). Identifying basic methodological or conceptual flaws typically does not require knowledge at or beyond the scientific frontier, and instead draws on codifiable, widely shared standards of rigor and reporting. Studies examining checklists, reporting requirements, error-injection designs, and the involvement of specialized methodologists suggest that review systems are generally effective at *classifying* submissions with basic problems as unsuitable, even when referees fail to enumerate all specific errors (e.g., Baxt et al. 1998; Godlee et al. 1998; Cobo et al. 2011; Speich et al. 2023; Sections 3.2 and 3.5). At the same time, the evidence indicates that explicit reporting of basic errors is often incomplete, pointing to scope for improving filtering reliability through better structuring of attention and responsibility—potentially via codification, routinization, or allocating specialist labor rather than scarce expert attention.

Especially promising—and still underdeveloped—direction for improving reliability in this filtering function concerns the role of automation and machine-based artificial intelligence. Precisely because filtering relies to a greater degree on codifiable knowledge not necessarily at the frontier, many of its component tasks are natural candidates for systematic automation. Prior experimental interventions involving checklists, reporting prompts, error-injection designs, and specialist reviewers (Sections 3.2 and 3.5) can be reinterpreted as early, human-centered approximations to this logic: they show that attention to codifiable elements affects outcomes, but that relying on individual referees to supply this attention is unreliable and incomplete. From this perspective, the opportunity presented by AI is not simply to extend existing intervention strategies, but to push the experimental literature toward a more explicitly theoretical and cumulative mode of inquiry. By explicitly theorizing the structure of knowledge, the nature of information processing task, the division of labor between human and machine-based intelligence, a new generation of studies can be much more than isolated A/B tests in informing cumulative advance of our understanding in this area.

Innovating "Ranking" Reliability. Successful use of machine-based intelligence and automation in scientific evaluation could plausibly result in human peer review constituting a smaller share of the overall division of labor, particularly for screening and diagnostic tasks. This prospect, however, throws into sharper relief the distinct—and far more difficult—problem of peer review as a mechanism for ranking high-quality, acceptable contributions. Unlike filtering, ranking operates precisely where uncertainty is greatest: close to, at, or beyond the knowledge frontier, where relevant expertise is highly specialized, unevenly distributed, and difficult to substitute or scale. In this region, evaluation necessarily relies on inference rather than verification, and disagreement reflects underlying epistemic structure rather than procedural failure. The implication is not that ranking can be readily automated, but that innovation in this domain must proceed with fundamentally different expectations and design principles—ones that acknowledge the limits of codification, the inelasticity of expert judgment, and the intrinsic difficulty of discriminating among frontier contributions.

What is striking, however, is how little the existing experimental literature has engaged in explicit theorizing about knowledge, judgment, and aggregation in this setting. This is notable given that the core function of scientific evaluation is precisely to assess new knowledge on the basis of an existing—and necessarily incomplete—stock of knowledge. With limited exceptions, most studies have focused on documenting variance, disagreement, or sensitivity to procedural features, without a clear account of how evaluators form predictions about scientific contribution, how those predictions

should be combined, or which sources of information are admissible and how they ought to be processed.

At the same time, the literature already contains important clues about how progress might be made. Research that examines intellectual distance and evaluator perspective, studies that contrast independent scoring with committee interaction and deliberation, and evidence showing that aggregation can stabilize outcomes even when individual judgments diverge all point toward the importance of modeling knowledge structure and information combination, rather than treating evaluation as a black box or a mere generator of noise. Section 2 of this chapter provides an initial framework for doing so by characterizing peer review as an information-processing institution operating under uncertainty. Building on this foundation, future experiments that explicitly theorize the knowledge space, the mapping between expertise and evaluative tasks, and the aggregation of heterogeneous beliefs offer a more promising path toward improving ranking reliability than further procedural refinements alone.

Innovating (Reducing) Bias. Bias in peer review has been studied more extensively than any other dimension of peer evaluation, with blinding experiments constituting the largest experimental literature in this area (Section 3.4; Figure 3). Yet despite this concentration of effort—roughly 18 blinding studies versus approximately 6 studies in each other intervention category—the literature provides surprisingly little basis for normative conclusions about bias (Section 3.4.2).

As documented in Section 3.4, blinding studies consistently establish that author-identity information causally affects evaluation outcomes: when authorship is visible, scores and decisions differ by author characteristics. However, these effects admit multiple interpretations. Section 2.2.5 established that with unobservable true quality (Q), reviewers may rationally draw on author attributes as informational cues when forming expectations about scientific merit—a practice once viewed as a feature rather than a flaw of peer review (Merton 1973; Csiszar 2019). The observed correlations between author characteristics and blinding effects (Section 3.4) cannot distinguish whether identity cues reflect inappropriate favoritism, rational statistical inference under uncertainty, or changes in reviewer scrutiny and effort. Further, as clarified in Section 2.2.3, author attributes are inherently bundled and correlated with other attributes, meaning that even clean experimental variation in identity visibility does not isolate effects of individual characteristics.

