SCIENTIFIC PEER REVIEW: INSTITUTIONAL WORKINGS AND PERFORMANCE

Kevin Boudreau*

Northeastern & NBER

Peer review stands at the center of modern science, certifying knowledge and allocating resources, yet it emerged only in the twentieth century and continues to evolve as an institution. This chapter begins with a detailed institutional review of peer review's history, organization, and performance across both journals and grant agencies, drawing together evidence on how the system actually works and what existing research reveals about its effectiveness.

Keywords: peer review, scientific evaluation, institutional design, science of science, multi-criteria decision-making, digitization, artificial intelligence, science policy, open science.

JEL Codes: C90, C93, I23, O30, O38, D70, D73, H83

* This draft was prepared as an early draft for the NBER Economics of Science meeting in September 2025 in Cambridge, MA. The current manuscript is developed to receive feedback and to organize and frame next steps of research. Correspondence: k.boudreau@northeastern.edu. All errors are my own.

1 INTRODUCTION

Modern peer review is a remarkable institution. It is anonymous, external, reliant on millions of hours of unpaid academic labor each year, and organized largely outside the boundaries of scientists' formal employers. Yet it enacts outcomes across an entire global community of scientists by harnessing decentralized actors. This is an extraordinary institutional response to perennial problems in science: how to allocate scarce resources, validate knowledge claims, maintain standards, ensure fairness, and confer credibility to both science and the gatekeeping process itself.

Peer review simultaneously seeks to identify the most promising proposals and papers, weed out flawed or infeasible work, signal quality and legitimacy to the wider scientific community, and uphold norms of transparency, rigor, and collective responsibility. In doing so, it allocates resources and prestige, polices standards of evidence and method, maintains efficiency and feasibility in execution, reinforces disciplinary norms, and—at least implicitly—underwrites the broader social contract that scientific knowledge will ultimately yield returns for society.

Remarkably, anonymous external peer review became a global standard only well into the twentieth century. It remains a recent innovation, now buffeted by new pressures and technologies: rising submission volumes, reproducibility crises, open science initiatives, growing digital infrastructures, and the emergence of artificial intelligence. Classical accounts of the "republic of science" emphasized peer review only indirectly—as part of broader institutional mechanisms by which science organizes itself without central authority (Merton, 1973; David, 1994; Stephan, 2012). These perspectives highlighted norms, incentives, and communal governance, but offered little systematic mapping of peer review's detailed operations.

In practice, modern peer review follows a broadly recognizable pattern. External experts, working anonymously and largely without pay, are asked to assess the originality, validity, and significance of work proposed or completed by their peers. Their judgments are then synthesized by editors, program officers, or panels into consequential decisions. While there is variation in procedures—scoring systems, deliberation formats, or editorial structures—the essential process is remarkably consistent across scientific venues. This chapter considers in particular how the institution operates in two of its most influential domains: the evaluation of research proposals by grant agencies and the evaluation of completed research by scholarly journals.

A wide body of scholarship examines the workings and performance of scientific peer review, spanning studies in medicine (e.g., Rothwell & Martyn, 2000; Pier et al., 2018), the natural sciences (e.g., Eyre-Walker & Stoletzki, 2013; Siler, Lee & Bero, 2015), psychology (e.g., Mahoney, 1977; Peters & Ceci, 1982), cognitive science (e.g., Ceci & Peters, 1984), economics and management (e.g., Blank, 1991; Boudreau et al., 2016), sociology (e.g., Cole, Cole & Simon, 1981), political science (e.g., Nielsen et al., 2021), computer science (e.g., Bravo et al., 2019), engineering (e.g., Squazzoni, Bravo & Takács, 2013), and information systems (e.g., Sarker et al., 2024).

The accumulated evidence paints a mixed picture. Inter-rater reliability is often only slightly above chance (Rothwell & Martyn, 2000; Cole, Cole & Simon, 1981). Bias has been reported in favor of prestigious affiliations, U.S.-based authors, and certain demographic groups (Peters & Ceci, 1982; Blank, 1991; Tomkins, Zhang & Heavlin, 2017; Squazzoni et al., 2021). Studies also find disparities by gender and nationality (Murray et al., 2019), and systematic differences in acceptance depending on the type of research performed (Lamont, 2009). Meanwhile, strains are visible in lengthening review times and the heavy reliance on a small pool of reviewers (Kovanis et al., 2016; Fox et al., 2017).

These patterns warrant serious attention, yet interpretation requires caution. Peer review remains an institution adopted and retained across fields despite its flaws, suggesting it may function better than feasible alternatives. At the same time, theoretical benchmarks are sparse, leaving performance to be judged largely against intuitive notions of "large" or "small" effects. Many findings—such as large discrepancies among referees—are meaningful but difficult to interpret without clearer causal frameworks. Moreover, weaknesses in one part of the process may be mitigated elsewhere; divergent referee opinions, for instance, are typically synthesized by editorial judgment.

What emerges is an institution that is both indispensable and imperfect: recent in historical terms, continually adapting under pressure, complex in its workings, and uneven in its performance. The following sections trace its evolution, map its present organization, and review existing evidence on its effectiveness—providing a detailed picture of how peer review operates today and where gaps in understanding remain.

2 HISTORICAL CONTEXT OF SCIENTIFIC PEER REVIEW

Although often treated as timeless, the practice of blinded external peer review is strikingly recent—a 20th-century invention.[1] For most of history, knowledge evaluation looked fundamentally different. Legitimacy rested on reputation, lineage, and performance rather than anonymous external review.

Different traditions developed sophisticated but personal evaluation systems. In the Islamicate world, hadith studies relied on chains of transmission (isnād), while education was validated through licensing from teachers (ijazah) (Makdisi 1981; Gutas 1998). South Asian monastic universities institutionalized public disputation (śāstrārtha, vāda) as mechanisms of adjudication and prestige (Bronkhorst 2011; Pollock 2006). China's keju examination system and the kaozheng movement emphasized mastery of canonical texts and factual precision (Elman 2000, 2013). In medieval Europe, authority came through oral disputations, scriptural commentary, and the licentia docendi (Rashdall 1895; de Ridder-Symoens 1992). Across all these contexts, evaluation was deeply institutionalized—but personal, performative, and tied to social authority, not anonymized manuscript review.

[1] For historical perspectives on the development of peer review see, for example: Baldwin (2015, 2018), Csiszar (2018), Fyfe et al. (2022), Kennefick (2005), and Moxham & Fyfe (2018).

Early modern Europe continued these traditions. Galileo's ties to the Medici illustrate how courtly reputation and public demonstrations validated claims (Biagioli 1993; Shapin 1994). The "Republic of Letters" circulated manuscripts for comment, but judgments were grounded in trust and reputation rather than systematic procedures (Goodman 1994; Goldgar 1995). Even the earliest scientific journals maintained this approach. The Philosophical Transactions (1665) and the Académie des Sciences' Mémoires occasionally sought colleague input, but evaluation remained editor-centered, reputational, and rarely anonymous (Csiszar 2018; Baldwin 2015; Moxham & Fyfe 2018). Before the mid-twentieth century, scholarly evaluation was established and consequential, but profoundly different from today's peer review.

The decisive shift came only in the twentieth century. For most journals, editorial discretion and name recognition had sufficed—eminent authors expected acceptance on reputation alone. A vivid illustration occurred in 1936, when Albert Einstein and Nathan Rosen submitted a paper on gravitational waves to Physical Review. Editor John Tate, uneasy about the paper's claims, quietly sent it to an anonymous referee (later identified as Howard Robertson). Einstein responded furiously, writing that he had "not authorized" such review and immediately withdrew the paper, never submitting to Physical Review again (Kennefick 2005). This episode was shocking precisely because anonymous review was so unusual—it was used only sparingly, when editors felt suspicious of a manuscript.

The decisive institutionalization of peer review came after World War II, driven by converging pressures. Explosive growth in research output, fueled by government investment, overwhelmed traditional editorial judgment (Larsen & von Ins 2010; Bornmann & Mutz 2015). Increasing subspecialization meant no single authority could assess work across rapidly narrowing subfields, creating demand for external experts. Public funding and Cold War competition heightened accountability demands—governments and taxpayers wanted transparent processes to justify resource allocation (Baldwin 2018). The U.S. National Science Foundation's establishment of external peer review for grant proposals in the 1950s created powerful incentives for journals to follow. By the 1970s, as scientists defended their autonomy against skeptical legislators, peer review was cast as the crucial safeguard of science's credibility.

Even the world's most prominent journals adopted systematic external review surprisingly late. Nature relied largely on personal networks and author reputation through most of its first century. Only in 1973, under editor John Maddox, did Nature begin requiring routine consultation with anonymous referees—prompted by rising submissions, increasing specialization, and demands for transparency (Fyfe et al. 2022). The adoption was not driven by abstract ideals of objectivity, but by practical needs to manage volume and defend decisions against accusations of favoritism.

The near-universal expectation that manuscripts and proposals undergo anonymous review by multiple external experts emerged only since the 1970s. The system did not evolve gradually from timeless principles—it consolidated rapidly as a pragmatic response to scale, specialization, and political pressure. Its very recency

underscores why its workings are not fully understood, and why experimental evidence is needed to evaluate whether the institution achieves its aims.

2.1 Contemporary Forces Continuing to Shape Peer Review

If the late twentieth century consolidated anonymous peer review as a global standard, the early twenty-first century has brought new waves of pressure that continue to pressure and transform the institution. These include digitization and new submission platforms, the surge of submissions and reviewer burden, political and funding pressures, reforms toward openness and accountability, and most recently, automation and artificial intelligence. Each of these forces has altered and continues to exert pressure on how peer review operates.

Digitization & Platforms. Beginning in the 1990s, digitization reshaped the infrastructure of peer review. Electronic submission systems replaced postal mail, enabling near-instant global circulation of manuscripts and reviews and sharply lowering transaction costs (Meadows 2000; Tenopir & King 2000; Weller 2001). Platforms such as ScholarOne and Editorial Manager allowed editors to coordinate much larger volumes of submissions with new efficiencies (Ware 2008). Automated tools—plagiarism detection software, formatting checks, and preprint servers like arXiv and later bioRxiv—introduced new channels for dissemination and upstream evaluation (Bourne et al. 2017). Digitization expanded the reach and speed of review, but also removed frictions that had once slowed submission and resubmission cycles.

International Submission Growth. These infrastructural changes coincided with growing academic incentives to publish, producing a sustained surge in submissions. Industry surveys estimate annual growth rates of journal submissions at over 6% since 2013 (Johnson, Watkinson & Mabe 2018; Publons/Clarivate 2019). Multiple factors contribute: rising co-authorship, the global expansion of research systems (particularly in Asia), and the proliferation of specialized and open-access journals (Xie, Zhang & Lai 2020). The numbers are striking. In biomedicine alone, annual output indexed by MEDLINE rose from roughly 300,000 articles in 1990 to over 900,000 by 2015, generating demand for an estimated 15 million reviews annually and consuming more than 60 million hours of reviewer labor (Kovanis et al. 2016). The burden is highly uneven: while the median researcher completes only about one review per year, the top 10% of reviewers perform roughly half of all reviews. In grant funding, declining success rates mean ever more applications per award (Guthrie, Ghiga & Wooding 2018). Survey data confirm the mounting strain: nearly three-quarters of reviewers report heavier workloads, with one in three declining more invitations than they accept (Clarivate/Publons 2019). Many describe a sustainability problem marked by reviewer fatigue, slower turnaround, and massive duplication of effort through cycles of rejection and resubmission (Squazzoni, Grimaldo & Marušić 2020).

Political & Funding Pressures. Alongside structural and workload pressures, political and funding dynamics have grown more salient. Competitive grant systems with low success rates have encouraged strategic framing of proposals, while funders increasingly emphasize criteria such as societal impact, interdisciplinarity, and translational relevance (Guthrie, Ghiga & Wooding 2018). In the U.S., NIH and NSF review criteria now

explicitly include "broader impacts," reproducibility safeguards, and—in some programs—requirements for data sharing (NIH CSR 2024; NSF PAPPG 2023). In Europe, the European Research Council (ERC) and Horizon programs similarly foreground interdisciplinarity and societal impact, often directing reviewers to weigh policy relevance and collaboration potential (European Commission 2023). At the same time, peer review has become more firmly organized by publishers, platforms, and agencies rather than universities, reflecting a modest shift in institutional authority (Csiszar 2018). Episodes of political contestation—ranging from disputes over climate science and COVID-19 policies to high-profile challenges to scientific institutions during the Trump administration—might also indirectly place peer review under pressure, given its role for allocating credit, credibility, and resources. Public scrutiny more generally has intensified: media coverage of retractions, fraud, and reproducibility controversies has made peer review more visible beyond academia. The effect has not been a wholesale erosion of credibility, but rather heightened contestation and closer attention to whether review practices can sustain trust in science (Oransky & Marcus 2016; Nature 2023).

Automation & AI. Most recently, automation and artificial intelligence have emerged as transformative forces. Building on digitized workflows, publishers now use AI for plagiarism detection, statistical screening, and reviewer–manuscript matching (Checco et al. 2021; Feng et al. 2022). Reviewer-matching algorithms have grown more sophisticated, incorporating semantic and bibliometric data (Squazzoni et al. 2020; Zhang et al. 2022). These tools promise efficiency gains, but they also introduce new vulnerabilities. Editorial guidelines caution against uploading confidential manuscripts into public large language models, while commentators warn that generative AI may fuel low-quality or fraudulent submissions, facilitate paper mills, and complicate plagiarism detection (COPE 2021, 2023; Doskaliuk et al. 2025). Moreover, current evidence suggests AI cannot substitute for human judgment on novelty, significance, or broader contribution (Lo Vecchio 2025). In this sense, AI both streamlines peer review operations and amplifies existing strains of volume, quality control, and legitimacy—underscoring the broader need for systematic evidence and experimental evaluation.

3 INSTITUTIONAL PERFORMANCE—AND QUESTIONS RAISED

While we should interpret the evolution of peer review as an institutional response to adapt while meeting its goals, a number of claims have been made to suggest that performance of this institution may be suboptimal.[2] The now relatively large body of research on this topic across medicine, psychology, economics, sociology, education, and the natural sciences teds to emphasize weaknesses in performance rather than strengths and proficiencies—focused particularly on reviewer scoring.

[2] For broad overviews of research evaluating the performance of peer review, see for example: Bornmann (2011), Tennant et al. (2017), Squazzoni et al. (2021), Jefferson et al. (2007), and Lee et al. (2013).

