CYNTHIA CUMMINGS

IN REFEREES WE TRUST?

Melinda Baldwin

The imprimatur bestowed by peer review has a history that is both shorter and more complex than many scientists realize.

Melinda Baldwin is PHYSICS TODAY's Books editor and the author of *Making "Nature": The History of a Scientific Journal* (University of Chicago Press, 2015).

In the early summer of 1936, Albert Einstein and his assistant Nathan Rosen submitted a paper on gravitational waves to *Physical Review*. In it they argued that gravitational waves did not exist—a controversial claim that went against the prevailing scientific consensus. Six weeks after the paper's submission, *Physical Review* editor-in-chief John Torrence Tate wrote back to Einstein with a copy of a critical referee report and asked for a response to the reviewer's comments.

So far, this story will sound familiar to most PHYSICS TODAY readers. Modern scientists expect that their submissions to journals will be read and criticized and will require revision before they are admitted into the corpus of published scientific literature. Einstein, however, did not share those expectations. In fact, he was surprised and offended by the idea that his paper had been sent out for external review. (See the article by Daniel Kennefick, PHYSICS TODAY, September 2005, page 43.) In his riposte to Tate, Einstein said that he and Rosen

> had sent you our manuscript for publication and had not authorized you to show it to specialists before it is printed. I see no reason to address the—in any case erroneous—comments of your anonymous expert. On the basis of this incident I prefer to publish the paper elsewhere.

Einstein kept his word. He would never again submit a research article for publication in *Physical Review*.

It might be tempting to view Einstein's reaction as a show of ego by a senior physicist who thought his fame would allow him to skip the peer review process. However, digging deeper into the history of peer review uncovers a more complicated picture. In 1936 refereeing was not a universal practice at the world's top scientific journals. It was not even a universal practice at *Physical Review*. Einstein's previous submission to that journal, the famous 1936 Einstein-Podolsky-Rosen (EPR) paper, was not sent out for referee reports despite its provocative anti-quantum conclusions.

So Einstein's bafflement at receiving an anonymous report criticizing his paper was hardly inexplicable. But 80 years later, peer review is an expected and established part of publishing for scientists and scholars in almost every academic discipline. How did this process become so ingrained in scientific life?

The origins of journal refereeing

Many academic and popular articles about peer review assign it the same origin story. In 1665 the Royal Society gave its secretary Henry Oldenburg permission to compile *Philosophical Transactions of the Royal Society of London*, generally regarded as the world's first scientific journal. Oldenburg immediately thought it wise to gather expert opinions on the papers he wanted to publish. Thus peer review was born and was ever after a consistent part of scientific publishing.

Or was it?

That origin story appears to have its roots in a famous 1971 sociology article—Harriet Zuckerman and Robert Merton's "Patterns of evaluation in science: Institutionalization, structure and functions of the referee system."[1] Zuckerman and Merton's article, based on an analysis of referee decisions at *Physical Review*, remains a foundational study of the sociology of peer review. It was so groundbreaking that PHYSICS TODAY printed a condensed version in its July 1971 issue (page 28). In crediting Oldenburg with the invention of peer review, Zuckerman and Merton implied that peer review's form and function had changed little since the 17th century.

More recent historical work, however, has called Zuckerman and Merton's history into question. In reality, Oldenburg rarely consulted outside opinions on what should be published in *Philosophical Transactions*, and he held such close control over the journal's contents that he occasionally referred to himself as its "author." There was not even a formal submission process. Oldenburg would simply print what interested him and what he thought might be of value to his readers, including not only experimental papers but secondhand accounts of others' experiments, discussions of recent books, and even his own personal correspondence.[2]

Although Oldenburg was indeed a pivotal figure in the history of science publishing, he was not peer review's inventor. That honor arguably belongs to William Whewell, a Cambridge University polymath who also coined the terms "physicist" and "scientist." In 1831 Whewell suggested that the Royal Society should commission written reports on papers submitted for publication in *Philosophical Transactions*. He thought those