More broadly, the bias literature illustrates the gap between causal identification and normative interpretation that runs throughout the current experimental evidence reviewed in this chapter. Until experimental designs explicitly link identity effects to efficiency outcomes—as Pleskac et al. begin to

do, or as non-experimental work has attempted through structural modeling (e.g., Li 2017)—strong claims about bias remain empirically unfounded, however prevalent the priors.

4.3 Conclusion

The experimental evidence reviewed in this chapter does not support the view—often implied in popular critiques and even within the scholarly literature—that peer review is highly ineffective or fundamentally broken (Section 1). Rather, the accumulated findings point to a functioning institution operating under binding constraints. Peer review reliably filters clearly inadequate submissions (Section 4.1, Claim 3), responds systematically to reviewer expertise and intellectual proximity (Section 4.1, Claim 2), and proves remarkably robust to many procedural interventions that alter behavior without degrading evaluative judgment (Section 4.1, Claim 1). Where peer review struggles—in fine-grained ranking of high-quality contributions near the knowledge frontier—the difficulties appear to reflect structural features of evaluating new knowledge under uncertainty (Section 2) rather than institutional dysfunction. The scope for improvement is substantial, particularly along dimensions of speed, cost, and the allocation of evaluative effort (Section 4.2), but innovation should proceed from realistic expectations about what peer review can and cannot accomplish. Drawing together the existing literature also pointed to opportunities to increase the inferential "yield" of interventions while minimizing costs and disruption in live review processes.

REFERENCES

Alam, M., Kim, N.A., Havey, J., Rademaker, A., Ratner, D., Tregre, B., West, D.P. and Coleman, W.P. (2011) 'Blinded vs. unblinded peer review of manuscripts submitted to a dermatology journal: a randomized multi-rater study', British Journal of Dermatology, 165(3), pp. 563-567.

Alberts, B., Kirschner, M.W., Tilghman, S. and Varmus, H. (2014). *Rescuing US biomedical research from its systemic flaws*. Science, 343(6176), pp. 1422–1425.

Arnau, C., Cobo, E., Ribera, J.M., Cardellach, F. and Selva-O'Callaghan, A. (2003) 'Effect of statistical review on manuscript quality in Medicina Clinica (Barcelona): a randomized study', Medicina Clinica, 121(18), pp. 690-694.

Baxt, W.G., Waeckerle, J.F., Berlin, J.A. and Callaham, M.L. (1998) 'Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance', Annals of Emergency Medicine, 32(3), pp. 310-317.

Blanco, D., Altman, D., Moher, D., Boutron, I., Kirkham, J.J., Cobo, E. and Gonzalez, J.A. (2020) 'Effect of an editorial intervention to improve the completeness of reporting of randomised trials: a randomised controlled trial', BMJ Open, 10(7), e036037.

Blank, R.M. (1991) 'The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review', American Economic Review, 81(5), pp. 1041-1067.

Bohannon, J. (2013) 'Who is afraid of peer review?', Science, 342(6154), pp. 60-65.

Bornmann, L. (2011) 'Scientific peer review', Annual Review of Information Science and Technology, 45(1), pp. 197–245.

Bornmann, L. (2011). Scientific peer review. Annual Review of Information Science and Technology, 45, pp. 199–245.

Boudreau, K.J. and Lakhani, K.R., 2016. Innovation experiments: Researching technical advance, knowledge production, and the design of supporting institutions. Innovation Policy and the Economy, 16(1), pp.135-167.

Boudreau, K.J., Guinan, E.C., Lakhani, K.R. and Riedl, C. (2016) 'Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science', Management Science, 62(10), pp. 2765-2783.

Bruce, R., Chauvin, A., Trinquart, L., Ravaud, P. and Boutron, I. (2016) 'Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis', BMC Medicine, 14(85), pp. 1-14.

Callaham, M.L. and Schriger, D.L. (2002) 'Effect of structured workshop training on subsequent performance of journal peer reviewers', Annals of Emergency Medicine, 40(3), pp. 323-328.

Callaham, M.L., Knopp, R.K. and Gallagher, E.J. (2002) 'Effect of written feedback by editors on quality of reviews: Two randomized trials', JAMA, 287(21), pp. 2781-2783.

Chetty, R., Saez, E. and Sandor, L. (2014) 'What policies increase prosocial behavior? An experiment with referees at the Journal of Public Economics', Journal of Economic Perspectives, 28(3), pp. 169-188.

Chi, M.T.H., Feltovich, P.J. and Glaser, R. (1981) 'Categorization and representation of physics problems by experts and novices', Cognitive Science, 5(2), pp. 121-152.

Cobey, K.D., Rice, D.B., Lalu, M.M., Abramowitz, D., Ahmadzai, N., Cunningham, H., Ayala, A.P., Raffoul, H., Khan, F., Shamseer, L. and Moher, D. (2020) 'Stress testing journals: A quasi-experimental study of rejection rates of a previously published paper', BMC Medicine, 18(1), p. 88.