Differences Across Reviews—and Meaningful Signals in Distinguish Good from Bad. One most basic question raised in prior research is whether evaluations contain much signal at all—particularly if reviewers disagree and differ in their assessments. Meta-analyses and single-journal or grant studies alike report strikingly low agreement between independent reviewers of the same manuscripts or proposals (Bornmann, Mutz & Daniel, 2010; Rothwell & Martyn, 2000; Pier et al., 2018). As Bornmann, Mutz and Daniel (2010, p. 5) put it in their meta-analysis, "the inter-rater reliability of peer assessments is quite limited, which means that the outcome of peer review is to a large extent a matter of chance." Boudreau et al. (2016) note that agreement is sometimes "barely beyond chance" (Kravitz et al., 2010, p. 1) and "perilously close to rates found for Rorschach inkblot tests" (Lee, 2012, p. 862). Also indicative of variance in reviews, Siler, Lee and Bero (2015) tracked a cohort of medical papers and found that 14 of the 15 most-cited articles were initially rejected by top journals. In economics, Gans & Shepherd (1994) surveyed 140 economists with papers in the field's "Top 5" journals and reported that roughly one in three of those articles had first been rejected at another journal before acceptance.

A series of audit and re-review studies document many instances of high variance across evaluations. Cicchetti (1991) reported that about half of NIH mental health proposals received consistent funding recommendations upon re-review, with the other half changing outcome. Nylenna, Riis and Karlsson (1994) sent the same two short, methodologically flawed manuscripts to 180 referees: only 8% recommended acceptance, 48% revision, and 44% rejection. Schroter et al. (2008) at the *BMJ* gave reviewers three test manuscripts deliberately seeded with nine major errors each; on average reviewers spotted only two, and recommendations ranged from rejection to acceptance with minor revision. Pier et al. (2018) re-ran NIH oncology grant reviews so that multiple reviewers scored the same proposals and found roughly zero correlation in reviewer scores. Fogelholm et al. (2012) compared two independent panels reviewing the same set of grant proposals and found no improvement in reliability through discussion. More recent studies echo the same pattern: Grimes et al. (2020) found large variation in reviews of identical seeded manuscripts; Grimm et al. (2020) found low concordance across German funding panels. At the same time, it is possible to observe variation even in simple description of review-level data. For example, at the U.S. National Science Foundation, Cole, Cole and Simon (1981) found that reviewers disagreed to the point that the score a proposal received from one set of reviewers was often more different from another set's score than the average differences between proposals themselves.

Yet, despite these suggestions of inconsistency, other lines of research suggest that reviewer scores are, at the same time, hardly meaningless. Bornmann and Daniel (2008), in a meta-analysis of 26 studies spanning elite journals (Nature, Science, American Economic Review) and major funding agencies (NSF, NIH), found that correlations between reviewer judgments and later citation impact were positive but modest, typically around $r \approx 0.2–0.3$. Li and Agha (2015) analyzed more than 130,000 NIH R01 grants and found that proposals with better review scores went on to produce more publications, more citations, and more high-impact work.

A one–standard deviation worse score was associated with about 15% fewer citations, 7% fewer publications, 19% fewer top-decile papers, and 14% fewer patents among funded grants. Li (2017) showed that reviewers who were closer in subject-matter expertise provided evaluations that were more predictive of which projects would later yield high-impact results. In related work, Li (2018) demonstrated that "higher-quality" committees — defined as panels made up of reviewers with stronger publication records and broader expertise — produced scores that were substantially better at forecasting future scientific impact than committees composed of weaker or narrower experts.

The body of prior studies suggests that reviewers tend to generate more reliable signals and scoring in identifying and distinguish poor work relative to distinguishing among very best submissions. Cole, Cole and Simon (1981) found reviewers often agreed on which proposals were clearly weak, but disagreed much more about those in the middle range. Cicchetti (1991) reached the same conclusion, arguing that peer review is good at rejecting inadequate studies but unreliable for ranking strong ones. Jayasinghe, Marsh and Bond (2003) reported that Australian Research Council panels could consistently weed out weak proposals but showed little agreement in ranking the top quartile. Graves, Barnett and Clarke (2011) found a similar pattern in Australian medical research grants: reviewers agreed more on weak applications than on strong ones. Siler, Lee and Bero (2015) showed that many papers which later became highly cited had received mixed early reviews, highlighting the same problem. Taken together, this evidence suggests that peer review gives useful signals when screening out poor-quality work, but is inconsistent when judging among the very best.

Submitter and Submission Characteristics—and Bias? Another prominent claim across the multiple literatures studying peer review is that correlations between evaluation outcomes and observable attributes of authors or characteristics of their submissions indicate bias in peer review. On the submitter side, for example, studies report disparities by gender, institutional prestige, and career stage (Gilbert, Williams & Lundberg, 1994; Ceci & Williams, 2011; Lee, Sugimoto, Zhang & Cronin, 2013; Helmer et al., 2017; Walker, Barros & Conejo, 2015). On the submission side, claims focus on topics such as how highly novel or unconventional papers and proposals are less likely to succeed, that interdisciplinary work faces systematic disadvantages, that "mainstream" topics and methods are favored, and that positive or statistically significant results are more readily published than null or negative findings (Langfeldt, 2006; Bromham, Dinnage & Hua, 2016; Gans & Shepherd, 1994; Laband & Piette, 1994; Fanelli, 2010; Dickersin et al., 1992).

Several studies attempt to take step towards greater control group comparisons, with methods such as audit-style resubmissions and blinding experiments. Peters and Ceci (1982) conducted the classic audit experiment in psychology, resubmitting 12 already published articles with fictitious low-prestige affiliations; only 3 were accepted while 9 were rejected, even though the content was identical. Blank (1991) ran a randomized blinding trial at the *American Economic Review*, finding that under single-blind review, submissions from prestigious institutions were significantly more likely to be recommended for acceptance than under double-blind, with prestige roughly doubling the odds of a favorable recommendation. Tomkins, Zhang and

Heavlin (2017) tested the same mechanism at a leading computer science conference by assigning 500 papers simultaneously to single- and double-blind reviewers, showing that single-blind reviewers were about 1.8 times more likely to recommend acceptance for famous authors and elite institutions. Witteman et al. (2019) exploited a policy shift at the Canadian Institutes of Health Research: when evaluations emphasized applicants rather than projects, women's success rates fell by 4 percentage points (22.5% vs. 26.0%), whereas shifting focus back to the science narrowed the gender gap. Boudreau et al. (2016) ran a triple-blind randomized review of biomedical proposals and found that evaluators gave lower scores to more intellectually proximate and novel submissions, even though these were later rated more favorably by broader expert audiences. Wang, Veugelers and Stephan (2017) analyzed patterns of journal placement and citation impact, showing that highly novel articles were more likely to appear in lower-impact journals at first publication, despite eventually receiving substantial recognition. Bromham, Dinnage and Hua (2016) demonstrated that Australian grant proposals spanning more disciplines had significantly lower success rates. Gallo et al. (2022) further observed that reviewers tended to prefer proposals they perceived as lower risk and more feasible. Large meta-analyses of publication outcomes also confirm a pervasive "file drawer" pattern in which positive results are more likely to be published than null or negative findings (Fanelli, 2010).

Other Measures of Peer Review Performance. Beyond reliability and bias, research has evaluated peer review against several other criteria. These include, for example, effort, efficiency, and reviewer burden, with reviewing found to be heavily concentrated among a minority of scholars—20% account for 69–94% of reviews in biomedical journals (Kovanis et al., 2016). Mulligan, Hall and Raphael (2013), in an international survey of over 4,000 authors, reviewers, and editors, found that respondents cited long delays as a leading frustration. Clarivate's *Global State of Peer Review* (2018) similarly estimated that individual reviews take about 19 days on average, with end-to-end publication lags stretching to a year or more, consuming tens of millions of hours annually. Delays are especially pronounced in the social sciences, where economics and management often require years due to multi-round revise-and-resubmit cycles (Ellison, 2002; Heckman & Moktan, 2020). Another performance indication that has been studied is whether peer review helps shape submissions. In journals, referees suggest changes to clarity, methods, or framing (Bornmann, 2011); in grant review, feedback has been documented to be typically shorter and more standardized, with applicants reporting mixed assessments of its usefulness (Pier et al., 2018).

Some Caution in Interpretation. Of course, there is considerable "signal" in these seemingly quite negative results, and they should be taken seriously. At the same time, several additional observations can help provide a little further context and point toward directions for future research—and, in turn, possible innovations and refinements in the practice of peer review.

*Modern peer review remains an endogenously chosen institutional response***.** However negative the indications from existing studies, the system has been adopted and retained across fields and countries as the dominant

governance mechanism of science. This persistence suggests that, despite flaws, it might still function better than the available alternatives.

There is an absence of theoretical benchmarks for performance. Just as engineers compare actual engine efficiency to the Carnot limit, or meteorologists measure relative humidity against saturation, we would benefit from analogous benchmarks for peer review. Without them, it is difficult to judge whether observed shortcomings reflect institutional failure or simply the limits of the task—in this case, one of information processing to evaluate a submission and render judgment. By analogy to economics, even if peer review falls short of a theoretical first-best solution, it may still represent a second-best institutional arrangement that persists because it performs better than feasible alternatives under real-world constraints. Thus, we might be left to judge performance relative to perfection or perhaps only intuitive notions what is a "large" or "small" effect.

Many claims based on descriptive rather than causal evidence. The bulk of empirical research on peer review is based on descriptive evidence and we might therefore also practice some caution in interpreting results. Of course, many claims are based on direct observation – such as direct observation of significant discrepancies in evaluation or high workloads – however, a combination of more discerning theory, as above, perhaps combined with more discerning causal tests might produce greater insights on the extent, nature, and possible solutions to any problems. Causal evidence often too comes in the form of expressing observed patterns in relation to a control group—an empirical benchmark. Empirical benchmarks might be especially valuable, especially in the absence of theoretical benchmarks, as mentioned above.

Interpretation of patterns is complicated by many possible mechanisms. Interpreting patterns such as differences among referees or evaluation outcomes correlated with submitter attributes may be especially complicated not just by usual problems of empirical inference, but also by a complex web of interesting factors, from institutional design (Section 5), to a range of behavior, incentive, and cognitive mechanisms shaping evaluators, acting within a complex institution (Section 4).

3.1.1 The Recent Public "Credibility Crisis."

After peer review performance had been studied for several decades acxross multiple literatures and disciplines with questions and claims of problems or at least limitations, by the 2000s these issues—and more—began to spill into the public sphere and media. Stories of fraud, failed replications, and retractions shifted what had been technical debates into headlines, and with them emerged the language of a broader "credibility crisis" in science. The term first gained traction in psychology in the early 2010s, where scholars wrote of a "crisis of confidence" after prominent replication failures and fraud cases (Pashler & Wagenmakers, 2012), and reformers called for a "credibility revolution" (Nosek, Spies & Motyl, 2012). In medicine, influential essays argued that structural features of research design and incentives meant that many published findings could not be trusted; Ioannidis (2005) famously contended that most claimed research findings are likely to be false. By the mid-

2010s, the same language had entered general news media, with *The Economist* (2013) describing reproducibility failures, retractions, and fraud as symptoms of science's growing "credibility crisis."

The first sparks came from spectacular frauds. In the mid-2000s, two *Science* papers on stem cell breakthroughs by Hwang Woo-suk were retracted after fabricated results were uncovered. The episode was covered worldwide and pointed directly at failures of editorial and peer review oversight. Retraction statistics reinforced the impression of a system under strain: Fang, Steen and Casadevall (2012) showed that of more than 2,000 biomedical and life science articles retracted through 2012, roughly two-thirds (67.4%) were due to misconduct rather than honest error. Retractions remained rare compared to the volume of publications, but their sharp rise was widely cited as a warning sign that peer review was not catching major flaws.

The early 2010s then brought replication projects that moved the conversation from scandal to systemic doubt. In preclinical cancer biology, Begley and Ellis (2012) reported that only 11% of landmark findings could be reproduced. In psychology, the Open Science Collaboration (2015) found that just 36% of studies replicated successfully. In economics, Camerer et al. (2016) replicated 18 experimental papers: 61% reproduced, but with effect sizes only about three-quarters as large as the originals. Studies of publication bias added to the picture: Brodeur et al. (2016) showed that too many results in top economics journals cluster just over the 5% significance threshold, consistent with selective reporting; follow-up work (Brodeur et al., 2020) confirmed that null results are systematically under-reported. These projects suggested that peer review was not reliably filtering the highest-quality science and, in some cases, was reinforcing incentives to prioritize "publishable" significance over robustness.

At the same time, striking cases drew headlines. In 2013, John Bohannon, a science journalist then working as a contributing correspondent for *Science*, published an investigation in which he submitted a deliberately flawed paper on a lichen-derived cancer drug to more than 300 open-access journals. The manuscript listed fictitious African researchers at invented institutions as the authors, but Bohannon himself acted as the submitting author. The paper contained obvious flaws—uncontrolled experiments, misused statistics, and conclusions that could not follow from the data—yet more than 150 journals accepted it or invited revisions, sometimes after only cursory review. The story went global and left the impression that "peer review" could be gamed or faked. In 2014, the "stimulus-triggered acquisition of pluripotency" (STAP) cell papers in *Nature* were hailed as a breakthrough, only to be retracted months later amid replication failures and evidence of data manipulation. These episodes underscored the perception that peer review was unable to catch even obvious problems before publication.

By the mid-2010s, the debate was out of specialist venues and into the mainstream press. *The Economist* (2013) called peer review "slow, expensive, prone to bias, easily abused, and poor at detecting fraud." *Nature* (2016) editorialized that the system was "perilously overstretched, yet no alternative has proved more credible." Such coverage crystallized peer review as the symbol of science's credibility problems as

essential, but also flawed. The COVID-19 pandemic then pushed the issue further. Preprints spread at unprecedented speed, illustrating both the power of rapid dissemination and the weakness of traditional peer review to vet urgent results. High-profile retractions followed, including two widely cited COVID-19 treatment papers in *The Lancet* and *NEJM* that collapsed when their underlying data proved fraudulent. By 2020–2021, mainstream outlets such as *The New York Times* and *The Guardian* were running regular stories questioning whether peer review could keep pace with a global crisis of trust in science.

4 TODAY'S ORGANIZATION OF PEER REVIEW

Having traced its historical evolution and the forces that have shaped it, we now take stock of how peer review is organized at the present moment. This section provides a systematic overview of its institutional architecture—covering governance and structure, the division of labor and roles, core process steps, and supporting infrastructure. In documenting this series of arrangements and practices, we can then begin to understand peer review as essentially an information processing organ, taking inputs from submissions, generating assessments, and aggregating them into judgments. We highlight key dimensions along which practices vary and review existing empirical evidence that examines these features. By clarifying how peer review currently operates, and where it diverges across contexts, we establish the necessary groundwork for later discussion of opportunities to apply randomized and experimental methods to evaluate and improve the system, as well as to advance our understanding of it.