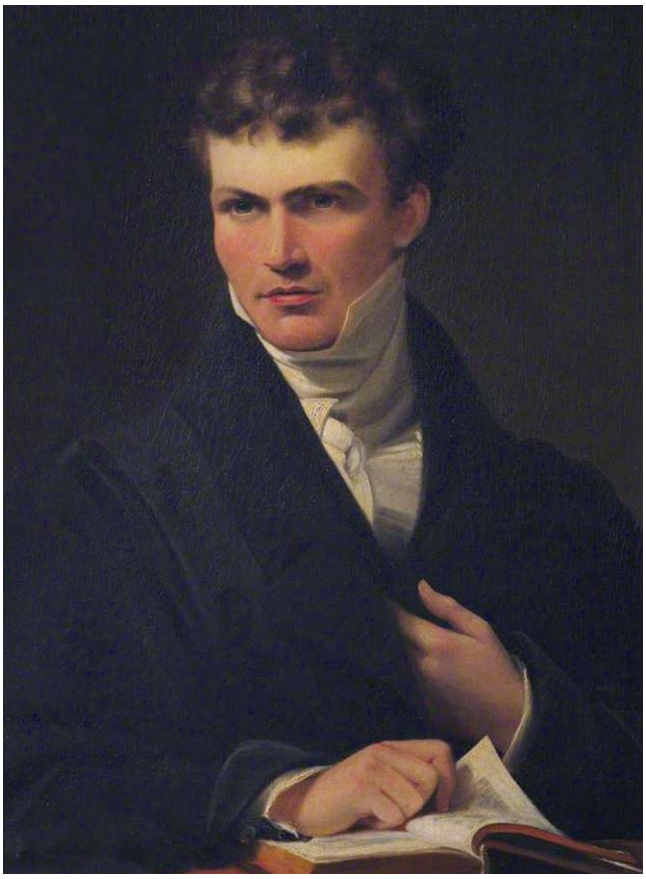

FIGURE 1. THE ENGLISH POLYMATH WILLIAM WHEWELL (1794–1866) proposed in 1831 that the Royal Society should collect and publish reports on *Philosophical Transactions* papers.

reports should then be published in the Society's new journal, *Proceedings of the Royal Society of London*, thereby fulfilling the dual purpose of fostering rich scientific discussions and providing material for the new publication.[3]

The Royal Society adopted Whewell's suggestion of soliciting reports but shifted quickly away from his vision of printing them for public discussion. A handful of reports did appear in *Proceedings*, but by the mid 1830s that practice had ceased. Instead, the society decided that referee opinions were mainly useful for helping it avoid printing anything embarrassing in its publications. By the mid 19th century, refereeing for *Philosophical Transactions* was almost entirely run by two secretaries, one in the physical sciences and one in the biological sciences. The secretaries were eminent members of the society, and they each worked with an assistant secretary to arrange refereeing for the papers submitted to *Philosophical Transactions*. The referee reports came to be seen as confidential documents for the internal use of the society. For many years, they were not made available to the authors of accepted or rejected papers.

Because authors did not see the reports, there was no real equivalent to today's common "revise and resubmit" decision. Submissions to *Philosophical Transactions* were either accepted or rejected. However, the secretaries did occasionally encourage authors of *Philosophical Transactions* papers to revise their articles before they went to print. Physicist George Gabriel Stokes, who served as the society's physical sciences secretary for more than 30 years, often suggested changes to authors via personal correspondence. Stokes would paraphrase useful comments from the *Philosophical Transactions* referees, and if Stokes himself had refereed the paper, he would often send the author a copy of his full, signed report.[4]

Refereeing in the early 20th century

At the end of the 19th century, an important shift began to take place in the scientific community's view of referees. With concerns growing about the overall quality of the scientific literature, the referee was no longer simply helping protect the reputation of a scientific society or journal. Instead, the referee was increasingly seen as someone whose work was to protect the reputation and trustworthiness of the entire scientific literature, to staunch a flood of "veritable sewage thrown into the pure stream of science," as physiologist and Member of Parliament Michael Foster put it.[3]

By the early 20th century, refereeing procedures had spread to most scientific societies in the English-speaking world. In theory the procedures were wide-ranging, but in practice the referees themselves tended to belong to small networks of elite scientists.[5] Early-20th-century refereeing procedures were less formal than the ones we now associate with scientific journals, and authors usually did not see referee reports.