Cobo, E., Cortes, J., Ribera, J.M., Cardellach, F., Selva-O'Callaghan, A., Kostov, B., Garcia, L. and Vilardell, M. (2011) 'Effect of using reporting guidelines during peer review on quality of final manuscripts: A cluster randomised trial', BMJ, 343, d6783.

Cobo, E., Selva-O'Callaghan, A., Ribera, J.M., Cardellach, F., Dominguez, R. and Vilardell, M. (2007) 'Statistical reviewers improve reporting in biomedical articles: A randomized trial', PLoS ONE, 2(3), e332.

Cole, S., Cole, J.R. and Simon, G.A. (1981) 'Chance and consensus in peer review', Science, 214(4523), pp. 881-886.

Cotton, C.S., Alam, A., Tosta, S., Buchman, T.G. and Maslove, D.M. (2025) 'Effect of monetary incentives on peer review acceptance and completion: A quasi-randomized interventional trial', Critical Care Medicine, 53(6), pp. e1181-e1189.

Csiszar, A. (2019) The Scientific Journal: Authorship and the Politics of Knowledge in the Nineteenth Century. Chicago: University of Chicago Press.

Das Sinha, S., Sahni, P. and Nundy, S. (1999) 'Does exchanging comments of Indian and non-Indian reviewers improve the quality of manuscript reviews?', National Medical Journal of India, 12(5), pp. 210-213.

DeGroot, M.H. (1970) Optimal Statistical Decisions. New York: McGraw-Hill.

Dewald, W.G., Thursby, J.G. and Anderson, R.G. (1986) 'Replication in empirical economics: The Journal of Money, Credit and Banking project', American Economic Review, 76(4), pp. 587-603.

Duflo, E., Glennerster, R. and Kremer, M. (2007), 'Using randomization in development economics research: A toolkit', in Schultz, T.P. and Strauss, J.A. (eds.), Handbook of Development Economics, Vol. 4, Amsterdam: Elsevier, pp. 3895–3962.

Ericsson, K.A. and Smith, J. (eds.) (1991) Toward a General Theory of Expertise: Prospects and Limits. Cambridge: Cambridge University Press.

Fisher, M., Friedman, S.B. and Strauss, B. (1994) 'The effects of blinding on acceptance of research papers by peer review', JAMA, 272(2), pp. 143-146.

Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A. and Vaananen, K. (2012) 'Panel discussion does not improve reliability of peer review for medical research grant proposals', Journal of Clinical Epidemiology, 65(1), pp. 47-52.

Fox, C.W., Meyer, J. and Aime, E. (2023) 'Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal', Functional Ecology, 37(5), pp. 1144-1157.

Gerber, A.S. and Green, D.P. (2012), *Field Experiments: Design, Analysis, and Interpretation*, New York: W.W. Norton & Company.

Godlee, F. (2012). *Making research more relevant, reproducible, and reliable*. British Medical Journal, 344, e4383.

Godlee, F., Gale, C.R. and Martyn, C.N. (1998) 'Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial', JAMA, 280(3), pp. 237-240.

Goldberg, A., Stelmakh, I., Cho, K., Oh, A., Agarwal, A., Belgrave, D. and Shah, N.B. (2025) 'Peer reviews of peer reviews: A randomized controlled trial and other experiments', PLoS ONE, 20(4), e0320444.

Harrison, G.W. and List, J.A. (2004), 'Field experiments', *Journal of Economic Literature*, 42(4), pp. 1009–1055.

Herbert, D.L., Barnett, A.G., Clarke, P. and Graves, N. (2015) 'Using simplified peer review processes to fund research: A prospective study', BMJ Open, 5(7), e008380.

Hirschauer, S. (2010) 'Editorial judgments: A praxeology of "voting" in peer review', Social Studies of Science, 40(1), pp. 71–103.

Hopewell, S., Boutron, I., Altman, D.G., Barbour, G., Moher, D., Montori, V., Schriger, D., Cook, J., Gerry, S., Omar, O. and Dutton, P. (2016) 'Impact of a web-based tool (WebCONSORT) to improve the reporting of randomised trials: Results of a randomised controlled trial', BMC Medicine, 14(1), p. 199.

Houry, D., Green, S. and Callaham, M. (2012) 'Does mentoring new peer reviewers improve review quality? A randomized trial', BMC Medical Education, 12, p. 83.

Ioannidis, J.P.A. (2005) 'Why most published research findings are false', PLoS Medicine, 2(8), e124.

Jefferson, T., Alderson, P., Wager, E. and Davidoff, F. (2002) 'Effects of editorial peer review: A systematic review', JAMA, 287(21), pp. 2784–2786.

John, L.K., Loewenstein, G., Marder, A. and Callaham, M.L. (2019) 'Effect of revealing authors conflicts of interests in peer review: Randomized controlled trial', BMJ, 367, l5896.