4.1 Peer Review Governance & Structure

A notable feature of peer review is that it largely operates as a para-organizational system with some remove from universities. It demands substantial labor and wields significant authority over careers and knowledge, yet it is mostly organized outside the scientist's formal employers—universities, research institutes, and laboratories (Csiszar 2018; Baldwin 2018).

4.1.1 Journals

Scientific journals are governed in three broad ways.

Society Journals. Society journals are owned and operated by disciplinary associations or academies, including high-profile examples such as Science (American Association for the Advancement of Science), PNAS (U.S. National Academy of Sciences), Journal of the American Medical Association or JAMA (American Medical Association), and the Journal of the American Chemical Society (American Chemical Society). These journals finance themselves through subscriptions, membership dues, Article processing charges, and conference surpluses (Fyfe et al. 2017). Academic editors—typically active faculty appointed by the society—select referees, synthesize reports, and make final publication decisions. Referees are also academics, who provide expert peer reviews assessing novelty, rigor, and contribution; they are unpaid and motivated by

professional norms of reciprocity and service (Rennie & Smith 1996; Smith 2006). Administrative staff support this process by managing submission systems, performing conflict-of-interest checks, handling correspondence, and ensuring deadlines are met. Because their governance rests with disciplinary associations, society journals tilt strongly toward serving members, enhancing disciplinary prestige, and advancing fields of study (Baldwin 2015). These journals typically serve more specialized disciplinary communities, with circulation and readership concentrated within specific academic fields, creating strong incentives to maintain rigorous standards that reflect well on the discipline.

University Press Journals. University press journals, such as the Quarterly Journal of Economics (Oxford University Press) or the Review of Economics and Statistics (MIT Press), are formally owned by presses attached to universities but typically operate as semi-autonomous publishing arms with their own budgets and staff (Guédon 2017). Editorial authority is delegated to academic editors or societies, who—like their counterparts in society journals—make the substantive evaluative decisions. Press-employed editorial assistants or coordinators provide administrative support. This semi-autonomous status creates a dual accountability: editors must satisfy both academic quality standards and the press's financial sustainability requirements, often leading to more conservative editorial decisions that balance intellectual risk-taking with market viability. University press journals typically have smaller circulations than commercial publishers but often maintain higher prestige within specific disciplines, particularly in humanities and social sciences where they dominate the most respected outlets. Their governance within universities tilts them toward maintaining academic credibility and reputational standing while also ensuring financial sustainability for the press.

Commercial Publishers. Commercial publishers such as Elsevier, Springer Nature, and Wiley represent a third model, with flagship titles that readers will recognize: Elsevier's *Cell*, *The Lancet*, and *Journal of Financial Economics*; Springer Nature's *Nature*, *Nature Medicine*, and *Scientific Reports*; and Wiley's *Advanced Materials*, *EMBO Journal*, and the *Journal of the American Chemical Society* (co-published). In these journals, salaried professional staff editors triage submissions, select referees, and make the final acceptance or rejection decisions; academic editors are usually absent apart from advisory boards (Larivière, Haustein & Mongeon 2015; Fyfe et al. 2022). Referees are again unpaid academics, providing expert reports as part of their professional obligations. Large administrative teams support the process by handling reviewer recruitment, chasing late reviews, coordinating production and typesetting, and ensuring ethical and reporting compliance. Corporate governance tilts commercial publishers toward objectives of scale, efficiency, and profitability, reflected in heavy investment in staff editors, large portfolios, and aggressive pursuit of subscription and APC revenues (Baldwin 2018). The market concentration among commercial publishers is substantial: Elsevier (with ~2,900 titles, including *Cell*and *The Lancet*), Springer Nature (*Nature*, *BMC* series, *Scientific Reports*), Wiley (*Advanced Materials*, *EMBO Journal*), Taylor & Francis (~2,500 journals), and SAGE (~1,000 journals in social

sciences) collectively control over half—and by some estimates nearly 70%—of the global output of published journal articles, giving them significant negotiating power with universities and enabling aggressive pricing strategies that have outpaced inflation for decades (Larivière et al. 2015; Johnson, Watkinson & Mabe 2018).

4.1.2 Fund Granting Bodies

Like journals, grant peer review is organized largely outside universities and takes three main institutional forms.

Government Agencies. Governmental agencies—such as the U.S. NIH and NSF, the European Research Council (ERC), UK Research and Innovation (UKRI), and China's National Natural Science Foundation—dominate in both scale and formality. Professional staff (scientific review officers, program managers) design review processes, recruit referees, convene panels, manage conflicts of interest, and enforce compliance (NIH CSR 2024; European Commission 2023). Academics provide written reviews and participate in panel discussions, assessing merit, feasibility, and potential impact; they are usually unpaid, or occasionally compensated with modest honoraria (Stephan 2012). Final award decisions are made by program officers and governing councils, who must balance peer review judgments with programmatic priorities, budgetary limits, and national accountability requirements (Guthrie et al. 2018). Because they are accountable to legislatures and taxpayers, government agencies tilt toward missions of scientific merit combined with national policy objectives and equitable resource distribution. Their operating budgets far exceed those of journals: the NIH's Center for Scientific Review spends US$300–400 million annually, NSF devotes about US$450 million to program management, and the ERC and UKRI operate at comparable scales—an order of magnitude greater than the largest journals. Beyond these operating costs, their grant allocations are vastly greater: NIH distributes ~US$47.7 billion, the ERC ~€2.2 billion, and UKRI ~£8 billion annually. The competitive intensity these budgets support is reflected in success rates: NIH R01 grants typically have success rates around 20%, NSF around 25%, and ERC Starting Grants approximately 13%, creating fierce competition that shapes how researchers design projects and institutions allocate internal resources (NIH 2023; NSF 2023; ERC 2023). Note too that standing review panels are periodically rotated, which distributes influence but also creates challenges of consistency and institutional memory (Siler et al. 2015).

Philanthropic Foundations. Philanthropic foundations such as the Wellcome (formerly Wellcome Trust), the Gates Foundation, and the Howard Hughes Medical Institute operate in a similar fashion, albeit on a smaller scale (Viergever 2016). Program officers design competitions, recruit reviewers, and integrate advisory input, while trustees or boards retain final authority. Academics again serve as reviewers and advisors, generally unpaid. Staff coordinate the process much like in government agencies, but philanthropic governance by trustees and donors gives greater discretion to tailor peer review to mission priorities: Wellcome emphasizes frontier science, Gates global health, and HHMI long-term career development. These governance differences translate into distinct review criteria and processes: the Gates Foundation emphasizes measurable impact

metrics and translational potential more heavily than traditional academic metrics, while HHMI's investigator-focused model evaluates people rather than specific projects, leading to longer review timelines but potentially greater intellectual freedom for recipients (Nicholson & Ioannidis 2012). Philanthropic agencies are also in many cases less bound by transparency requirements than government funders (Viergever 2016). Their operating budgets remain substantial—Wellcome reports ~£8 billion annually—even if smaller than those of national agencies (Wellcome 2023).

University Seed Grants. University grant competitions, by contrast, tend to operate on much smaller scales and remain closely embedded in the academic community. In addition to providing non-competitive institutional support (faculty salaries, start-up packages, lab space), universities also run peer-reviewed competitions for seed, bridge, or pilot grants (Stephan 2012). Virtually every major research university operates some form of internal seed grant program—from University of Georgia's Faculty Research Grants (up to $25,000), to Florida State University's Seed Grants (up to $100,000), to University of Utah's college-specific competitions funded by their Vice President for Research. Columbia University offers multiple tiers, from $10,000 awards for doctoral students to $150,000 interdisciplinary grants, while Utah State University runs two competitive cycles annually. These competitions, which at large research universities may collectively distribute several million dollars each year, are primarily geared toward seed funding that enables preliminary studies, proof-of-concept work, and pilot data collection to strengthen future external proposals. The decentralized nature of these programs creates remarkable variety in organization and orientation: some universities emphasize interdisciplinary collaboration (UGA's Presidential Interdisciplinary Seed Grants), others focus on specific career stages (Columbia's junior faculty diversity grants), while still others target particular research themes (precision agriculture, global health, aging research) or methodological approaches (computational research, community engagement). This diversity reflects local institutional priorities, donor interests, and strategic research directions rather than the standardized criteria of national agencies. National academies such as the U.S. National Academy of Sciences and the UK's Royal Society administer comparable fellowship and prize schemes. In these settings, administrative staff provide minimal support, while academics themselves review, deliberate, and make the final decisions. Success rates for internal university competitions are often substantially higher than national competitions—frequently 30-50%—reflecting their role as stepping stones rather than ultimate funding destinations. Governance within the academic community tilts these competitions toward nurturing capacity, advancing disciplinary prestige, and helping researchers progress toward more substantial external funding.

Corporate Research Programs. Large technology firms such as Microsoft, Google, LinkedIn, and Meta also operate competitive research grant schemes. These programs mimic academic peer review in form—soliciting proposals through open calls, recruiting panels of company researchers and external academics, and evaluating submissions on novelty, rigor, and contribution—but governance and final decision-making remain firmly under corporate control. Review criteria blend scientific merit with alignment to strategic priorities such

as artificial intelligence, digital health, or labor market analytics. Awards are typically smaller than those of government or major philanthropies (often US$25,000–250,000) but may include privileged access to proprietary data or infrastructure. While program descriptions indicate the use of panels and external referees [Microsoft Research 2024; Google Research 2024; Meta Research 2022; LinkedIn 2019], independent scholarly analysis of how these review systems function remains scarce. Existing research on industry sponsorship suggests that such programs can shape research agendas toward corporate interests (Fabbri et al. 2018; Hirschhorn & Bialous 2001).

4.2 Objectives & Submission Selection Criteria

Journals, funding agencies, and other evaluative venues articulate their objectives in different ways, but certain themes recur. (See The Appendix for examples to complement the discussion here.) In considering the institution of peer review as an information processing organ, these objectives influence the kinds of signals evaluators are asked to extract.

Quality & Scientific Merit. The most universal of criteria emphasized by peer review bodies is general soundness and contribution—which we might think of as general scientific merit and "quality." Across contexts, reviewers are expected first to check that work is methodologically valid and appropriately reported. Yet validity alone is rarely enough: most venues also expect a judgment of contribution—whether the work adds something non-trivial to existing knowledge. *Nature* specifies that papers should "represent an advance in understanding likely to influence thinking in the field" (Nature, 2023), and *Science*requires authors to motivate the significance of their findings, signalling that mere correctness is insufficient (Science, 2023). Clinical flagships like *The Lancet* and *NEJM* set the bar as "practice-changing" or policy-relevant research (The Lancet, 2023; NEJM, 2023). Among funders, the NSF explicitly instructs reviewers to assess "Intellectual Merit"—defined as the potential to advance knowledge—while NIH reviewers score "Significance" as a separate criterion from "Approach" or "Innovation" (NSF, 2023; NIH CSR, 2024). The exception is mega-journals such as *PLOS ONE* and *BMJ Open*, which deliberately bracket contribution away: they instruct reviewers to evaluate "technical soundness, not significance," ensuring that validity alone determines acceptance (PLOS ONE, 2023; BMJ, 2023).

Feasibility. A second recurrent objective is feasibility and execution. Feasibility is fundamentally a prospective criterion: it asks whether proposed research can realistically be carried out if funded. Granting agencies make this explicit. At the U.S. National Institutes of Health, reviewers must score not only the proposed methods but also the *Investigators* and *Environment*, considering "the institutional support, equipment, and other physical resources available to the investigators" as part of the Overall Impact score (NIH CSR, 2024). UK Research Councils likewise list "Resources and Management" among their main assessment criteria (EPSRC, 2023), and Horizon Europe devotes one of its three review pillars to "Quality and efficiency of the implementation," requiring evaluation of work plans, risk management, and team capacity (European

Commission, 2021). Mission-oriented programs elevate this emphasis even further. DARPA, ARPA-H, and BARDA all require applicants to specify detailed milestones, deliverables, and credible transition strategies, treating executional capacity as co-equal with scientific merit (DARPA, 2022; ARPA-H, 2023; BARDA, 2023).

By contrast, feasibility does not normally apply to journals, where the research has already been completed at the time of review. The closest analogue is an assessment of the adequacy of execution after the fact. Reviewer guidelines for *JAMA* emphasize whether "methods are appropriate and sample size adequate" (AMA, 2023), and *BMJ* directs referees to judge whether "the study design and statistical analyses are appropriate to the research question" (BMJ, 2023). Even general guidance from COPE on journal reviewing identifies "soundness of study design" as a central responsibility for referees (COPE, 2017). A further innovation is the registered reports format, in which feasibility is assessed prospectively at the journal stage: protocols are peer-reviewed before data collection, with acceptance conditional on methodological adequacy and practical viability rather than results (Chambers & Tzavella, 2022). In sum, while funders assess feasibility as the likelihood of successful execution, journals assess whether execution was adequate to support claims once the research is complete.

Novelty. A third common dimension is novelty and potential for path-breaking advance. At selective generalist journals, reviewers are asked to determine whether findings are original and consequential enough to warrant broad attention. Nature's criteria emphasize advances "likely to influence thinking" across disciplines (Nature, 2023), while Science places significance at the center of its editorial gatekeeping (Science, 2023). Clinical journals translate importance into practice-changing potential (The Lancet, 2023; NEJM, 2023). In the funding domain, NSF directs reviewers to assess originality, NIH maintains "Innovation" as a separate criterion, and the ERC defines "scientific excellence" (its sole criterion) in terms of ambition, originality, and feasibility of the scientific approach (NSF, 2023; NIH CSR, 2024; ERC, 2023). Other funding schemes are even more explicit in prioritizing novelty: the NIH Pioneer Awards and New Innovator Awards require reviewers to evaluate "high-risk, high-reward" potential; DARPA solicitations routinely call for "revolutionary" or "disruptive" advances with step-change impact (DARPA, 2022); and the U.S. Department of Energy's ARPA-E similarly mandates "transformational" innovation criteria (Bonvillian & Van Atta, 2011). By contrast, broad-scope soundness-only journals such as PLOS ONE and BMJ Open explicitly reject novelty as a criterion, focusing on technical validity and transparency alone (PLOS ONE, 2023; BMJ, 2023). The spectrum here ranges from maximizing novelty at selective journals and frontier "high-risk, high-reward" programs to eliminating it altogether in soundness-only venues.