At *Physical Review*, for example, referees knew that the editor would paraphrase their comments for authors and often submitted brief, casual, and occasionally sarcastic reports. Frequent referee Howard P. Robertson (1903–61) once suggested that a paper could be improved "if it were written in invisible ink." It was not until 1935 that *Physical Review* offered referees a standard questionnaire about papers. And not until the 1960s did systematic refereeing for all papers become an official policy.[6]

Commercial journals printed by for-profit publishers were even less likely to employ systematic refereeing before the Cold War. Indeed, publications like the *Philosophical Magazine* or *Nature* would continue to keep editorial deliberations in-house well into the 20th century. Those periodicals placed a high value on printing issues quickly, and many were run by ambitious editors who saw little reason to consult anyone outside a small circle of trusted advisers to decide whether or not a paper was good.

Likewise, many prominent journals outside the English-speaking world relied heavily on the judgment of their editors to select content. Such journals often counted some of the country's most respected scientists among their editorial staff. For instance, Max Planck was a longtime member of the editorial board at the revered physics journal *Annalen der Physik*. Few physicists would have questioned Planck's ability to decide, with or without any outside opinions, which papers belonged in *Annalen*.

The story of external refereeing at grant organizations is similar to the story for journal refereeing. Private grant organizations such as the Rockefeller Foundation generally left funding decisions in the hands of trusted middle managers long after World War I.[7] Grant organizations associated with governments or scientific societies were more likely to use external refereeing, although the practice was by no means universal.

When the US government formed the National Institutes of Health in 1948, NIH's division of research grants initially evaluated grant applications with little or no consultation with outside referees. Instead, small "study sections" composed of NIH-affiliated scientific experts were the first to review proposals.

Final authority over funding decisions rested with institute directors—the heads of NIH's constituent institutions, such as the National Cancer Institute.

The National Science Foundation, established by federal law in 1950, was more reliant than NIH was on outside experts for opinions on proposals. Some proposals were sent out ad hoc for mail review: Copies of the proposal were mailed to scientists who submitted their comments by return mail. Other proposals were evaluated by special panels of experts assembled in Washington, DC.

As was the case with NIH, however, decisions about funding at NSF were largely in the hands of NSF employees. Directors were responsible for deciding which proposals to fund, and referee opinions were seen as only one piece of their decision—an important piece, but not the determining factor for whether NSF would award or withhold funding. Furthermore, at both NSF and NIH, referee reports were not shared with grant applicants. Scientists who submitted proposals would receive only a short summary prepared by a government employee that stated the major reasons for acceptance or rejection.

