THE GENERATION OF KNOWLEDGE · TRUST, QUALITY AND THE FUTURE

Peer Review, Verification and Quality Control of Knowledge

How peer review developed, what evidence shows about its strengths and weaknesses, why it cannot catch every error, and how replication, data checking, preregistration, correction and professional standards add further layers of quality control for research of every kind.

ModQu research article · 2,142 words · 22 references (APA 7th edition)

Every claim to knowledge needs to be checked. Researchers make mistakes in design, calculation and interpretation; they may overstate their findings or, rarely, invent them. Societies that rely on research therefore need ways of catching errors before they spread and correcting them once they are found. The best-known of these is **peer review**, the evaluation of research by other experts before it is published. Peer review is often described as the gold standard of quality control. This article examines how it developed, what evidence shows about how well it works, and the other checks that together make up the quality control of knowledge.

The origins of peer review

Learned journals began in the seventeenth century. Henry Oldenburg, the first secretary of the Royal Society of London, launched the *Philosophical Transactions* in 1665 and relied on members of the Society to advise on what should appear. Yet formal refereeing, in which external experts write reports on submitted manuscripts, developed much later. Csiszar (2016) traced the practice to the 1830s, when members of the Royal Society proposed that reports on papers be written and, at first, published. The system spread unevenly through the nineteenth and early twentieth centuries, and many journals continued to rely mainly on the judgement of editors.

Baldwin (2018) showed that the term "peer review" and the idea that it is essential to scientific legitimacy became common only in the second half of the twentieth century, particularly in the United States. In the 1970s, debates in the United States Congress about the accountability of publicly funded science led scientists to present peer review as the mechanism that guaranteed the quality and integrity of research. Some leading journals adopted systematic external refereeing only in this period.

Zuckerman and Merton (1971) offered an early sociological study of refereeing. They examined the records of the physics journal *Physical Review* and found that the referee system helped to maintain standards by giving authority to judgements made by recognised experts. They also noted that the system depended on trust and on the willingness of scientists to give their time, unpaid, for the common good.

What peer review is meant to do

Peer review has several aims. It is meant to assess whether research is original, important and sound; to detect errors and weaknesses; to improve papers through criticism; and to help editors choose what to publish. Smith (2006), a former editor of the *BMJ*, noted that despite its central place in science, peer review had for a long time been studied very little, and argued that it is in large part untested. When it was studied, the results were often disappointing.

How well does it work?

Detecting errors

Several experiments have tested whether reviewers find errors deliberately inserted into manuscripts. Godlee et al. (1998) took a paper about to be published in the *BMJ*, inserted eight weaknesses into it and sent it to more than 400 potential reviewers. Of those who reviewed it, the average reviewer identified only about two of the eight weaknesses, and some found none. Schroter et al. (2008) repeated the approach with three papers, each containing nine major errors, sent to more than 600 reviewers. On average, reviewers detected fewer than three of the nine major errors in the first paper. Training reviewers improved their performance only slightly.

Errors that require access to data are particularly hard to detect, because reviewers usually see only the written report. Nuijten et al. (2016) used an automated program to check statistical results reported in more than 30,000 articles in psychology journals. About half of the articles contained at least one reported result whose figures were internally inconsistent, and about one in eight contained an inconsistency large enough to change whether a result was statistically significant. These errors had passed peer review.

Agreement between reviewers

If peer review measured quality reliably, different reviewers judging the same manuscript would reach similar conclusions. Bornmann et al. (2010) combined the results of 48 studies of agreement between reviewers of journal manuscripts. They found that agreement was low: on average, reviewers agreed only modestly more often than would be expected by chance. Peters and Ceci (1982) carried out a famous test. They resubmitted twelve articles that had recently been published by leading psychology journals, changing the authors' names and institutions to obscure ones. Only three resubmissions were recognised. Of the nine that were not, eight were rejected, often on the grounds of serious methodological flaws, even though the same articles had previously been accepted by the same journals.