Other Priorities. Review systems also sometimes embed further objectives that reflect institutional identity and policy environments. Some are highly idiosyncratic. Clinical journals often stress immediate *practice relevance*—whether results are likely to change treatment or policy (The Lancet, 2023; NEJM, 2023). Public

health and disease-focused foundations (e.g. the Michael J. Fox Foundation) direct reviewers to translational potential for specific conditions (Sampat, 2012). U.S. mission agencies such as DARPA, ARPA-H, and BARDA explicitly equate success with credible transition to adoption, national security, or public-health payoff, and require proposers to articulate milestones and end-user engagement (DARPA, 2022; ARPA-H, 2023; BARDA, 2023). Funders also introduce political and societal priorities: NSF's *Broader Impacts* stresses education, diversity, and inclusion (NSF, 2023); NIH's UNITE initiative integrates equity considerations into review (NIH, 2021); Horizon Europe ties calls to missions on climate neutrality, cancer, and digital transition (European Commission, 2021). Universities and foundations align internal schemes with strategic themes, such as Harvard's Climate Solutions Fund or Stanford's Impact Labs, both of which embed community and policy engagement into their evaluation (Harvard University, 2024; Stanford University, 2023). Open-access soundness-focused journals such as *PLOS ONE* and *BMJ Open* adopt their own distinctive mission by bracketing out novelty and importance and instead requiring only technical validity, transparency, and data availability in review (PLOS ONE, 2023; BMJ, 2023).

Another example of a particular objective emphasized by at least some venues is interdisciplinary or multidisciplinary research. The European Research Council states under its Scientific Excellence criterion that proposals are expected to show "ground-breaking nature, ambition and feasibility, including where relevant novel concepts and approaches, and interdisciplinary proposals" (ERC, 2023). Horizon Europe makes interdisciplinarity explicit by requiring reviewers to assess "extent to which the proposed work is beyond the state of the art, novel, and where relevant interdisciplinary" as part of Excellence, and by embedding cross-disciplinary collaboration in its mission-oriented calls (European Commission, 2021). UK Research and Innovation similarly lists interdisciplinarity as a priority in its strategic plan and calls, instructing applicants to "demonstrate contributions across disciplines and sectors" (UKRI, 2023). In the United States, NSF and NIH encourage it mainly through cross-cutting programs (e.g., NSF's "Convergence Research"; NIH's "Common Fund") rather than as a default review criterion (NSF, 2023; NIH, 2020). Universities and philanthropic seed funds also promote it by requiring or rewarding cross-school collaborations in internal competitions (Stephan, 2012). By contrast, most journals are structurally organized around fields and disciplinary communities. At journals, research that crosses disciplinary boundaries can be disadvantaged, with evidence of difficulties in assigning referees and higher risks of rejection (Rhoten & Parker, 2004). Related analyses of grant review also suggest consistent penalties, with interdisciplinary proposals showing lower success rates and slower review processes (Bromham, Dinnage & Hua, 2016), patterns that echo the structural challenges of journal peer review.

Another example of an objective outside the big three of quality, feasibility, and novelty is demographic diversity. At the U.S. National Science Foundation, the requirement to assess Broader Impacts was introduced in 1997; at first this was loosely defined, but in the 2007 *Grant Proposal Guide* reviewers were explicitly instructed to consider how projects would "broaden participation of underrepresented groups" (NSF, 2007). Since then,

every proposal has been scored on this dimension alongside *Intellectual Merit*, making NSF the earliest funder to institutionalize EDI in review. At the National Institutes of Health, the five scored criteria (Significance, Innovation, Investigators, Approach, Environment) remained unchanged from the 1980s through 2020, with no explicit reference to EDI. In 2021, however, NIH launched the UNITE initiative and simultaneously began issuing RFAs that required reviewers to evaluate attention to health disparities and diversity in research teams (NIH, 2021). In Europe, Horizon 2020 (2014–2020) introduced gender balance as a "cross-cutting issue" but did not make it a scored element; with Horizon Europe (2021–2027), the European Commission went further by requiring every applicant institution to have a Gender Equality Plan as an eligibility condition for most calls, and by embedding EDI into the Impact pillar of review (European Commission, 2021). UKRI promoted diversity rhetorically from around 2017, but only in 2022 did it adopt a formal EDI Strategy requiring applicants and reviewers to account for equality and inclusion in the evaluation process (UKRI, 2022). Journals have only moved recently, and mainly through editorials rather than referee guidance: The Lancet launched its Gender Equity in Research series in 2019, and *Nature* and *Science* announced diversity pledges covering editorial boards, reviewer pools, and authorship policies in 2020.

A final observation on objectives is that the expectation of returns to scientific work—whether societal or economic—is rarely an explicit criterion in mainstream journal or grant review. Instead, the link from new knowledge to downstream impact is generally treated as an implicit social contract: advancing fundamental understanding is presumed to yield benefits over time, even if the pathway is uncertain and indirect. Exceptions arise in mission-driven contexts, where agencies such as DARPA, ARPA-E, ARPA-H, and BARDA make transition to practice central to their review criteria, requiring applicants to specify milestones, adoption strategies, and potential end-users (Bonvillian & Van Atta, 2011; DARPA, 2022; ARPA-H, 2023; BARDA, 2023). In Europe, Horizon Europe embeds "Impact" as one of its three review pillars, explicitly extending evaluation beyond intellectual merit to include economic, social, and policy relevance (European Commission, 2021). In the broader peer review system, however, "returns" remain a background justification rather than an overt evaluative standard, reflecting recognition that downstream benefits typically unfold over longer horizons than reviewers can credibly assess.

4.3 Peer Review Process

Across contexts, peer review proceeds through a recognizable sequence: intake and prescreening, triage, reviewer assignment, reviewer evaluation, deliberation and judgment, aggregation and decision, and feedback and iteration. This process divides across multiple roles: submitters (authors, principal investigators), editors or program officers, academic reviewers, administrative staff, and digital infrastructures (increasingly with AI tools) such as journal platforms ScholarOne and Editorial Manager; and grant management systems including NIH's eRA Commons (Electronic Research Administration), NSF's Research.gov, the

ERC's ESub electronic submission system, and UKRI's Je-S (Joint Electronic Submissions system) (NIH CSR 2024; NSF PAPPG 2023; European Commission 2023).

4.3.1 Submission Intake & Requirements

A first step in any review is submission. Historically, authors mailed paper manuscripts to journals and thick proposal packets to funders, where clerical staff manually logged, stamped, and routed them. Today, intake is almost universally digital, with the submitter—an author in journals, a principal investigator (PI) and team in grants—entering information into online portals. The shift to digital platforms has created a far greater level of standardization. For journals, systems such as ScholarOne and Editorial Manager require standardized fields for titles, abstracts, affiliations, keywords, conflict-of-interest statements, funding acknowledgments, and sometimes suggested reviewers. For grants, portals like NIH's eRA Commons, NSF's Research.gov, and the ERC's ESub demand far more extensive inputs: budget justifications, biosketches, compliance certifications, and institutional assurances, alongside project narratives (NIH CSR 2024; NSF PAPPG 2023; European Commission 2023). Across both journals and grants, digital intake forces information into structured fields—minimizing missing elements and creating standardized datasets for later prescreening, reviewer matching, and compliance checks.

Despite this standardization, different journals now impose different additional requirements, with a general trend toward expanding demands. Perhaps most consequentially, journals vary dramatically in their requirements for open materials—data, code, and other research outputs. A 2019 survey found that 31% of journals had mandatory data sharing policies, with considerable variation by field: 85% of journals in genomics require data deposition, compared to just 15% in psychology (Tedersoo et al. 2021). Science and Nature require data availability statements and may mandate deposition in recognized repositories, while PLOS journals require underlying data for all figures and tables to be made available (PLOS 2020). Some journals, particularly in computational fields, now require code availability—for instance, Nature Machine Intelligence mandates that "custom code that is central to the conclusions must be available" (Nature Machine Intelligence 2024). These requirements can impose substantial additional costs and time burdens: preparing datasets for public release, documenting code, and ensuring compliance with privacy regulations can require weeks of additional work and potentially expensive data management infrastructure (Christensen et al. 2019).

Journals also vary substantially in their fee structures. Many journals charge no submission fees, but others impose charges that vary by field and publisher type. Economics, finance, and accounting journals commonly charge submission fees (Elsevier 2024). Society journals often tie submission fees to membership (dues): the Econometric Society charges $125 for regular members and $50 for student members for Econometrica submissions, and requires membership for first-time submissions (Econometric Society 2024). The Royal Economic Society charges £75 for submissions to The Econometrics Journal, with reduced rates for RES members (Royal Economic Society 2024). Beyond submission fees, journals impose various

publication-related charges: traditional subscription journals may charge page fees of $100-250 per page and color figure fees up to $1,000 per image, while some society journals like Econometrica charge publication page fees at $12 per page. Open access journals charge substantially higher Article Processing Charges (APCs)—Nature Communications charges $5,200 and PLOS ONE charges $1,895 for publication (Nature Portfolio 2024; PLOS 2024). Additionally, some journals require costly formatting, professional language editing, or specific image resolution standards that can add perhaps hundreds to thousands of dollars in preparation costs. Beyond data sharing and fees, journals increasingly impose additional credentialing and verification requirements: many now mandate ORCID identifiers for authors, detailed biographical statements, extensive ethics declarations covering everything from AI tool usage to institutional affiliation verification, and clinical trial registration numbers where applicable. Grant applications, in contrast to journals, typically involve no submission fees in the case of major funding agencies.

Grant applications face particularly extensive requirements, given a need to evaluate prospective, uncertain research that has yet to be completed and associated investigator capacity, and significant financial commitments in projects (Guthrie, Ghiga & Wooding 2018; Stephan 2012). Unlike journal submissions where reviewers can assess completed results, grant evaluations must predict success based on detailed documentation of investigator qualifications, institutional resources, and research plans. This has created a marked shift from earlier eras when applications consisted primarily of a research proposal and basic budget to today's comprehensive packages requiring: research narratives, detailed budgets with line-item justifications, principal investigator and co-investigator biosketches, institutional facilities and resources descriptions, letters of collaboration, compliance certifications for human subjects or animal research, detailed project timelines with specific milestones and deliverables, and data management and sharing plans—now mandatory for all NIH applications regardless of funding amount as of 2023, a significant departure from previous policies that applied only to large grants, while NSF requires such plans for all proposals requesting more than $500,000 (NIH 2023; NSF 2023). In addition, requirements now often include broader impact statements detailing societal benefits, diversity and inclusion plans, research integrity training certifications, and plans for public engagement and knowledge transfer activities, with the European Research Council mandating open access publication as part of its Horizon Europe program (European Commission 2021). These data management requirements alone can demand 10-20 pages detailing data types, storage protocols, sharing mechanisms, and long-term preservation strategies. In the case of NIH R01s, Wellcome Trust Investigator Awards, European Research Council grants, and many Department of Defense research programs, pilot study results demonstrating feasibility are required, including proof-of-concept experiments, feasibility studies showing methods work, or pilot data from small-scale versions of proposed studies.

These extensive requirements typically imply considerable investment for grant applications, beyond designing the research itself. NIH R01 applications typically require 100-200 hours of PI time plus additional effort from co-investigators and administrative staff (Jacob & Lefgren 2011), while smaller grants may still

demand 20-40 hours of preparation time. In large or complex applications, universities may provide pre-award or grants-management staff to help investigators navigate these digital systems, though this support is uneven across institutions and not universal (Chubin & Hackett 1990; Stephan 2012). The availability and quality of this institutional support can significantly affect both submission success rates and the true cost of proposal preparation, with well-resourced universities enjoying considerable advantages (Danthi et al. 2014).

4.3.2 Pre-screening

Once submitted, manuscripts and grant proposals pass through automated portal checks and basic administrative pre-screening before any external review. Journal systems (e.g., ScholarOne, Editorial Manager) verify required metadata, file types, declarations, and word/page limits, and route missing items to staff (Fyfe et al., 2017; COPE, 2017). Grant portals (NIH eRA Commons, NSF Research.gov, ERC ESub) automatically flag missing forms, budget elements, or ethics attestations for correction before referral to review (NIH CSR, 2024; NSF PAPPG, 2023; European Commission/ERC, 2023).

Most large publishers also run text-similarity screening (e.g., Crossref Similarity Check/iThenticate) at submission as a triage signal; editors then manually adjudicate whether flagged overlap reflects legitimate reuse (methods, quotations, self-overlap) or potential plagiarism (Aries Systems/Editorial Manager, n.d.; Crossref, 2023; COPE, 2019). Evidence from audits and method comparisons underscores this point: raw similarity percentages are **not** determinations of misconduct and require human review (Foltýnek et al., 2020); in one medical-journal audit, ~17% of submissions were initially flagged as "unacceptable" overlap at a chosen threshold, with a smaller fraction judged serious after editorial follow-up (Higgins, Lin & Evans, 2016). In parallel, submitters complete identity/COI and policy-compliance statements (e.g., ORCID, conflicts, data/code availability, clinical-trial registration where applicable), which are checked administratively at intake (COPE, 2017; Nature Portfolio, 2024).[3]

4.3.3 Initial Rejections and Triage.

Editors or associate editors in journals perform desk rejections, eliminating 50–80 percent of manuscripts before external review (Baldwin 2015). This gatekeeping function serves the crucial institutional objective of scalable filtering—managing volume while directing the limited time and attention of expert reviewers toward the most promising submissions. As submission volumes have grown exponentially, triage has become a central institutional safeguard, combining scope assessment, completeness checks, ethical screening, and tools like plagiarism detection to protect integrity before expert judgment is even sought. In commercial publishing, professional staff editors make these calls, sometimes with the support of systems

[3] At least today, data-integrity and computational reproducibility retraction checks are typically performed after acceptance during production quality control and are treated in the (separate) verification section, if they are performed at all (Thorp, 2024; Taylor & Francis, 2022).

like *ScholarOne* or *Editorial Manager*, which can automatically flag missing metadata, formatting errors, or obvious scope mismatches. But the decision to desk reject remains predominantly human, relying on expert judgment rather than automation. Program officers at grant agencies rely less on desk triage, filtering instead through stage-one reviews or preliminary reviewer scores (NIH CSR 2024; European Commission 2023). The workload here often allocates highly trained academics or staff to repetitive low-value filtering, justified as necessary to protect scarce reviewer capacity. AI-driven triage has been piloted — for example, by Elsevier and Springer Nature, which have tested machine-learning models to flag low-quality submissions — but these tools remain proprietary, limited in uptake, and far from displacing human editorial judgment (Checco et al. 2021).

4.3.4 Reviewer Selection and Invitations.

At journals, reviewer assignment (more accurately: identifying and inviting) is a multi-step process: editors or associate editors identify potential referees (traditionally through professional networks and judgment), issue invitations, and await confirmations. Because many invitations are declined, editors typically contact more candidates than needed before securing two to four reports. Administrative staff track invitations, chase responses, and enforce conflict-of-interest declarations (Bornmann, Mutz & Daniel 2010). In commercial publishing, staff editors increasingly rely on recommender systems embedded in submission platforms, such as Elsevier's *Reviewer Recommender* or Clarivate's *Reviewer Locator*, though editor judgment remains central (Feng et al. 2022). Journal assignment is therefore largely one-off and asynchronous: each submission is matched to reviewers individually, with separate invitation and acceptance cycles.