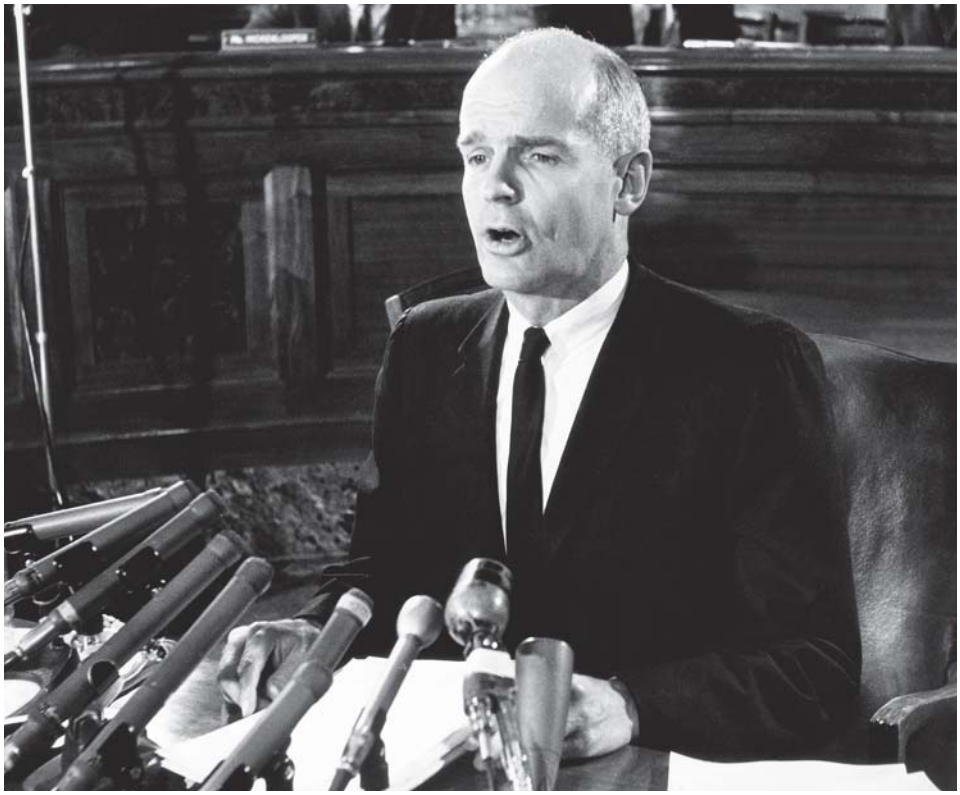

FIGURE 2. DEMOCRATIC SENATOR WILLIAM PROXMIRE (1915–2005) was a vocal critic of NSF in the 1970s and a master of the acerbic press release. (Image reproduced by permission of the Wisconsin Historical Society.)

Before the Cold War, journals or grant organizations that eschewed refereeing or placed significant power in the hands of editors and directors were not seen as less reliable or less scientific than ones that depended on referees. And the story of Einstein's clash with *Physical Review* shows that researchers who were accustomed to editors or foundation directors making decisions did not necessarily see external refereeing as a superior system. Why, after all, should an author trust the word of an anonymous referee rather than a respected editor or program director who was willing to sign his name to his remarks?

Public trust and peer review

The term "peer review" first began to appear in the scientific press in the 1960s. Interestingly, the term does not seem to have originated at journals. Instead, "peer review" was originally used to describe review committees at grant organizations—most often NIH—and in the medical community.

"Peer review means different things to different people," physician and researcher Irvine H. Page explained in a 1973 editorial for the *Journal of the American Medical Association*. He continued:

> To most American physicians it means PSRO [the Professional Standards Review Organization, which reviewed compliance with American Medicare laws], to the British House of Lords it means Peers examining other Peers for moral turpitude, and to the scientific community, it means Study Sections and Councils that determine a grantee's financial and possibly research future.[8]

Significantly, journal refereeing was not one of the definitions Page offered, although scientists and editors slowly adopted the term for that purpose over the course of the 1970s.

One episode that brought the term into more common use was a 1975 controversy about funding at NSF—a controversy that would both highlight and solidify peer review's increasing importance to the research community. Scientists in the US, particularly physicists, entered the Cold War riding on the success of the Manhattan Project. By 1953 US government spending on science had increased by a factor of 25 from its prewar numbers—and science's public profile only increased after the Soviet Union beat the US into space with the launch of *Sputnik* in 1957.[9]

The US enthusiasm for science funding proved finite, however. As early as 1966, a study by the Department of Defense concluded that DoD spending on basic research had not yielded significant progress on the department's goals, such as new weapons. The study was published in a document called the Project Hindsight report, whose findings caused some legislators and commentators to begin questioning scientific spending more broadly. Project Hindsight was an early hint that the social and financial status scientists had acquired in the early Cold War might be at risk.

By 1975 the Cold War had entered a relatively calm period of official détente between the two superpowers. The goal of keeping up with the USSR seemed less crucial. Furthermore, the US was suffering from an economic crisis. Oil and gas supplies shrank when several major oil-producing countries refused to sell oil to the US in retaliation for the country's support for Israel in the 1973 Arab–Israeli war. Economic growth stalled. Inflation and unemployment soared. With Congress under pressure to trim expenditures from dwindling tax revenues, a handful of lawmakers set their sights on NSF.