Johnston, S.C., Lowenstein, D.H. and Ferriero, D.M. (2007) 'Early editorial manuscript screening versus obligate peer review: A randomized trial', Annals of Neurology, 61(4), pp. A10-A12.

Jones, B.F. (2009) 'The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?', Review of Economic Studies, 76(1), pp. 283–317.

Justice, A.C., Cho, M.K., Winker, M.A., Berlin, J.A. and Rennie, D. (1998) 'Does masking author identity improve peer review quality? A randomized controlled trial', JAMA, 280(3), pp. 240-242.

Keeney, R.L. and Raiffa, H. (1976) Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: Wiley.

Keynes, J.M. (1937) 'The general theory of employment', Quarterly Journal of Economics, 51(2), pp. 209–223.

Knight, F.H. (1921) Risk, Uncertainty, and Profit. Boston: Houghton Mifflin.

Kuhn, T.S. (1962) The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Lamont, M. (2009) How Professors Think: Inside the Curious World of Academic Judgment. Cambridge, MA: Harvard University Press.

Lane, J.N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E.C. and Lakhani, K.R. (2022) 'Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation', Management Science, 68(6), pp. 4478-4495.

Leamer, E.E. (1983) 'Let's take the con out of econometrics', American Economic Review, 73(1), pp. 31-43.

Li, D. (2017) 'Expertise versus bias in evaluation: Evidence from the NIH', American Economic Journal: Applied Economics, 9(2), pp. 60–92.

Liu, R., Shah, N.B. and Fiez, T. (2024) 'Testing for reviewer anchoring in peer review: A randomized controlled trial', PLoS ONE, 19(3), e0301111.

Mahoney, M.J. (1977) 'Publication prejudices: An experimental study of confirmatory bias in the peer review system', Cognitive Therapy and Research, 1(2), pp. 161-175.

Mayo, N.E., Brophy, J., Goldberg, M.S., Klein, M.B., Miller, S., Platt, R.W. and Bhupsingh, J. (2006) 'Peering at peer review revealed high degree of chance associated with funding of grant applications', Journal of Clinical Epidemiology, 59(8), pp. 842-848.

McNutt, R.A., Evans, A.T., Fletcher, R.H. and Fletcher, S.W. (1990) 'The effects of blinding on the quality of peer review: A randomized trial', JAMA, 263(10), pp. 1371-1376.

Merton, R.K. (1973) The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press.

National Academies of Sciences, Engineering, and Medicine (2018). Peer Review in Scientific Publishing: Procedures, Policy, and Promise. Washington, DC: National Academies Press.

Neuhauser, D. and Koran, C.J. (1989) 'Calling medical care reviewers first: A randomized trial', Medical Care, 27(6), pp. 664-666.

Okike, K., Hug, K.T., Kocher, M.S. and Leopold, S.S. (2016) 'Single-blind vs double-blind peer review in the setting of author prestige', JAMA, 316(12), pp. 1315-1316.

Peters, D.P. and Ceci, S.J. (1982) 'Peer-review practices of psychological journals: The fate of published articles, submitted again', Behavioral and Brain Sciences, 5(2), pp. 187-255.

Pier, E.L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M.J., Ford, C.E. and Carnes, M. (2018) 'Low agreement among reviewers evaluating the same NIH grant applications', Proceedings of the National Academy of Sciences, 115(12), pp. 2952-2957.

Pitkin, R.M. and Burmeister, L.F. (2002) 'Identifying manuscript reviewers: Randomized comparison of asking first or just sending', JAMA, 287(21), pp. 2795-2796.

Pleskac, T.J., Kyung, E.J., Chapman, G.B. and Urminsky, O. (2025) 'Blinded versus unblinded review: A field study on the equity of peer-review processes', Management Science (ahead of print).

Polanyi, M. (1966) The Tacit Dimension. London: Routledge & Kegan Paul.

Popper, K.R. (1959) The Logic of Scientific Discovery. London: Hutchinson.

Rethlefsen, M.L., Schroter, S., Bouter, L.M., Kirkham, J.J., Moher, D., Ayala, A.P., Blanco, D., Brigham, T.J., Grossetta Nardini, H.K., Kirtley, S., Nyhan, K., Townsend, W. and Zeegers, M. (2025) 'Improving peer review of systematic reviews and related review types by involving librarians and information specialists as methodological peer reviewers: A randomized controlled trial', BMJ Evidence-Based Medicine, 30(4), pp. 241–249.

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F. and Smith, R. (2004) 'Effects of training on quality of peer review: Randomised controlled trial', BMJ, 328(7441), p. 673.

Shah, N.B. (2022) 'Challenges, experiments, and computational solutions in peer review', Communications of the ACM, 65(6), pp. 76–87.

Shapin, S. (1994) A Social History of Truth: Civility and Science in Seventeenth-Century England. Chicago: University of Chicago Press.

Smith, R. (1999). *Opening up BMJ peer review.* British Medical Journal, 318, pp. 4–5.

Smith, R. (2006). *Peer review: a flawed process at the heart of science and journals.* Journal of the Royal Society of Medicine, 99(4), pp. 178–182.