Grant agencies employ more formalized approaches to finding and assigning reviewers. NIH uses its *Reviewer Finder* to identify referees based on bibliometric similarity, then assigns them to study sections, which may include 20–30 standing members plus additional ad hoc experts recruited for a given cycle (NIH CSR 2024). NSF assembles large ad hoc panels from its reviewer databases, while ERC panels are drawn from central rosters designed to balance expertise and geography (European Commission 2023). Here, assignment means constructing whole rosters or panels that review dozens of proposals together during a single meeting cycle**.** Administrative staff and review officers coordinate invitations, enforce COI policies, and ensure coverage across disciplines.

Reviewer assignment serves several institutional objectives simultaneously: to ensure evaluations are carried out by individuals with relevant subject-matter expertise, to reduce the influence of idiosyncratic biases through multiple independent assessments, and to spread the evaluative burden across the community (Bornmann, Mutz & Daniel 2010; Pier et al. 2018).

4.3.5 Review and evaluation.

Academic reviewers carry out the substantive evaluation, reading submissions, writing critiques, and scoring against specific criteria. Editors or program officers provide guidance, while administrative staff enforce

deadlines and platform compliance. Journal reviewers typically receive manuscripts (PDF plus source files), associated data and code repositories, and venue-specific guidance through platforms like ScholarOne or Editorial Manager. Anonymization practices vary substantially across disciplines: single-blind review predominates, double-blind review is standard in certain fields (notably economics and management), while a growing minority of journals now publish reviews alongside accepted articles (Baldwin 2015). In most cases of journal reviews, the identity of reviewers is also masked from other reviewers. Analogously, grant reviewers access comprehensive proposal packets—including research narratives, investigator biosketches, budgets, and institutional resources—through centralized portals such as NIH's Internet Assisted Review system, NSF's Research.gov, or the European Research Council platform. Unlike journal submissions, grant proposals typically identify applicants, though NIH has piloted anonymization procedures in select programs (NIH CSR 2024; European Commission 2023).

A crucial distinction in peer review systems lies between scoring-focused versus recommendation-focused evaluation approaches, which reflect different institutional philosophies about the role of expert judgment. In terms of guiding evaluation via frameworks and rubrics, journals often request narrative assessments alongside categorical recommendations (accept/revise/reject), anchored by broad criteria such as "quality," "novelty," "methodological rigor," "significance," and other categories, depending on the venue. The extent to which structured parameter-by-parameter evaluation may vary considerably. For example, journals such as Nature and Science instruct reviewers to give particular weight to dimensions like novelty and broad significance, but they do so through general guidance rather than formalized scoring rubrics; reviewers provide narrative comments and an overall recommendation, with no standardized weighting across criteria (Baldwin 2015). Importantly, many journals prioritize the reviewer's explicit recommendation (accept, revise, or reject) as the primary input for editorial decision-making, treating numerical assessments, when present, as secondary to expert judgment about the manuscript's ultimate fate.

Funding agencies, by contrast, typically mandate highly structured scoring systems that often carry more weight than explicit recommendations. NIH requires reviewers to provide bulleted strengths and weaknesses across three domains (Significance; Investigator and Environment; Innovation/Approach) using standardized 1-9 scoring scales, while the ERC evaluates proposals on Excellence, Impact, and Implementation with mandatory written justifications (NIH CSR 2024; European Commission 2023). Platform constraints enforce this standardization—for instance, NIH's system will not accept reviews in which these scoring fields are not completed. These numerical scores become the primary basis for funding decisions, with applications ranked by their overall impact scores and funded in strict numerical order until budgets are exhausted. While reviewers may offer general assessments of a proposal's merit, their explicit recommendations about whether to fund or not fund carry less institutional weight than their numerical evaluations.

To a greater degree than journals, funding agencies tend to more strictly impose standardized rubrics within mandatory digital frameworks, producing structured, quantifiable assessments suitable for comparison

and aggregation (NIH CSR 2024; European Commission 2023; Guthrie, Ghiga & Wooding 2018). These more structured and standardized outputs can then be more easily assimilated and understood during the judgement and decision-making stage of the process, which typically involves a multitude of perspectives and voices—rather than, say, a single editor's synthesis and discretion. This scoring-versus-recommendation distinction has profound implications for how expert judgment operates: journal systems preserve more space for holistic, contextual evaluation where a reviewer's overall assessment can override concerns about specific criteria, while grant systems channel expert judgment into standardized, quantifiable metrics that enable systematic comparison but may constrain the expression of nuanced professional judgment. By contrast, journals generally rely on narrative reports with minimal formal rubric guidance, leaving reviewers greater interpretive discretion and preserving flexibility in editorial judgment (Rennie & Smith 1996; Baldwin 2015; Lee et al. 2013).

Some systems occupy a middle ground, asking reviewers for both detailed scores and explicit recommendations, though institutional practice varies in which input receives greater weight in final decisions. NSF, for instance, requests both categorical ratings and written assessments but emphasizes the integration of both rather than strict numerical ranking, while some journal systems collect numerical scores primarily for editorial guidance rather than as binding inputs to accept/reject decisions.

To date, the review process has been little automated or using digital tools. However, early-stage pilots have explored the use of AI-assisted manuscript summarization and automated detection of methodological deficiencies, though implementation remains experimental and uneven across publishers and disciplines (Checco et al. 2021; Doskaliuk et al. 2025). Recent guidance from the Committee on Publication Ethics mandates that reviewers disclose AI usage and prohibits uploading confidential manuscripts to public language models (COPE 2023).

4.3.6 Deliberation and judgment.

At journals, deliberation is typically concentrated in the editor's office. Editors (in commercial houses, professional staff editors) weigh referee reports, assess fit and contribution, and make the final call, sometimes consulting boards or senior editors (Siler, Lee & Bero 2015). Outcomes are not mechanically derived from reviewer recommendations: editors routinely synthesize arguments, override split recommendations, or invite multiple rounds of revision when a paper appears promising (Bornmann & Daniel 2008; Heckman & Moktan 2020). Digital submission systems such as ScholarOne and Editorial Manager document the decision trail and correspondence. Some publishers are piloting AI-assisted summaries of reviewer input to support decision letters (Checco et al. 2021).

At grant institutions, deliberation is more collective and procedurally codified than in the case of journals. Panels discuss proposals after designated reviewers present critiques; review officers enforce COI rules and manage meetings; and specialized digital infrastructures (e.g., NIH's Internet Assisted Review system, NSF's Research.gov, the ERC's ESub platform) handle voting, scoring, and audit trails, ensuring transparency

and compliance at scale (Pier et al. 2018; European Commission 2023). These systems typically record score changes, link critiques to reviewers, and generate standardized outputs such as panel summaries, which both guide funding councils and provide formal feedback to applicants.[4]

The mode of aggregation of scores in the deliberation process varies by agency: NIH converts preliminary and post-discussion scores into an Overall Impact score; NSF panels provide ranked or tiered recommendations with bounded program officer discretion; the ERC compiles consensus-based ranked lists (NIH CSR 2024; NSF PAPPG 2023; European Commission 2023). Despite these structures, evaluation and aggregation often blur or are arguably not quite as distinct and separate steps even in relation to journals for several reasons. Designated reviewers both evaluate and sit on decision-making panels, so their critiques directly feed into the aggregation stage. Oral presentations can sway unassigned scorers who had little prior exposure to a proposal (Pier et al. 2018; Lamont 2009).

Several practices and trainings have been attempted to increase agreement and efficacy of the judgment stage in collective judgments within grant review process. For example, calibration exercises at the Austrian Science Fund (Sattler et al. 2015) train reviewers to apply scoring scales more consistently. Cross-review, as used in ERC panels and in computer science conferences such as NeurIPS, require reviewers to comment on each other's reports, making reasoning more visible. When reviewers cannot agree and many proposals are judged to be of similar quality, some agencies have experimented with using randomization to break ties. The Swiss National Science Foundation, for instance, now places proposals that meet a minimum quality threshold but rank just below the funding line into a lottery. In this system, chance—not further debate—determines which of these "gray-zone" proposals receive funding (Bieri et al. 2021).

4.3.7 Feedback and iteration.

Feedback is a core but unevenly developed part of peer review. In journals, feedback usually comes via referee reports and decision letters, with revise-and-resubmit cycles common in many fields. Economics and management appear to have strong incentives related to journal prestige that can lengthen publication timelines (Heckman & Moktan, 2020), though direct data on numbers of revision rounds across disciplines are limited. In medicine and the natural sciences, decisions are often more binary: many manuscripts are rejected outright,

[4] The details of these the enabling digital infrastructure can also imply procedural differences. For example, at the NIH, preliminary scores are "locked in" before the panel meeting, but reviewers may revise them after discussion; both the original and final scores are recorded in the Internet Assisted Review system (NIH CSR 2024). By contrast, at the ERC, external reviewers submit written evaluations in advance, but panel members are free to revise their individual assessments during the consensus meeting in Brussels (European Commission 2023). Interfaces also control whether panelists see only their own scores or the distribution of all panelists' scores: at the Austrian Science Fund (FWF), reviewers submit 0–100 scores individually without seeing others' inputs, whereas in many NSF divisions, panelists can view the spread of preliminary ratings before deliberation begins (Mutz, Bornmann & Daniel 2012; NSF PAPPG 2023). In some agencies, scores cannot be submitted without filling in structured fields—for example, NIH requires bulleted strengths and weaknesses in three mandated domains (Significance, Rigor/Feasibility, Expertise/Resources)—effectively structuring deliberation within standardized categories (NIH CSR 2024).

and resubmission may happen to different outlets rather than after multiple rounds at the same journal. Editorial management systems like ScholarOne and Editorial Manager help standardize correspondence, but the nature of feedback and revision varies sharply by discipline.

Grant agencies follow a different model. Applicants typically receive panel summaries, numerical scores, and sometimes percentile rankings (e.g. NIH via eRA Commons) (Pier et al., 2018). Feedback in grant review is often viewed by applicants as inconsistent or only partially helpful (Gallo et al., 2021). There is little systematic evidence that feedback leads to improved outcomes upon resubmission.

Studies do suggest that journal review improves manuscripts, helping clarify framing or strengthen methods, but results are mixed regarding whether these improvements lead to higher citation impact. Variation in tone, constructiveness, and usefulness of reviews has been documented (e.g. Squazzoni et al., 2021).

Digital and AI tools are at an early stage. COPE has discussed AI-assisted decision-letters and feedback summaries. Some journals/publishers report piloting systems to streamline letters and structure reviewer feedback, though publicly documented examples are rare. In 2025, Nature moved to publishing anonymized peer review files alongside accepted articles, increasing transparency in the iterative exchange (Nature Editorial Policies, 2025).

4.3.8 Verification Processes prior to Final Acceptance

For most of its history, peer review relied on the discretion of referees and the wider community to verify research claims. Problems discovered after publication could trigger corrections or retractions, but systematic checks at review or acceptance were rare. That has begun to change. In recent years, publishers and funders have institutionalized verification steps in editorial workflows, extending the process beyond novelty or importance to include technical validation. These efforts span computation, statistical reporting, data integrity, references, researcher credentials, and compliance.

Computational verification is the most ambitious but least widespread. A 2021 survey by Stodden et al. found only a handful of journals—most prominently Journal of Statistical Software and Biostatistics—with formal policies requiring editors to run code prior to acceptance, while most simply required code and data availability statements without testing them. PLOS ONE piloted code-execution checks for computational biology in the mid-2010s but never scaled the practice. eLife's "Executable Research Articles" remain a flagship experiment, limited to fewer than 100 papers between 2020–2023 (eLife, 2020a; 2020b). In short, while high-profile pilots exist, systematic computational verification is still absent in most journals.

Statistical and data-integrity checks are further advanced, especially in psychology. Nuijten et al. (2016) found roughly half of psychology articles contained at least one inconsistency between reported test statistics and p-values. Journals such as Psychological Science and Behavior Research Methods began trialing automated statcheck in the late 2010s. By 2024 adoption remained modest—dozens of journals rather than hundreds—but the matched-journals study showed a large impact where implemented. In medicine, Carlisle's (2017)

statistical methods for detecting fabricated baseline characteristics have been piloted in anesthesiology but remain resource-intensive and contested. Image-forensics software is used by some large publishers (e.g., Elsevier, Wiley), yet still touches only a small fraction of submissions; a 2020 STM survey estimated less than 20% of biomedical submissions were screened. In their follow-up, Nuijten et al. (2024) analyzed more than 7,000 articles across four psychology journals, extracting 147,000 tests. Two journals had introduced statcheck (Psychological Science and Journal of Experimental Social Psychology), while two had not (Journal of Experimental Psychology: General and Journal of Personality and Social Psychology). After adoption, the share of papers with at least one inconsistency fell from about half to under one-third, while error rates in controls remained flat. In short, when journals used the tool, authors stopped making so many basic statistical mistakes.

Reference verification has diffused more widely through centralized infrastructure. By 2022, over 18,000 journals were depositing references with Crossref, enabling automated matching against DOIs. Integration of Retraction Watch data in 2023 made it possible to flag citations to retracted work. Still, implementation remains uneven: COPE (2024) noted that while many publishers have the technical capacity, only a minority systematically screen for retractions before acceptance.

Credential verification is now routine at the level of identifiers but remains shallow. ORCID reported 16 million registered researchers by 2024, and major publishers (Elsevier, Springer Nature, Wiley) require ORCID iDs for corresponding authors. Most journals stop there; they do not verify CVs, positions, or records. Funders such as NIH and NSF, by contrast, cross-reference applications with PubMed, Scopus, and institutional databases, rejecting non-compliant or unverifiable proposals outright (NIH CSR, 2024; NSF PAPPG, 2024).

Compliance checks show the sharpest contrast between grants and journals. In grant review they are comprehensive: NIH and NSF state that applications missing IRB approvals or budget documentation will not be reviewed, and agencies return thousands per year as "non-compliant." Journals rarely go beyond requiring a statement of IRB or ethics approval, with little independent verification.

4.4 Division of Labor & Capital

Peer review constitutes one of the largest global systems of knowledge work, involving tens of thousands of editors, hundreds of thousands of reviewers, and billions of dollars in both direct expenditures and uncompensated academic labor. It follows a recognizable sequence—intake, prescreening, triage, reviewer assignment, evaluation, deliberation, and feedback—but distributes labor across distinct roles with divergent qualifications, incentives, and working conditions.

Submitters initiate the process through digital platforms, providing increasingly extensive structured data. Administrative staff coordinate intake, enforce compliance, run plagiarism and metadata checks, and manage reviewer invitations, with responsibilities ranging from part-time clerical support at society journals to highly trained scientific review officers in government agencies. Editors exercise decision-making authority,

either as volunteer academic faculty appointed by societies (a prestige-based but unpaid role) or as full-time professional staff at commercial publishers with doctoral training and career pathways. Reviewers contribute the core intellectual evaluation—unpaid, often fatigued, recruited through networks or algorithmic recommender systems.