The most prominent NSF opponent was Senator William

Proxmire, a colorful Wisconsin Democrat with a knack for publicity. In March 1975 Proxmire began issuing his famous Golden Fleece Award, which he gave to the government project that he deemed the month's worst use of taxpayer money. The first two Golden Fleece Awards went to NSF projects: a University of Wisconsin sociological study about interpersonal attraction and psychologist Ronald Hutchinson's work on why humans, rats, and monkeys clench their jaws in moments of stress. Proxmire called on NSF to "get out of the love racket" and declared that Hutchinson's "nonsense" had "made a monkey out of the American taxpayer."

Meanwhile, an ambitious Republican congressman named John Conlan began criticizing NSF's spending on its education programs, particularly Man: A Course of Study (MACOS) and the Individualized Science Instructional System (ISIS). MACOS was a social sciences curriculum that had been controversial since the early 1970s in Conlan's home state of Arizona, where critics claimed that it promoted moral relativism. ISIS, a program aimed at fourth-graders, was accused of being too explicit about reproductive education.

In his quest to discover why MACOS and ISIS had received government funding, Conlan came into conflict with NSF leadership, including the foundation's director, H. Guyford Stever. Conlan requested full copies of NSF's referee reports, along with the names of the reviewers. Stever replied that referees submitted their reports under an "implied promise of confidentiality" and that releasing the text of the reports or the names of the reviewers would violate NSF policy. Conlan, however, was not persuaded.

> I would again remind you that I am a Member of Congress on a Committee charged with the oversight of the National Science Foundation. . . . Consequently, I do again demand that you make available the peer reviewer comments originally demanded by me—in their original and complete form, not paraphrased.[10]

The public debate and private exchanges over NSF grants led to the National Science Foundation Peer Review Special Oversight Hearings, held before the House subcommittee on science, research, and technology in July 1975. Over the course of six days, Congressional questioners and witnesses discussed NSF's peer review process at length.

In his testimony, Conlan argued that NSF's system placed too much decision-making power in the hands of NSF directors, and did not give enough weight to referee reports. He claimed that the only way to hold the foundation accountable was to make referee reports public, along with the names of the referees.

The NSF team came to the hearing prepared to make changes in response to the criticisms. Director Stever announced that as of 1 January 1976, applicants would be given full copies of their referee reports instead of just the summaries. However, Stever insisted that referees must remain anonymous to ensure their candor. NSF leaders also indicated that in the future, a new audit office would ensure that directors were placing appropriate weight on positive and negative referee reports—in other words, placing more decision-making power in the hands of referees.

Following the hearings, NSF's education programs were significantly downsized, and funding for MACOS and ISIS was almost entirely eliminated. However, NSF's peer review reforms quieted the fiercest criticisms, at least temporarily, and the controversy soon faded from public view. Proxmire, meanwhile, became embroiled in a lawsuit when Golden Fleece