Squazzoni, F., Bravo, G., Farjam, M., Marušić, A., Mehmani, B., Willis, M. and Birukou, A. (2021). Peer review and competition for resources. Research Policy, 50(9), 104292.

Van Rooyen, S., Delamothe, T. and Evans, S.J.W. (2010) 'Effect on peer review of telling reviewers that their signed reviews might be posted on the web: Randomised controlled trial', BMJ, 341, c5729.

Van Rooyen, S., Godlee, F., Evans, S., Black, N. and Smith, R. (1999) 'Effect of open peer review on quality of reviews and on reviewers recommendations: A randomised trial', BMJ, 318(7175), pp. 23-27.

van Rooyen, S., Godlee, F., Evans, S., Black, N. and Smith, R. (1998) 'Effect of blinding and unmasking on the quality of peer review: A randomized trial', JAMA, 280(3), pp. 234-237.

Vinther, S., Nielsen, O.H. and Rosenberg, J. (2012) 'Same review quality in open versus blinded peer review in Ugeskrift for Laeger', Danish Medical Journal, 59(8), A4479.

von Winterfeldt, D. and Edwards, W. (1986) Decision Analysis and Behavioral Research. Cambridge: Cambridge University Press.

Walsh, E., Rooney, M. and Appleby, L. (2000) 'Open peer review: A randomised controlled trial', British Journal of Psychiatry, 176(1), pp. 47-51.

TABLES
Table 1 List of Published Experiments (1977-2025)