Journals and grant agencies diverge in how this division of labor is organized. Journals rely on narrative reports with high editorial discretion, while grant agencies impose standardized rubrics, numeric scores, and panel deliberation. Journals operate through dispersed volunteerism, while agencies rely on large-scale, federally funded infrastructures such as NIH's *eRA Commons* and NSF's *Research.gov*. This yields sharp contrasts in both labor models and decision mechanisms.

Underlying all of this is the capital investment required to sustain peer review. Commercial publishers devote tens of millions annually to proprietary platforms and AI-assisted tools; government agencies commit hundreds of millions to manage compliance and transparency at scale; society journals operate with the leanest resources, dependent on member dues and shared infrastructures like ScholarOne. These disparities in funding and organization produce uneven system capacities—sophisticated automation and audit trails at some institutions, basic manual processes at others.

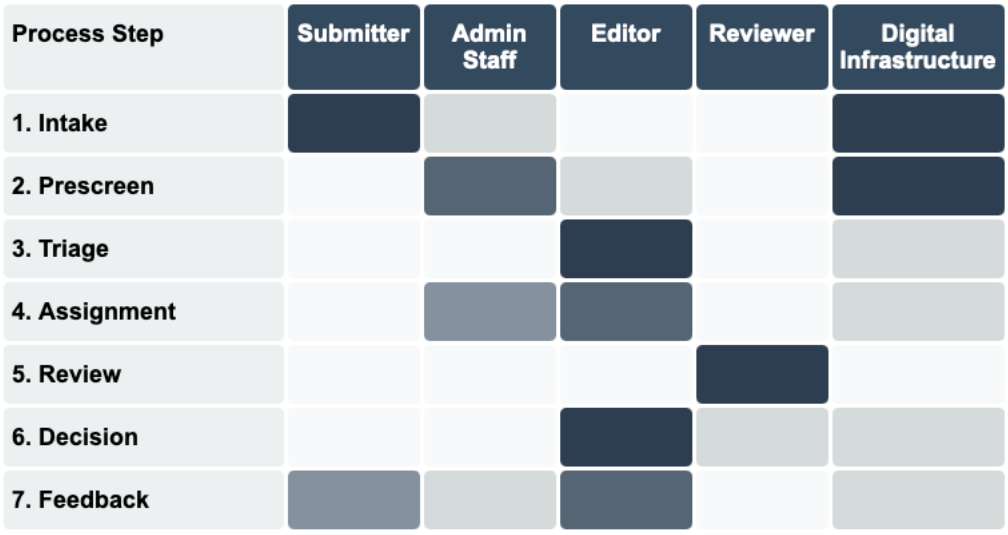


Figure 1 Illustration of the Peer Review Process Steps & Division of Labor

4.5 Incentives, Motivations, and Norms

Peer review governs careers, allocates scarce resources, and legitimizes knowledge, yet it operates outside scientists' primary employers. Its durability depends on mobilizing vast unpaid labor through norms, prestige, and professional identity rather than direct pay. Understanding these drivers helps explain both the

resilience of the institution and recurring strains—timeliness, uneven reviewer supply, and inconsistent evaluation.

For authors, incentives are straightforward: publication and funding are the currencies of academic careers (Merton 1973; Stephan 2012). Administrative staff and publisher platforms are salaried or capitalized, with professionalized incentives tied to reliability and efficiency (Resnik & Elmore 2016; Hames 2012). The puzzle lies in editors and reviewers, who provide the evaluative core. Editors average *10–20 hours per week* of work (Ware & Monkman 2008), while reviewers spend *5–10 hours per report*, with median reviewers handling *6–8 manuscripts annually* (Warne 2016; Publons 2018). Because *20–50% of invitations are declined* (Fox et al. 2017), securing two reports often requires multiple solicitations. The estimated value of this unpaid labor exceeds *$1.5 billion globally each year* (Publons 2018).

Financial and Material Incentives. Editors and reviewers in many cases receive financial or otherwise material compensation, but this is almost always symbolic relative to the time and opportunity cost required. For editors at society and university press journals, compensation may take the form of light stipends, course relief, or travel support. For example, the Seismological Society of America specifies *"a modest honorarium ($500 per year)"* for associate editors, while the American Geophysical Union offers *"an honorarium and a travel stipend, [plus] free access to all AGU journals, and complimentary registration to AGU annual meetings."* In medical physics, calls note that *"some academic journals offer a nominal honorarium (< $10,000) … others offer a stipend more commensurate to a full-time position."* In contrast, commercial publishers employ full-time salaried editors, whose incentives are tied to efficiency, rejection rates, and journal impact (Ware & Monkman 2008; Resnik & Elmore 2016).

For reviewers, journals usually provide no pay. Some offer modest tokens such as vouchers for journal article processing charge vouchers (for reviewers who are later submitters in another instance): the Royal Society provides credits *"equivalent to £200/$254/€233"* per review, and MDPI, publisher of open access journals, offers discounts for "timely and substantial" reports. Grant agencies typically provide small honoraria and reimbursements. For example, the NIH standard schedule specifies *"One day meeting: $200 honorarium, $80 per diem … $235 for miscellaneous expenses."*

Career Concerns, Status, and Recognition. Beyond modest material compensation, review and editorial service matter because they can shape careers—mainly as signals of expertise, networks, and stewardship rather than as direct, countable outputs. For reviewers, verified records of service can now be included in promotion, tenure, and funding applications, turning what was once invisible labor into an auditable contribution. This is the logic behind systems like Publons and ORCID that "record, verify, and showcase" peer-review work (Publons Global State of Peer Review 2018; da Silva 2015). Agencies also frame reviewing as explicit career development: the U.S. NIH Early Career Reviewer (ECR) program is designed to "help early career scientists become more competitive as grant applicants" by giving first-hand study-section experience and training (NIH/CSR 2015).

On the editorial side, board membership and editorship operate as prestige signals and expand networks. Editor and board roles are widely described as recognition of a scholar's standing and as positions that influence a field's trajectory and visibility (Fox & Petchey 2010; Lee et al. 2013; University Affairs 2017; SAGE 2019). Editorial board representation is also interpreted as a measure of institutional influence and visibility within a discipline (Lamont 2009). Qualitative accounts highlight the gatekeeping and stewardship responsibilities associated with these roles as central to professional identity (Karhulahti 2021).

At the university level—the direct employer of reviewers and editors—most promotion and tenure frameworks prioritize research and teaching, with service, including peer review and editorial work, treated as a third criterion that is expected and valued but typically secondary to research outputs. Promotion and tenure guidance generally treats review and editing as evidence of standing and contribution, rather than as a primary basis for advancement (Stephan 2012).

More broadly, service as an editor or reviewer confers prestige and visibility. Editors gain authority and influence over their field's trajectory, often describing their role in terms of stewardship (Fox & Petchey 2010; Lee et al. 2013). SSH editors emphasize fairness and legitimacy: *"double-blind… the most fair way of selecting what gets published"* (Karhulahti 2021). For reviewers, invitations to evaluate manuscripts signal expertise and standing. Recognition schemes—annual acknowledgment lists, "outstanding reviewer" awards, Publons/ORCID credit—formalize this status. Advocates for open peer review argue that *"reviewers are recognised because they are openly helping improve each other's work"* (Ross-Hellauer et al. 2017). Evidence from randomized studies in select journals (e.g., Bravo et al. 2019) suggests that transparency and recognition can increase reviewer willingness and in some cases improve report quality, without reducing participation.

Intrinsic Rewards, Learning, and Intellectual Stimulation. Reviewers often describe reviewing as intellectually rewarding: an opportunity to learn, stay current, and sharpen their own scholarship (Ravetz 2004; Squazzoni et al. 2013). Engaging closely with manuscripts exposes reviewers to new methods and literatures ahead of publication, while the act of formulating critiques can deepen their own analytical skills. Recent empirical evidence shows that reviewers measurably learn from evaluating others' work, improving their own subsequent research output (Ayoubi, Dumlao, & Teplitskiy 2023).

The learning dimension is increasingly recognized in how institutions and publishers structure review. Studies show that exposing reviewers to co-reports and editorial feedback strengthens these intrinsic benefits, improves review quality, and helps reviewers benchmark their own judgments (Gropp et al. 2017). Many participants describe this feedback loop as one of the most valuable aspects of reviewing, particularly for early-career researchers. Open peer review practices can reinforce these motives by making reviews visible, fostering dialogue among reviewers and authors, and providing accountability. As one study concluded, *"accountability and feedback improve the review experience"* (Ross-Hellauer et al. 2017).

Academic Norms, Professional Identity, and Culture. Norms and cultural expectations generate some of the most powerful non-pecuniary motivations for participating in peer review. The most

straightforward of these is reciprocity: reviewers and editors commonly describe their service as "doing their fair share" or "giving back." Surveys confirm that many scientists view reviewing as part of their professional duty, a way of reciprocating for reviews they themselves receive (Ware & Monkman 2008; Publons 2018). Grant panelists echo this ethos, emphasizing fairness, stewardship, and community obligation (Guthrie et al. 2018). NIH even appeals directly to these motives in its recruitment campaigns, urging reviewers to "help shape the future of science" (NIH 2015). Reciprocity thus supplies a moral obligation and a practical mechanism for sustaining reviewer labor.

A related source of motivation lies in professional identity and service expectations. Editors frequently describe their role as integral to academic life, akin to mentoring or committee work. Some universities provide course release, administrative assistance, or symbolic stipends to support editorship (Resnik & Elmore 2016). As one editor reflected, *"It's not so much about the pay, it's about being part of the conversation, being visible, doing what our discipline expects of us"* (Karhulahti 2021). Professional editors in commercial publishing and review officers in grant agencies operate under somewhat different logics — efficiency, impact, and compliance — but their m[5]otivations are likewise rooted in professional role expectations rather than material incentives (Chubin & Hackett 1990; NIH CSR 2024).

Beyond reciprocity and recognition, peer review also offers scholars a way to shape the direction of their fields and take part in the collective conversation of science. Editors often frame their responsibilities in terms of stewardship: ensuring fairness, guiding journals toward inclusivity, and influencing intellectual trajectories (Fox & Petchey 2010; Lee et al. 2013; Schultz et al. 2025). Reviewers likewise report that evaluation connects them to the ongoing dialogue of their discipline and provides a sense of visibility and belonging. NIH appeals to the same ideals when it invites reviewers to "help shape the future of science" (NIH 2015). These perceptions turn peer review into more than a gatekeeping exercise: it becomes a direct mode of influence and participation in the collective enterprise of science.

These non-pecuniary motivations are further reinforced by codified professional rules that make cultural ideals actionable. Editorial and funding bodies require confidentiality, disclosure of conflicts of interest, timeliness, civility, constructive tone, and scope discipline (COPE 2017; ICMJE 2019). NIH requires reviewers to sign legally binding confidentiality agreements (NIH CSR 2024). Blinding policies similarly operationalize cultural values: single- and double-blind review aim to reduce bias and embody universalism, while open-identity models emphasize accountability and transparency (Ross-Hellauer et al. 2017).

Seen in this light, participation in peer review is not just another academic task but also a way of living out the ethos of science. Merton (1942) described communality, universalism, disinterestedness, and organized

[5] As one editor explained, *"It's not so much about the pay, it's about being part of the conversation, being visible, doing what our discipline expects of us"* (Karhulahti 2021). Professional editors in commercial publishing and civil servants in grant agencies operate under different logics—efficiency, impact, compliance—but their motivations are likewise rooted in professional role expectations rather than material rewards (Chubin & Hackett 1990; NIH CSR 2024).

skepticism as central norms of scientific life, and peer review is perhaps the most direct means by which they are instantiated. To review is to contribute to collective knowledge (communality), to judge work on its merits (universalism), to set aside personal gain (disinterestedness), and to subject claims to critical scrutiny (organized skepticism). Peer review thus functions as a ritual of belonging — a practice in which scholars embody the values of science and affirm their membership in its collective enterprise.

REFERENCES

Baldwin, M. (2015). Making "Nature": The History of a Scientific Journal. University of Chicago Press.

Baldwin, M. (2018). Scientific autonomy, public accountability, and the rise of "peer review" in the Cold War United States. Isis, 109(3), 484–503.

Begley, C.G. and Ellis, L.M. (2012). Raise standards for preclinical cancer research. Nature, 483(7391), 531–533.

Benos, D.J., Bashari, E., Chaves, J.M., Gaggar, A., Kapoor, N., LaFrance, M., … Zotov, A. (2007). The ups and downs of peer review. Advances in Physiology Education, 31(2), 145–152.

Biagioli, M. (1993). Galileo, Courtier: The Practice of Science in the Culture of Absolutism. University of Chicago Press.

Bieri, F., Heinimann, H.R. & Reinhart, M. (2021). Random selection in research funding: An innovative approach to grant allocation. Research Evaluation, 30(2), 231–239.

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), 1041–1067.

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

Bornmann, L. & Daniel, H.D. (2008). What do peer reviewers do? A review of empirical research on peer review. Scientometrics, 77(1), 45–63.

Bornmann, L. & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis. Journal of the Association for Information Science and Technology, 66(11), 2215–2222.

Bornmann, L., Mutz, R. & Daniel, H.D. (2010). A reliability-generalization study of journal peer reviews. PLOS ONE, 5(12), e14331.

Boudreau, K.J., Guinan, E., Lakhani, K.R. & Riedl, C. (2016). Field experimental evidence on evaluation bias in innovation. Management Science.

Bourdieu, P. (1975). The specificity of the scientific field. Social Science Information, 14(6), 19–47.

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B. & Squazzoni, F. (2019). The effect of publishing peer review reports. Nature Communications, 10, 322.

Brodeur, A., Cook, N. and Heyes, A. (2020). Methods matter: P-hacking and publication bias in economics. American Economic Review, 110(11), 3634–3660.

Bronkhorst, J. (2011). Buddhism in the Shadow of Brahmanism. Brill.

Burnham, J.C. (1990). The evolution of editorial peer review. JAMA, 263(10), 1323–1329.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.H., Huber, J., Johannesson, M. et al. (2016). Evaluating replicability of laboratory experiments in economics. Science, 351(6280), 1433–1436.

Card, D. and DellaVigna, S. (2020). What do editors maximize? Review of Economics and Statistics, 102(1), 195–217.

Carlisle, J.B. (2017). Data fabrication and other reasons for non-random sampling. Anaesthesia, 72(8), 944–952.

Chambers, C. (2019). The Seven Deadly Sins of Psychology. Princeton University Press.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S. & Bianchi, G. (2021). AI-assisted peer review. Humanities and Social Sciences Communications, 8(1), 1–11.