Bias

Reviewers are also subject to bias. Mahoney (1977) sent versions of the same manuscript to reviewers, varying only the results. Reviewers rated the methods more highly when the results supported their own theoretical views than when they contradicted them. Lee et al. (2013) reviewed evidence of bias in peer review and found that judgements could be affected by the prestige, institution, nationality, language and gender of authors, as well as by the content of their

conclusions. Tomkins et al. (2017) tested this directly at a computer science conference by assigning some reviewers to see authors' names and others not. Reviewers who knew who the authors were were significantly more likely to recommend acceptance of papers by famous authors and by authors from highly ranked institutions.

Overall effect

Jefferson et al. (2002) systematically reviewed studies of the effects of editorial peer review and concluded that there was little empirical evidence to support its use as a mechanism for ensuring the quality of biomedical research, while noting that the absence of evidence was not evidence that it did not work. Rennie (2016), one of the founders of a series of international congresses on peer review, called for peer review itself to be studied scientifically, so that its practices could be improved on the basis of evidence rather than tradition.

Peer review also places heavy demands on a small group. Kovanis et al. (2016) estimated the distribution of reviewing in biomedical research and found that a minority of researchers carried out most of the reviews: about one in five researchers performed most of the reviewing work.

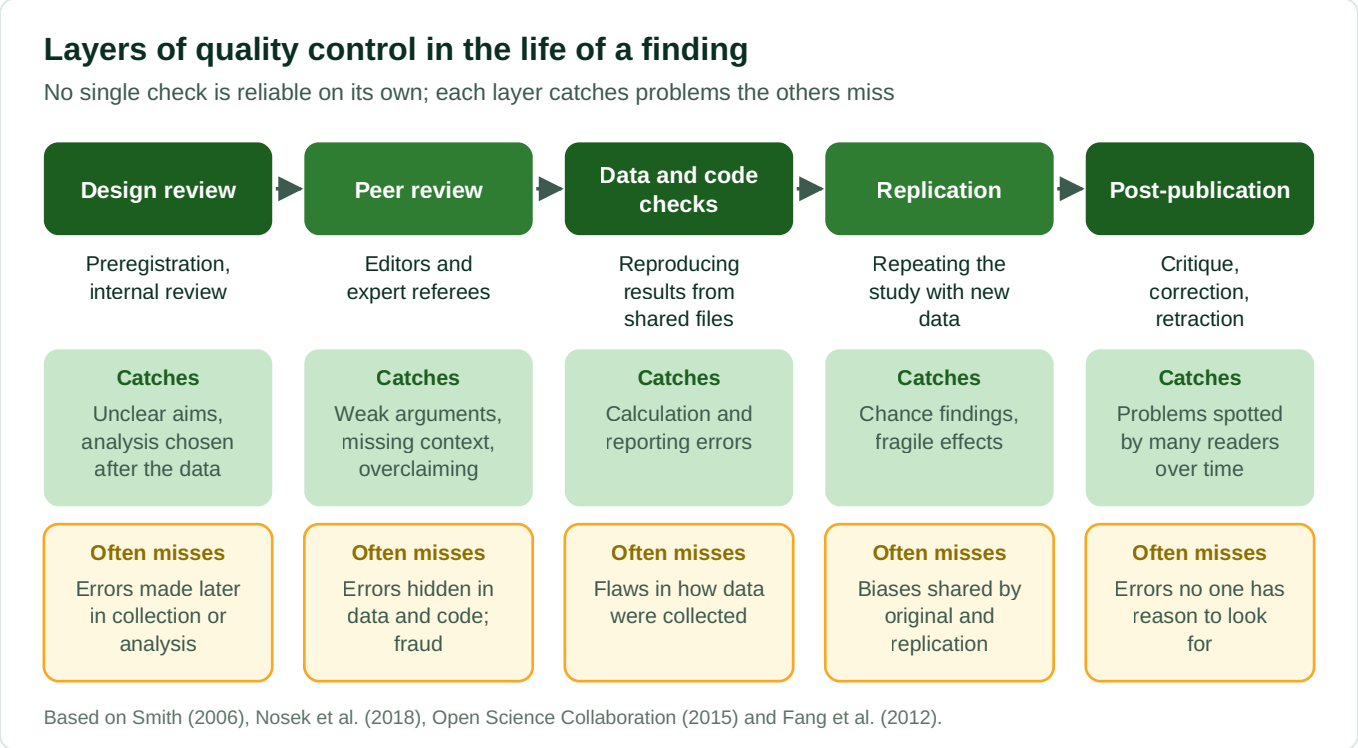


Figure 1. Layers of quality control in the life of a finding. Based on Smith (2006), Nosek et al. (2018), Open Science Collaboration (2015) and Fang et al. (2012); diagram by ModQu.