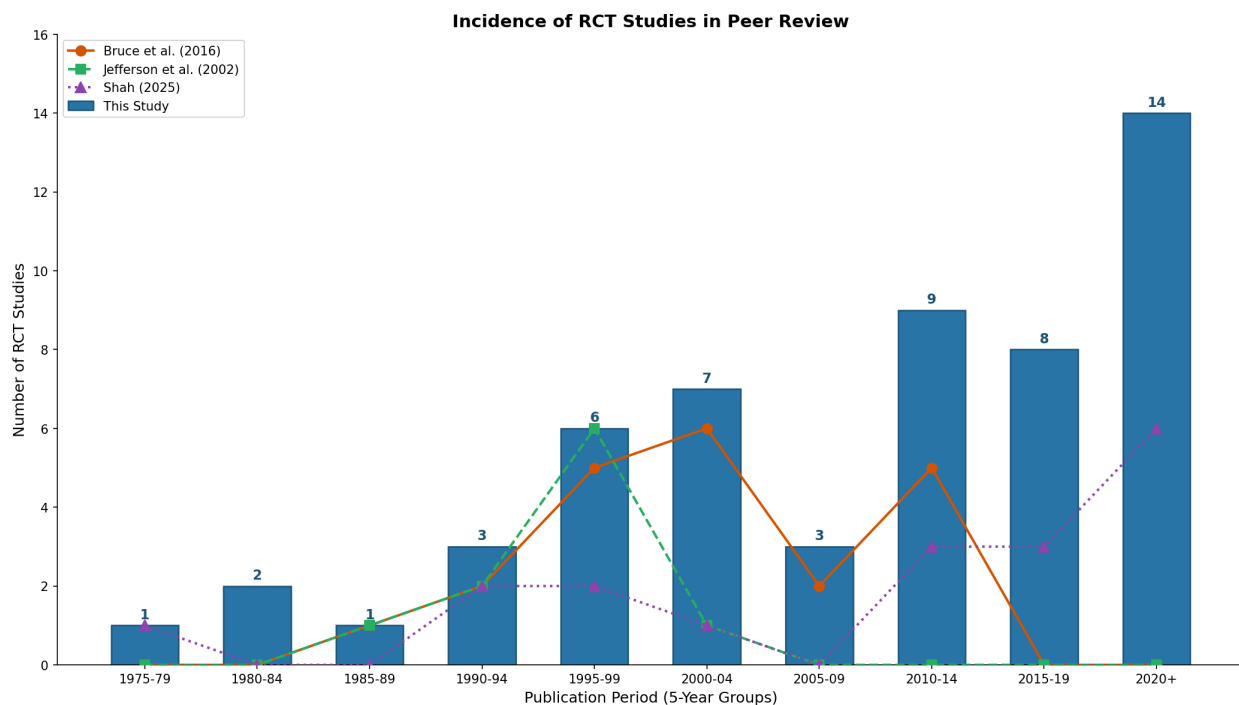| Primary Section within Chapter	Year	Study	Citations	Page Count	Journal
3.1 Re-Evaluation (Re-Submission) Studies	1981	Cole et al. (1981)	879	6	*Science*
3.1 Re-Evaluation (Re-Submission) Studies	1982	Peters and Ceci (1982)	1450	9	*Behavioral and Brain Sciences*
3.1 Re-Evaluation (Re-Submission) Studies	2006	Mayo et al. (2006)	92	7	*Journal of Clinical Epidemiology*
3.1 Re-Evaluation (Re-Submission) Studies	2018	Pier et al. (2018)	229	5	*PNAS*
3.2 Audit and Sting Studies	1977	Mahoney (1977)	1486	15	*Cognitive Therapy and Research*
3.2 Audit and Sting Studies	1998	Baxt et al. (1998)	186	8	*Annals of Emergency Medicine*
3.2 Audit and Sting Studies	1998	Godlee et al. (1998)	474	4	*JAMA*
3.2 Audit and Sting Studies	2013	Bohannon (2013)	1678	6	*Science*
3.2 Audit and Sting Studies	2020	Cobey et al. (2020)	8	10	*BMC Medicine*
3.2 Audit and Sting Studies	2010	Emerson et al. (2010)	230	6	*Archives of Internal Medicine*
3.2 Audit and Sting Studies	2000	Resch et al. (2000)	85	4	*Journal of the Royal Society of Medicine*
3.3 Reviewer Composition Studies	2003	Arnau et al. (2003)	10	5	*Medicina Clínica*
3.3 Reviewer Composition Studies	2007	Cobo et al. (2007)	125	9	*PLoS ONE*
3.3 Reviewer Composition Studies	2016	Boudreau et al. (2016)	633	19	*Management Science*
3.3 Reviewer Composition Studies	2021	Rethlefsen et al. (2021)	12	9	*BMJ Evidence-Based Medicine*
3.4 Author Blinding Studies	1990	McNutt et al. (1990)	472	6	*JAMA*
3.4 Author Blinding Studies	1991	Blank (1991)	668	27	*American Economic Review*
3.4 Author Blinding Studies	1994	Fisher et al. (1994)	228	4	*JAMA*
3.4 Author Blinding Studies	1998	van Rooyen et al. (1998)	373	4	*JAMA*
3.4 Author Blinding Studies	1998	Justice et al. (1998)	354	3	*JAMA*
3.4 Author Blinding Studies	1999	Van Rooyen et al. (1999)	459	5	*BMJ*
3.4 Author Blinding Studies	2000	Walsh et al. (2000)	319	5	*British Journal of Psychiatry*
3.4 Author Blinding Studies	2010	Van Rooyen et al. (2010)	191	7	*BMJ*
3.4 Author Blinding Studies	2011	Alam et al. (2011)	65	5	*British Journal of Dermatology*
3.4 Author Blinding Studies	2012	Vinther et al. (2012)	39	4	*Danish Medical Journal*
3.4 Author Blinding Studies	2016	Clarke et al. (2016)	21	5	*Journal of Clinical Epidemiology*
3.4 Author Blinding Studies	2016	Okike et al. (2016)	194	2	*JAMA*
3.4 Author Blinding Studies	2017	Tomkins et al. (2017)	614	6	*PNAS*