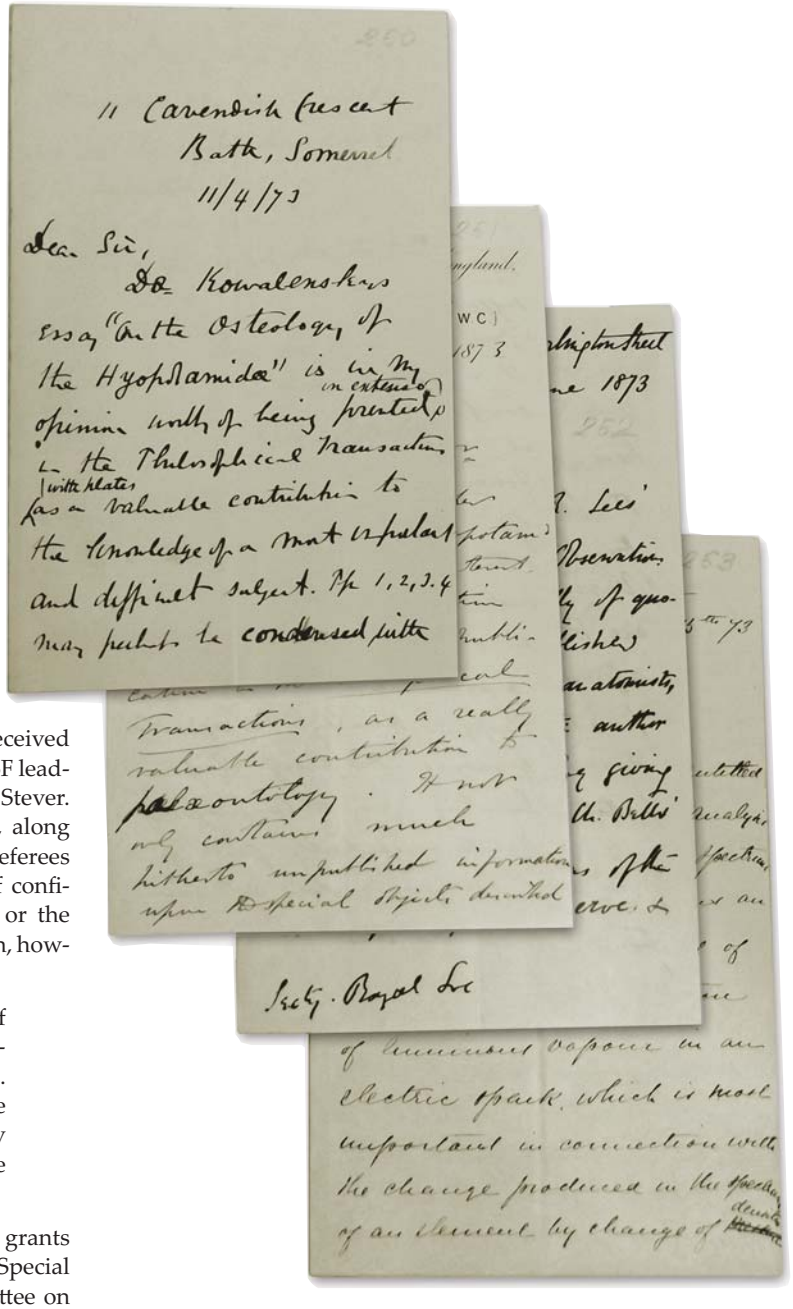

FIGURE 3. SEVERAL SHORT, HANDWRITTEN REFEREE REPORTS submitted to the Royal Society in 1873. It was common for referees to simply recommend publication or rejection with only a few explanatory comments. These reports would not have been seen by the paper's author. (Image reproduced by permission of the Royal Society Library and Archives, item RR_7_176.)

awardee Hutchinson sued him for libel. Proxmire eventually made a public apology to the psychologist and omitted individual names from future Golden Fleece press releases.

Although most of the criticisms were leveled at the social sciences, scientists from across disciplines followed the controversy. PHYSICS TODAY reported closely on the hearings and on NSF's policy changes. Editor-in-chief Harold Davis argued in an editorial that the hearings demonstrated "that peer review is by far the best means we have for deciding how funding should be distributed in a given area." (See PHYSICS TODAY, September 1975, page 96.) In the same editorial, Davis went on to announce that PHYSICS TODAY would be sending complimentary issues to every member of Congress to illuminate the inner workings of the scientific community. As Davis put it, "In an age in which the issues of society cannot avoid being ever more closely involved with science and technology we are going to need more peer review, not less."

The 1960s and 1970s seem to have been a crucial period of transition for ideas about peer review. In the mid 20th century, external refereeing was simply one of several methods a journal or grant-issuing organization could use to choose which submissions to accept or reject. By the end of the Cold War, peer review was a prerequisite for scientific respectability.

The NSF controversy strongly suggests that one reason for the increased emphasis on peer review, at least in the US, was a shifting relationship between scientists and the public during the Cold War. Spending on both basic and applied research had increased dramatically in the 1950s and 1960s—but when doubts began to creep in about the public value of the work that money had funded, scientists were faced with the prospect of losing both public trust and access to research funding. Legislators wanted publicly funded science to be accountable; scientists wanted decisions about science to be left in expert hands. Trusting peer review to ensure that only the best and most essential science received funding seemed a way to split the difference.

Peer review in crisis?