Chressanthis, G.A. and Chressanthis, J.D. (1994). The determinants of journal submission fees. Southern Economic Journal, 61(2), 425–434.
Christensen, G., Freese, J. and Miguel, E., 2019. Transparent and reproducible social science research: How to do open science. University of California Press.
Chubin, D.E. & Hackett, E.J. (1990). Peerless Science: Peer Review and US Science Policy. SUNY Press.
Cicchetti, D.V. (1991). The reliability of peer review for manuscript and grant submissions. Behavioral and Brain Sciences, 14(1), 119–135.
Cole, S., Cole, J.R. & Simon, G.A. (1981). Chance and consensus in peer review. Science, 214(4523), 881–886.
COPE Council (2017). Ethical Guidelines for Peer Reviewers. Committee on Publication Ethics.
COPE Council (2021). Artificial intelligence (AI) in decision making: Discussion document. Committee on Publication Ethics.
Cortes, C. & Lawrence, N.D. (2021). Inconsistency in conference peer review. Communications of the ACM, 64(3), 120–128.
Crossref (2023). Crossref acquires Retraction Watch data. Crossref Blog.
Crossref (2025). Retraction metadata in the Crossref REST API. Crossref Blog/Docs.
Danchev, V., Rzhetsky, A. and Evans, J.A. (2021). Meta-research: Data sharing and reuse in the biomedical sciences. eLife, 10, e57067.
Danthi, N., Wu, C.O., Shi, P. and Lauer, M. (2014). Percentile ranking and citation impact of NIH-funded cardiovascular R01 grants. Circulation Research, 114(4), 600–606.
DARPA (2022). Broad Agency Announcement (BAA) guidance: High-risk, high-reward research. DARPA.
David, P.A. (2004). The historical origins of "open science." In Handbook of Science and Technology Studies. SAGE.
Dawson, S. and Law, D. (2024). Digital transformation of peer review. Learned Publishing, 37(2), 135–148.
Deaton, A. & Cartwright, N. (2018). Understanding and misunderstanding RCTs. Social Science & Medicine, 210, 2–21.
DeTora, L., et al. (2022). Good Publication Practice (GPP 2022) guidelines. Annals of Internal Medicine.
Dinnage, R., Bromham, L. and Hua, X. (2016). The impact of interdisciplinarity on early career research. PLOS ONE, 11(8), e0161426.
Doskaliuk, B., Zimba, O., Yessirkepov, M., Klishch, I. & Yatsyshyn, R. (2025). Artificial intelligence in peer review. Journal of Korean Medical Science, 40(7), e92.
Edwards, W. & Barron, F. (1994). SMARTS and SMARTER. Decision Analysis methods.
eLife (2020a). Executable Research Articles: press pack.
eLife (2020b). Welcome to a new ERA of reproducible publishing.
Elman, B.A. (2000). A Cultural History of Civil Examinations in Late Imperial China. University of California Press.
Elman, B.A. (2013). Civil Examinations and Meritocracy in Late Imperial China. Harvard University Press.
EPSRC (2023). Assessment criteria for proposals. EPSRC.
Erosheva, E.A., Martinková, P. & Lee, C.J. (2021). When zero may not be zero. Journal of the Royal Statistical Society: Series A, 184(3), 904–919.
European Commission (2021). Horizon Europe: Open access and data.
European Commission (2023). Horizon Europe: Proposal evaluation forms.
European Commission (2023). ERC Evaluation Guide for Applicants.
Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. PLOS ONE, 5(4), e10068.
Fang, F.C., Steen, R.G. & Casadevall, A. (2012). Misconduct retractions: Lessons. PNAS, 109(42), 17028–17033.
Feng, Y., Liang, S., Zhang, Y., Chen, S., Wang, Q., Huang, T., … Pan, H. (2022). Automated literature screening using AI. Journal of the American Medical Informatics Association, 29(8), 1425–1432.
Foltýnek, T., Meuschke, N. & Gipp, B. (2019). Academic plagiarism detection. ACM Computing Surveys, 52(6), 112.
Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A. and Väänänen, 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., et al. (2023). Double-blind peer review affects reviewer ratings. Functional Ecology, 37(1), 2–13.
Fox, F. & Paine, C. (2019). Peer Review and Diversity. Sense About Science.
Fyfe, A., Coate, K., Curry, S., Lawson, S., Moxham, N. & Røstvik, C.M. (2017). Untangling Academic Publishing. Zenodo.
Fyfe, A., et al. (2022). Publishing at Nature, 1869–2019. Nature.
Gallo, S.A., Thompson, L.A., Schmaling, K.B. & Glisson, S.R. (2019). Grant review formats and scoring. Research Evaluation, 28(2), 101–113.
Gaudino, M., et al. (2021). Effects of interventions to improve biomedical peer review. Journal of the American Heart Association, 10(15), e019903.
Goldberg, A., Stelmakh, I., Cho, K., Oh, A., Agarwal, A., Belgrave, D. & Shah, N.B. (2024). Peer reviews of peer reviews. PLOS ONE, 19(5), e0320444.
Goldgar, A. (1995). Impolite Learning. Yale University Press.
Goodman, D. (1994). The Republic of Letters: A Cultural History of the French Enlightenment. Cornell University Press.
Goodman, S. N., Berlin, J., Fletcher, S. W. & Fletcher, R. H. (1994) 'Manuscript quality before and after peer review and editing at Annals of Internal Medicine', Annals of Internal Medicine, 121(1), pp. 11–21.
Hatch, C. L. & Goodman, S. N. (1998) 'Perceived value of providing peer reviewers with abstracts and preprints of related published and unpublished papers', JAMA, 280(3), pp. 273–274.
Gruber, J. and MacLean, H. (2021). Marginal awardees: Grant funding and productivity. Research Policy, 50(1), 104133.
Gutas, D. (1998). Greek Thought, Arabic Culture. Routledge.
Guthrie, S., Ghiga, I. & Wooding, S. (2018). What do we know about grant peer review? RAND.
Hamermesh, D.S., Myers, C.S. and Schmieder, J.F. (2024). Resubmissions and the quality of published work. Journal of Economic Literature, 62(2), 377–402.
Hames, I. (2012). Peer Review and Manuscript Management in Scientific Journals. Wiley.
Harrison, G.W. & List, J.A. (2004). Field experiments. Journal of Economic Literature, 42(4), 1009–1055.
Heckman, J.J. & Moktan, S. (2020). Publishing and promotion in economics. Journal of Economic Literature, 58(2), 419–470.
Herbert, R.D., Sherrington, C. and Maher, C.G. (2021). Reproducibility under data-sharing policies. BMJ Open, 11(7), e045123.
Higgins, J.P.T., Lin, L. and Evans, J.P. (2016). Audit of plagiarism detection. Research Integrity and Peer Review, 1, 14.
Horbach, S.P.J.M. & Halffman, W. (2019). Journal peer review and editorial management. Research Integrity and Peer Review, 4, 8.
Hosier, A. and Cantwell-Jurkovic, L.P. (2025) 'AI and Library and Information Science Publishing: A Survey of Journal Editors', Library Trends, 73(3), pp. 243-266.
Huisman, J. & Smits, J. (2017). Time to accept? Publication delays. Scientometrics, 113, 1227–1248.
Hünermund, P. & Czarnitzki, D. (2019). Evaluation of funding: RDD. Research Policy, 48(1), 63–74.
Hwang, C.-L. & Yoon, K. (1981). Multiple Attribute Decision Making. Springer.
ICMJE (latest version). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals.
Ioannidis, J.P.A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124.
Jacob, B.A. and Lefgren, L. (2011). NIH postdoctoral training grants and productivity. Research Policy, 40(6), 864–874.
Jahn, N. and Tullney, M. (2016). Institutional spending on open access publication fees. PeerJ, 4, e2323.
Jasanoff, S. (2005). Designs on Nature. Princeton University Press.
Jayasinghe, U.W., Marsh, H.W. & Bond, N.W. (2003). A multilevel approach to grant review. Journal of the Royal Statistical Society: Series A, 166(3), 279–300.
Johnson, R., Watkinson, A. & Mabe, M. (2018). The STM Report (5th ed.). International STM Publishers.
Journal of Open Source Software (2025). Review criteria.
Kahneman, D., Sibony, O. & Sunstein, C.R. (2021). Noise. Little, Brown.

Kaltenbrunner, W., De Rijcke, S. and Rushforth, A. (2022). Evaluating journal transitions. Science and Public Policy, 49(1), 92–104.
Karhulahti, V.M. (2021) 'Double-blind peer review and the fairness of science publishing: Evidence from philosophy of science journals', Science and Engineering Ethics, 27(5), 67. doi:10.1007/s11948-021-00326-2
Keeney, R.L. (1992). Value-Focused Thinking. Harvard University Press.
Keeney, R.L. & Raiffa, H. (1976). Decisions with Multiple Objectives. Wiley.
Kennefick, D. (2005). Einstein versus the Physical Review. Physics Today, 58(9), 43–48.
Kovanis, M., Porcher, R., Ravaud, P. & Trinquart, L. (2016). The global burden of peer review. PLOS ONE, 11(11), e0166387.
Kovera, M.B. & McAuliff, B.D. (2000). The impact of peer-reviewed status. Journal of Applied Psychology, 85(1), 67–75.
Lalli, R. (2016). "Dirty work": Howard P. Robertson and refereeing practices. Notes and Records, 70, 151–174.
Lamont, M. (2009). How Professors Think. Harvard University Press.
Lane, J., et al. (2021). Group discussion vs independent scoring. Management Science.
Larivière, V., Haustein, S. & Mongeon, P. (2015). The oligopoly of academic publishers. PLOS ONE, 10(6), e0127502.
Larsen, P.O. & von Ins, M. (2010). The rate of growth in scientific publication. Scientometrics, 84(3), 575–603.
Lee, C.J., Sugimoto, C.R., Zhang, G. & Cronin, B. (2013). Bias in peer review. Journal of the Association for Information Science and Technology, 64(1), 2–17.
Lee, J. & Schunn, C.D. (2011). Social influences in peer assessment. Learning and Instruction, 21(2), 225–235.
Li, D. (2017). Expertise and assignment in grant peer review. American Economic Journal: Applied Economics, 9(2), 167–201.
Li, D. & Agha, L. (2015). Big names or big ideas? Science, 348(6233), 434–438.
Liu, R., et al. (2024). Reviewer anchoring in peer review: An RCT. PLOS ONE, 19(3), e0301111.
Lo Vecchio, N. (2025). AI-generated peer reviews: A case study. Research Integrity and Peer Review, 10(1), 4.
Long, C., et al. (2013). Reviewer assignment optimization. Operations Research.
Makdisi, G. (1981). The Rise of Colleges. Edinburgh University Press.
Marsh, H.W. & Ball, S. (1989). Peer review of manuscripts. Evaluation Review, 13(2), 118–145.
Marsh, H.W., Jayasinghe, U.W. & Bond, N. (2008). Improving the peer-review process. American Psychologist, 63(3), 160–168.
Matthews, K.R.W. and Ho, V. (2008). Philanthropic research funding. Journal of the Royal Society of Medicine, 101(11), 560–563.
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.
Meadows, A.J. (2000). Communicating Research. Academic Press.
Merriman, B. (2020). Peer review as organizational response. The American Sociologist, 51(3), 543–568.
Merton, R.K. (1942/1973). The normative structure of science. In The Sociology of Science. University of Chicago Press.
Merton, R.K. (1973). The Sociology of Science. University of Chicago Press.
Mimno, D. & McCallum, A. (2007). Expertise modeling for reviewer assignment. KDD Conference Proceedings.
Misra, D.P. and Ravindran, V. (2021). Plagiarism in medical research. Journal of Korean Medical Science, 36(23), e163.
Moxham, N. & Fyfe, A. (2018). The Royal Society and peer review, 1665–1965. The Historical Journal, 61(4), 863–889.
Murray, D., Siler, K., Larivière, V., Chan, W.M., Collings, A.M. & Sugimoto, C.R. (2019). Gender and diversity in peer review. Nature, 573(7774), 29–32.
Nature Editorial Policies (2024). ORCIDs at Nature Portfolio.
Nature Editorial Policies (2025). Transparent peer review: Publication of reviewer reports.
NIH (2021). UNITE: Ending structural racism in biomedical research. NIH.
NIH (2023). Data Management and Sharing Policy. NIH.

NIH CSR (2024). Center for Scientific Review Guidelines. NIH.
Nosek, B.A. & Lakens, D. (2014). Registered reports. Social Psychology, 45(3), 137–141.
NSF (2007). Grant Proposal Guide (GPG). NSF.
NSF (2023/2024). Proposal & Award Policies & Procedures Guide (PAPPG). NSF.
Nuijten, M.B. & Wicherts, J.M. (2024). Statcheck during peer review. Advances in Methods and Practices in Psychological Science, 7(2), 25152459241258945.
Nuijten, M.B., Hartgerink, C.H.J., van Assen, M.A.L.M., Epskamp, S. & Wicherts, J.M. (2016). Statistical reporting errors in psychology. Behavior Research Methods, 48(4), 1205–1226.
Nylenna, M., Riis, P. and Karlsson, Y. (1994). Multiple blinded reviews. JAMA, 272(2), 149–151.
OMB (2024). 2 CFR Revisions: Single Audit threshold to $1,000,000.
Open Science Collaboration (2015). Estimating reproducibility of psychology. Science, 349(6251), aac4716.
Oransky, I. & Marcus, A. (2016). Retraction Watch reporting and reproducibility.
Pashler, H. and Wagenmakers, E.J. (2012). Crisis of confidence. Perspectives on Psychological Science, 7(6), 528–530.
Peters, D.P. & Ceci, S.J. (1982). Peer-review practices in psychology. Behavioral and Brain Sciences, 5(2), 187–195.
Pier, E.L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M.J. & Carnes, M. (2018). Low agreement among reviewers. PNAS, 115(12), 2952–2957.
Pina, D.G., Hren, D. & Marušić, A. (2015). Peer review evaluation of EU's Seventh Framework. PLOS ONE, 10(6), e0130753.
PLOS (2020). Data availability policy.
Pollock, S. (2006). The Language of the Gods in the World of Men. University of California Press.
Porter, T.M. (1995). Trust in Numbers. Princeton University Press.
Price, A., et al. (2018). Patient and Public Involvement statements. BMJ Open.
Proofig (2022). Taylor & Francis trials Proofig software for image duplication detection.
Publons (2018). Global State of Peer Review Report.
Rashdall, H. (1895). The Universities of Europe in the Middle Ages. Clarendon Press.
Rastogi, C., Song, X., Jin, Z., Stelmakh, I., Daumé III, H., Zhang, K. & Shah, N.B. (2024). An RCT on anonymizing reviewers. arXiv:2403.01015.
Reinhart, M. (2010). Peer review practices in transdisciplinary research. Research Evaluation, 19(5), 317–329.
Rennie, D. & Smith, R. (1996). Editorial peer review. JAMA.
Resnik, D.B. and Elmore, S.A. (2016). Ensuring research integrity. Accountability in Research, 23(5), 291–301.
Richards, T., et al. (2020). Patient and public involvement reporting. BMJ.
Ross-Hellauer, T. (2017). What is open peer review? F1000Research, 6, 588.
Royal Economic Society (2024). The Econometrics Journal submission fees and policy.
Saaty, T.L. (1980). The Analytic Hierarchy Process. McGraw-Hill.
Sarker, S., Susarla, A., Gopal, R.D. & Thatcher, J.B. (2024). Human–AI collaboration in academic peer review. Journal of the Association for Information Systems, 25(1), Article 2.
Sattler, D.N., McKnight, P.E., Naney, L. & Mathis, R. (2015). Grant peer review: Training to improve reliability. PLOS ONE, 10(6), e0130450.
Schönfelder, W. (2020). Article processing charges and open access. Learned Publishing, 33(1), 89–103.
Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L. & Smith, R. (2008). What errors do peer reviewers detect? BMJ, 337, a1931.
Shapin, S. (1994). A Social History of Truth. University of Chicago Press.
Siebert, S., Machesky, L.M. & Insall, R.H. (2020). ICMJE data-sharing statements. BMJ Open, 10(4), e034002.
Siler, K., Lee, K. & Bero, L. (2015). Effectiveness of scientific gatekeeping. PNAS, 112(2), 360–365.
Smith, R. (2006). Peer review: A flawed process. Journal of the Royal Society of Medicine, 99(4), 178–182.
Snodgrass, R. (2006). Single- versus double-blind reviewing. ACM SIGMOD Record, 35(3), 8–21.
Squazzoni, F., Grimaldo, F. & Marušić, A. (2020). Journals could share peer-review data. Nature, 578(7796), 512.
Squazzoni, F., et al. (2021). Bias in peer review across 1.7M reviews. Science Advances.
Stephan, P.E. (2012). How Economics Shapes Science. Harvard University Press.