3.4 Author Blinding Studies	2019	John et al. (2019)	37	8	*BMJ*
3.4 Author Blinding Studies	2021	Nakamura et al. (2021)	25	15	*eLife*
3.4 Author Blinding Studies	2022	Huber et al. (2022)	164		*PNAS*
3.4 Author Blinding Studies	2023	Fox et al. (2023)	98	12	*Functional Ecology*
3.4 Author Blinding Studies	2025	Pleskac et al. (2025)	0	20	*Management Science*
3.5 Reviewer Training, Feedback, and Checklists	2002	Callaham et al. (2002)	103	3	*JAMA*
3.5 Reviewer Training, Feedback, and Checklists	2002	Callaham and Schriger (2002)	55	6	*Annals of Emergency Medicine*
3.5 Reviewer Training, Feedback, and Checklists	2004	Schroter et al. (2004)	339	6	*BMJ*
3.5 Reviewer Training, Feedback, and Checklists	2012	Houry et al. (2012)	91	7	*BMC Medical Education*
3.5 Reviewer Training, Feedback, and Checklists	2023	Speich et al. (2023)	28	14	*JAMA Network Open*
3.5 Reviewer Training, Feedback, and Checklists	2025	Goldberg et al. (2025)	31	15	*PLOS ONE*
3.6 Manipulations of the Peer Review System	1989	Neuhauser and Koran (1989)	6	3	*Medical Care*
3.6 Manipulations of the Peer Review System	1999	Das Sinha et al. (1999)	19	4	*National Medical Journal of India*
3.6 Manipulations of the Peer Review System	2002	Pitkin and Burmeister (2002)	29	2	*JAMA*
3.6 Manipulations of the Peer Review System	2007	Johnston et al. (2007)	28	3	*Annals of Neurology*
3.6 Manipulations of the Peer Review System	2011	Cobo et al. (2011)	296	8	*BMJ*
3.6 Manipulations of the Peer Review System	2012	Fogelholm et al. (2012)	96	6	*Journal of Clinical Epidemiology*
3.6 Manipulations of the Peer Review System	2015	Herbert et al. (2015)	31	8	*BMJ Open*
3.6 Manipulations of the Peer Review System	2016	Hopewell et al. (2016)	60	11	*BioMed Central Medicine,*
3.6 Manipulations of the Peer Review System	2020	Blanco et al. (2020)	43	9	*BMJ Open*
3.6 Manipulations of the Peer Review System	2021	Lane et al. (2021)	74	18	*Management Science*
3.6 Manipulations of the Peer Review System	2024	Liu et al. (2024)	8	18	*PLOS ONE*
3.6 Manipulations of the Peer Review System	2025	Struthers et al. (2025)	5	12	*Research Integrity and Peer Review*
3.6 Manipulations of the Peer Review System	2023	Stelmakh et al. (2023)	15	20	*PLOS ONE*
3.7 Incentives and Motivation Manipulations	2014	Chetty et al. (2014)	146	20	*Journal of Public Economics*
3.7 Incentives and Motivation Manipulations	2025	Cotton et al. (2025)	10	8	*Critical Care Medicine*