Today peer review is an expected part of publishing any scientific article or obtaining grants. However, few would argue that it is a perfect process. Many observers have lamented that fraudulent or flawed results still reach the pages of peer-reviewed journals. Others complain that the peer review system favors established ideas and institutions and stifles scientific innovation.

In 2014 Michael Eisen, a cofounder of the publisher Public Library of Science (PLOS), told the *Wall Street Journal* that scientists and nonscientists need to discard the notion "that peer review of any kind at any journal means that a work of science is correct. What it means is that a few (1–4) people read it over and didn't see any major problems."[11]

Another drawback with the current peer review system is that the work reviewers put in generally does not count toward tenure or promotion. Overburdened scientists face little incentive to write long, careful, and detailed reports that go beyond discharging their minimum duty as good scientific citizens.

The shift to online publication and reading seems to suggest alternative methods for vetting articles, such as allowing scientists to post comments about those they read. Physicists have long relied on the non-peer-reviewed arXiv.org to find the latest publications in their field, although readers may regard a paper posted to the arXiv but never published in a journal as somewhat questionable.

Other journals have been experimenting with slightly altered peer review systems. *PLOS One*, a well-known open access journal, instructs its referees to judge only the quality of the science in the paper, not the work's perceived importance or impact. The reasoning behind *PLOS One*'s policy is that working scientists will determine which papers are most important after publication. Another journal, *eLife*, puts referees and editors in communication with each other to arrive at a single joint decision on a paper's future, rather than sending authors multiple reports that might disagree wildly with one another.

As the scientific community considers peer review's future, it may be instructive to consider its past. We often speak of refereeing as something that has been a stable and unchanging part of science ever since the age of Isaac Newton, but peer review's story is both shorter and more complex than we often assume. It is also littered with criticism. As early as 1845, the scientific referee was described as "full of envy, hatred, malice, and all uncharitableness."[3] Complaints about reviewer uselessness and bias, in other words, are hardly new.

It also seems significant that refereeing procedures were not initially developed to detect fraud or to ensure the accuracy of scientific claims. Whewell thought referee reports would spur scientific discussion, and scientific societies adopted refereeing to ensure that nothing obviously embarrassing reached print. Authors, not referees, were responsible for the contents of their papers. It was not until the 20th century that anyone thought a referee should be responsible for the quality of the scientific literature, and not until the Cold War that something had to be peer-reviewed to be seen as scientifically legitimate.

Peer review's role in the scientific community has never been static. Its form and purpose have been shaped and reshaped according to what scientists needed from the practice—whether it was credibility for a scientific society, improvements in the scientific literature, or assurances to public funders that their money was being spent responsibly. If scientists are to tranform peer review's future, they must consider what purpose they want it to serve—and whether that purpose can indeed be fulfilled by reports from two or more referees.

REFERENCES

1. H. Zuckerman, R. Merton, *Minerva* **9**, 66 (1971).
2. A. Fyfe, J. McDougall-Waters, N. Moxham, *Notes Rec.: R. Soc. J. Hist. Sci.* **69**, 227 (2015).
3. A. Csiszar, *Nature* **532**, 306 (2016).
4. M. Baldwin, in *The Age of Scientific Naturalism: Tyndall and His Contemporaries*, B. Lightman, M. Reidy, eds., Pickering and Chatto (2014), p. 171.
5. I. Clarke, *Isis* **106**, 70 (2015).
6. R. Lalli, *Notes Rec.: R. Soc. J. Hist. Sci.* **70**, 151 (2016).
7. R. Kohler, *Partners in Science: Foundations and Natural Scientists, 1900–1945*, U. Chicago Press (1991), p. 68.
8. I. H. Page, *J. Am. Med. Assoc.* **225**, 1240 (1973).
9. D. Kaiser, *Nature* **505**, 153 (2014).
10. J. Conlan to H. G. Stever, 15 May 1975, in National Science Foundation Peer Review, Special Oversight Hearings, US House of Representatives, p. 21.
11. M. Eisen, quoted in H. Campbell, *Wall Street Journal*, 13 July 2014, http://www.wsj.com/articles/hank-campbell-the-corruption-of-peer-review-is-harming-scientific-credibility-1405290747. PT