Sterling, T.D. (1959). Publication decisions and significance tests. Journal of the American Statistical Association, 54(285), 30–34.
Superchi, C., et al. (2019). Tools for assessing the quality of peer review reports. Research Integrity and Peer Review, 4, 2.
Taylor & Francis (2022). T&F trials Proofig to detect image manipulation.
Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä. et al. (2021). Data availability mandates across scholarly journals. PLOS Biology, 19(11), e3001415.
Tenopir, C. & King, D.W. (2000). Towards Electronic Journals. SLA Publishing.
The Economist (2013). Trouble at the lab. The Economist, 19 October.
Thorp, H.H. (2024). Genuine images in 2024. Science, 383. doi:10.1126/science.adn7530.
Tite, L. & Schroter, S. (2007). Why do peer reviewers decline to review? Journal of Epidemiology & Community Health, 61(1), 9–12.
Tomkins, A., Zhang, M. & Heavlin, W.D. (2017). Reviewer bias in single- versus double-blind peer review. PNAS, 114(48), 12708–12713.
UKRI (2022). Equality, diversity and inclusion strategy. UKRI.
UKRI (2023). Policies and transparency statements. UKRI.
Van Rooyen, S., Delamothe, T. & Evans, S.J. (2010). Effect of posting signed reviews online. BMJ, 341, c5729.
Van Rooyen, S., Godlee, F., Evans, S., Smith, R. & Black, N. (1999). Effect of open peer review. BMJ, 318(7175), 23–27.
Viergever, R.F. (2016). Mismatch between health research funding and global disease burden. Health Research Policy and Systems, 14, 12.
Vines, T.H., et al. (2013). Mandated data archiving. Current Biology, 23(1), R9–R10.
Walker, R. & Rocha da Silva, P. (2015). Emerging trends in peer review. Journal of Informetrics, 9(3), 717–732.
Wang, W. and Wan, X. (2018). Text mining for peer review. Journal of Informetrics, 12(4), 998–1011.
Ware, M. (2008). Peer review: Benefits, perceptions and alternatives. Publishing Research Consortium.
Weiner, B.K., Weiner, J.P. & Smith, H.E. (2010). Reviewer agreement by chance? The Spine Journal, 10(4), 308–312.
Wellcome (2023). Annual Report.
Weller, A.C. (2001). Editorial Peer Review. Information Today.
Whitley, R. (2007). The Intellectual and Social Organization of the Sciences. OUP.
Witteman, H.O., Hendricks, M., Straus, S. & Tannenbaum, C. (2019). Gender gaps at a national funding agency. The Lancet, 393(10171), 531–540.
Xie, Y., Zhang, C. & Lai, Q. (2020). China's rise in science and technology. PNAS, 117(17), 9499–9504.
Yuen, J., Lunsford, C., Carter, M. & Mullins, C. (2019). Economic aspects of journal charges. Scientometrics, 121(2), 1001–1019.
Ziman, J. (2000). Real Science. Cambridge University Press.
Zuckerman, H. & Merton, R.K. (1971). Patterns of evaluation in science. Minerva, 9(1), 66–100.

Domain	Typical intervention	Literatures / settings	Process stage	Outcomes	Examples
Identity & blinding	Toggle visibility of author identity; single- vs double-blind; open vs closed reports	Economics journals; CS/IS conferences; biomedical pilots	External review	Bias by prestige/gender/nationality; review tone/length; recommendations	**Blank (1991, *AER*)** randomized blinding during review; **Tomkins, Zhang & Heavlin (2017, WSDM)** conference trial on single- vs double-blind; **Van Rooyen, Godlee et al. (1999/2010, *BMJ*)** RCTs on open identities and effects on review quality/rec.
Audit resubmissions	Re-submit **identical**manuscripts/proposals with altered identity/affiliation signals or to new	Psychology; econ/policy; biomed grants	External review; grant scoring	Accept/reject; inter-rater reliability; prestige effects	**Peters & Ceci (1982, *Behavioral & Brain Sciences*)** classic journal audit; **Cole, Cole & Simon (1981, *Science*)** NSF re-review of identical proposals across panels; **Pier et al. (2018, *PNAS*)** independent NIH re-review exercise.
Reviewer assignment & workload/noise	Random allocation to reviewers/panels; vary reviewer load/deadlines	Sci-of-science; econ/management ops; biomedical agencies	Triage; reviewer selection	Score dispersion; reviewer fixed effects; time-to-decision; noise vs bias	**Rothwell & Martyn (2000, *Brain*)** on reliability/noise; **Siler, Lee & Bero (2015, *PNAS*)** same-paper trajectories across journals (quasi-experimental view on inconsistency); panel-allocation designs in grants (agency reports; see also Pier et al.).
Reviewer incentives & transparency	Recognition/badges; payments/charity donations; open reports/identities	Biomedical journals; cross-publisher pilots; IS venues	External review; decision letters	Willingness to review; turnaround; report length/depth; participation	**Van Rooyen et al. (1999/2010, *BMJ*)** open identities RCTs; **Bravo et al. (2019)** (cross-publisher study on open peer review practices and participation); **Ross-Hellauer (2017)**synthesis on open review modalities (context for where RCTs sit).
Editorial triage & workflow	Desk-reject thresholds; structured rubrics; reporting checklists; automated screening; dedicated	Multidisciplinary & biomed journals; management/econ pilots	Initial screening; R&R cycles	False-neg/false-pos; cycle time; quality proxies	**Cobo et al. (2011, *PLOS ONE*)** RCT adding reporting-guideline prompts/statistical review to peer review; **Hopewell et al. (2010, *BMJ*)**CONSORT flow-diagram interventions; **Schroter et al. (2004/2008, *BMJ*)**randomized editorial process changes.
Panel composition & deliberation (grants)	Vary expertise mix/seniority/gender; scoring scales; discussion order; voting rules; split-	NIH/NSF/ERC and social-science funders	Panel scoring & ranking	Reliability; order/composition effects; stability of funding decisions; downstream	**Cole, Cole & Simon (1981, *Science*)**split-panel re-review; **Pier et al. (2018, *PNAS*)** NIH re-review reliability; **Jacob & Lefgren (2011, *J. Labor Econ.*)** RD on NIH paylines for downstream impact (quasi-exp context).
Reform experiments in situ	Registered Reports; mandatory stats review; replication badges; revision deadlines/decision	Psychology/open-science; biomedical; spreading to management/IS	Pre-, external, post-review	Methodological rigor; reproducibility; timelines; author satisfaction	**Chambers (2013, *Cortex*)** Registered Reports model (policy innovation, with subsequent journal-level adoption studies); **Cobo et al. (2011)** and **Hopewell et al. (2010)** as direct RCTs on workflow features within review.

APPENDIX: JOURNAL AND GRANT OBJECTIVES

Venue	Overall Scientific Quality	Feasibility	Novelty	Other	Source
Nature	"advance in understanding… with strong evidence for conclusions"	n/a	"novelty… reason why the work deserves the visibility of a Nature Portfolio journal"	"broad significance/interest beyond specialist journals"	Nature Portfolio — Editorial Policies, Peer Review [https://www.nature.com/nature-portfolio/editorial-policies/peer-review]
Science	"scientific rigor" and quality of analysis	n/a	"novelty… indicate whether the conclusions are novel"	"significance and broad interest"	Science — Information for Reviewers [https://www.science.org/content/page/science-information-reviewers]
American Economic Review	"quality of analysis"	n/a	"innovation over the literature"	"importance of the questions, clarity of presentation, general interest"	AER — Editorial Policy [https://www.aeaweb.org/journals/aer/submissions]
Management Science	"scientific rigor" (sound design, valid inference, consideration of alternatives)	n/a	"broad interest"	"relevance to theory or practice"	INFORMS — Management Science Submission Guidelines [https://pubsonline.informs.org/page/mnsc/submission-guidelines]
MIS Quarterly	"methodological rigor" (replicable, valid design)	n/a	"strong theoretical contribution"	"relevance to the IS field"	MIS Quarterly — Policies & Reviewer Guidance [https://misq.umn.edu/pages/publication_policies]
PNAS	"scientific merit," "methodology," "quality"	n/a	"originality"	"objectivity and impartiality"	PNAS — Reviewer Information [https://www.pnas.org/reviewer]
The Lancet	"peer reviewed by experts and a statistical reviewer" (validity, quality)	n/a	—	"likely to change clinical practice, policy or substantially change thinking"	The Lancet — Peer Review Policy [https://www.thelancet.com/peer-review]
New England Journal of Medicine	"validity and quality of the research"	n/a	—	"importance to practice / clinical relevance"	NEJM — Editorial Policies [https://www.nejm.org/about-nejm/editorial-policies]
British Medical Journal	"original, important, and reliable articles"	n/a	"original"	Importance to medicine and public health	BMJ — Resources for Reviewers [https://www.bmj.com/about-bmj/resources-reviewers]
BMJ Open	"methodological soundness and transparency"	n/a	Novelty not required	"reviewers will not be asked to judge importance or breadth of appeal"	BMJ Open — Author Information [https://bmjopen.bmj.com/pages/about/]
PLOS ONE	"scientific rigor" (including compliance with ethics and data availability)	n/a	Novelty not considered	"ethical standards" and "data availability"	PLOS ONE — Reviewer Guidelines [https://journals.plos.org/plosone/s/reviewer-guidelines]

Venue	Overall Scientific Quality	Feasibility	Novelty	Other	Source
NIH R01 (US)	"Significance" and overall impact	"Approach" (design, analysis)	"Innovation"	"Investigator(s), Environment"	NIH Peer Review Criteria — Simplified Framework [https://grants.nih.gov/policy/peer/criteria.htm]
NSF Panels (US)	"Intellectual Merit" (advances knowledge)	Feasibility embedded in IM (methods, plan)	Novelty captured under IM	"Broader Impacts"	NSF Merit Review Criteria [https://www.nsf.gov/bfa/dias/policy/merit_review/]
NHS/NIHR (UK)	"quality of the research proposed"	"feasibility"	—	"significance and potential benefit to the NHS and its patients"	NIHR Peer Review Guidance [https://www.nihr.ac.uk/documents/peer-review-guidance]
ERC (EU)	"scientific excellence … sole criterion"	"Implementation" (quality and efficiency)	Originality/ambition under "Excellence"	Impact (in panel stage)	ERC Guide for Peer Reviewers [https://erc.europa.eu]
Marie Skłodowska-Curie Actions (EU)	"Excellence"	"Implementation" (quality/efficiency)	Novelty within "Excellence"	"Impact"	MSCA Evaluation Criteria [https://marie-sklodowska-curie-actions.ec.europa.eu/]
Horizon Europe (EU)	"Excellence"	"Quality and efficiency of the Implementation"	Novelty within "Excellence"	"Impact"	Horizon Europe Evaluation Criteria [https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/expert-area]
Wellcome Trust (UK)	"high-quality research"	"feasibility," environment, team	"bold and creative ideas"	Importance; suitability of applicants	Wellcome Written Expert Review Guidance [https://wellcome.org/grant-funding/guidance]
UKRI / EPSRC (UK)	"vision" (research excellence/importance)	"approach," "resources and cost justification"	Novelty under "vision"	"applicant and team capability," "ethics and responsible research and innovation"	UKRI Funding Service — EPSRC Guidance [https://www.ukri.org/apply-for-funding/how-we-make-decisions/how-reviewers-use-the-ukri-funding-service/]
UKRI / MRC (UK)	"Importance," "Scientific potential"	"Resources requested" proportionate	Novelty within "importance/potential"	Ethics, value for money	MRC Guidance for Peer Reviewers [https://www.ukri.org/councils/mrc/guidance-for-reviewers/]
FWF (Austria)	"quality of the proposed research"	"approach and feasibility"	"innovation and originality"	Researchers' qualifications; ethics/gender	FWF PI Projects Guidelines [https://www.fwf.ac.at/en/research-funding/fwf-programmes]
ARC (Australia)	"Project quality and innovation"	"Feasibility" (10% weighting)	"innovation" explicitly weighted	"Investigator(s)/Capability," "Benefit"	ARC Discovery Projects Selection Criteria [https://www.arc.gov.au/policies-strategies/peer-review]
NHMRC (Australia)	"Scientific Quality"	"Team Quality & Capability" (delivery feasibility)	"Significance … and/or Innovation"	"Impact"	NHMRC Peer Review Guidelines [https://www.nhmrc.gov.au/funding/nhmrc-peer-review]
CIHR (Canada)	"Significance and Impact of the Research" (Concept)	"Approaches and Methods"; "Expertise, Experience and Resources"	Novelty under "Concept"	Relevance (program-specific)	CIHR Project Grant Adjudication Criteria [https://cihr-irsc.gc.ca/e/51645.html]