Figure 1 Published Studies Reporting Experiments Embedded in Scientific

Evaluation Processes

Note. Bars indicate the number of published studies included in this chapter, grouped by five-year publication periods, reflecting my attempts to identify RCTs and controlled comparisons embedded in live peer review processes that have been published in academic journals to year end 2025. Connected markers show numbers of studies appearing in three prior reviews that fit the same set of criteria. Jefferson et al. (2002) identify19 comparative studies "in which some attempt to control for confounding had been made." Of these, 10 meet my criteria of being an embedded RCT. Bruce et al. (2016) identify 21 studies reporting embedded RCTs in published in academic journals. Shah (2022), published in Communications of the ACM, surveys experiments in peer review with a particular emphasis on computer science conferences. Of studies discussed, 2 meet the selection criteria here of being experiments published in journals rather than in conference proceedings or working papers. An extended online version of the study (Shah, 2025) covers additional work, of which 18 meet the selection criteria here.
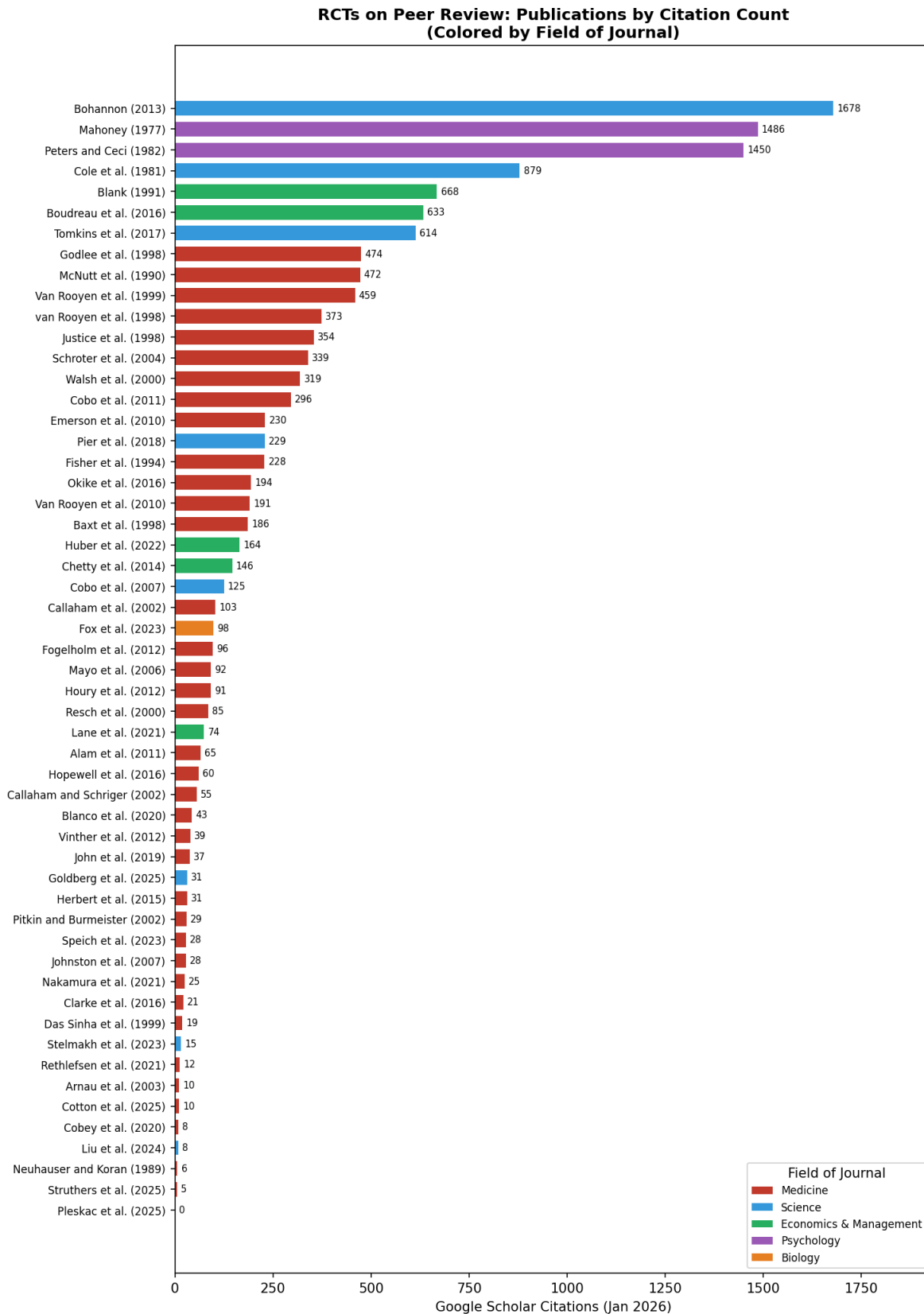
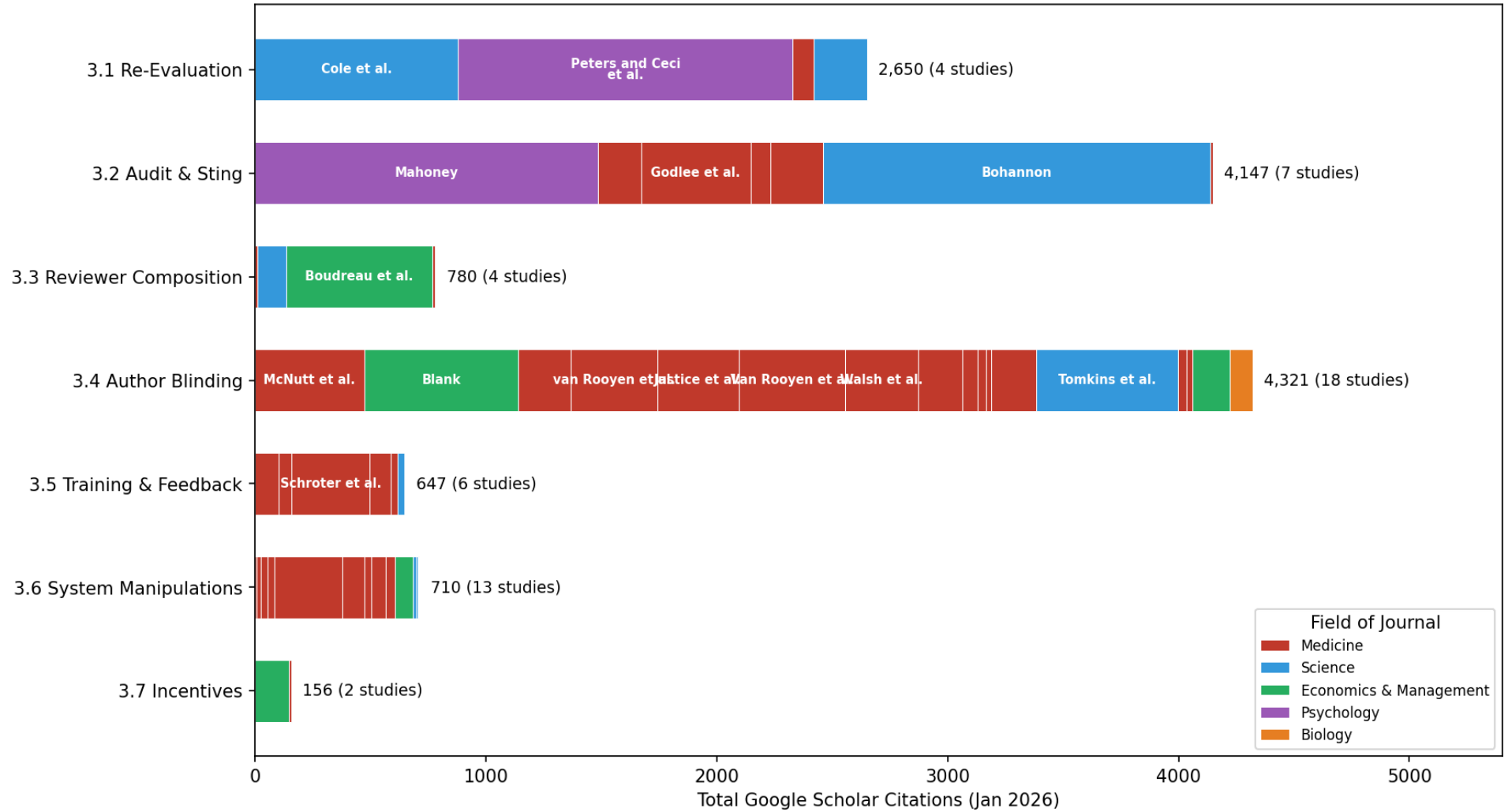**Figure 2 Experimental Studies by Citations**

Figure 3 Studies and Citations by Study Type

Note. In certain cases, alternative categorizations were plausible. Where there was discretion, the categorizations shown here and in Table 1 prioritized exposition purposes.