Beyond peer review: layers of verification

Because peer review is imperfect, the reliability of knowledge depends on several further layers of checking, each of which can catch problems the others miss.

Checking before the study

Preregistration means recording a study's hypotheses, design and analysis plan before the data are collected or examined. Nosek et al. (2018) argued that preregistration helps to distinguish genuine tests of predictions from analyses chosen after seeing the data, which can easily produce false positive results. **Registered reports** go further. Chambers (2013) introduced a format in which journals review the design of a study before it is carried out and commit to publishing the results, whatever they turn out to be, if the study is conducted as planned. This shifts peer review to the stage at which criticism can still improve the research, and it removes the incentive to publish only striking results.

Checking the data

Sharing data and analysis code allows others to reproduce published results. When data are shared, errors that reviewers cannot see may be found later, by readers who rerun the analyses. Some journals now check that the results in an article can be reproduced from the files the authors provide before they publish it.

Replication

Repeating a study with new data is the most direct test of whether a finding is reliable. The Open Science Collaboration (2015) attempted to replicate 100 studies published in three leading psychology journals. Although 97% of the original studies had reported statistically significant results, only 36% of the replications did, and the average size of the effects in the replications was about half that of the originals. Ioannidis (2005) had argued on statistical grounds that, in fields where studies are small, effects are modest and analyses flexible, many published findings are likely to be false. Replication is how such findings are eventually identified.

Correction and retraction

When serious problems are found after publication, journals may publish corrections or retract articles. Fang et al. (2012) examined more than 2,000 retracted articles in the biomedical and life sciences and found that about two-thirds had been retracted because of misconduct, including fraud or suspected fraud, duplicate publication and plagiarism, and about one-fifth because of error. They also found that the proportion of articles retracted because of fraud had increased substantially since the 1970s. Retraction is an important safeguard, but it often comes years after publication, by which time the article may have been widely cited.

Fraud is particularly hard for peer review to detect, because the system rests on trust. Reviewers assume that the data described in a manuscript were collected as reported and that the analyses were carried out as described. They rarely have the means or the time to check. As Zuckerman and Merton (1971) observed, the referee system depends on the integrity of authors as well as the expertise of reviewers. Detecting fabrication usually requires access to the raw data, or the attention of colleagues who notice that results are too consistent, or that they cannot be reproduced.

Open and post-publication review

Ross-Hellauer (2017) reviewed the many forms of **open peer review**, including publishing reviewers' names, publishing their reports alongside articles, and allowing comments from the wider community after publication. Open review aims to make the process more accountable and to give credit to reviewers. Post-publication discussion, on platforms and in journals, means that review does not end at publication but continues as long as a finding is read and used.

Quality control in survey and market research

Most survey and market research is not published in academic journals and is never peer reviewed in the formal sense. Its quality control relies on different layers: internal review of questionnaires and designs; testing of questions before fieldwork; checks on the quality of interviewing and data collection; validation of data and weighting; and review of reports before release. International standards set requirements for these processes. ISO 20252 specifies requirements for organisations that carry out market, opinion and social research, covering matters such as project planning, questionnaire design, fieldwork, data processing and reporting (International Organization for Standardization, 2019).

Clients add a further layer. A client who commissions research and knows the market well may notice findings that do not fit other evidence, and a good research organisation treats such questions as an opportunity to check its work rather than as a challenge to defend against. Independent audits against professional standards provide yet another check, by examining whether an organisation's processes are followed in practice and not only described on paper.

The lessons of research on peer review apply directly to such work. A single reviewer is unlikely to catch every problem, so important findings should be checked in several ways. Errors in data and calculations are the hardest to see in a finished report, so they need checks by someone who works with the data themselves. Reviewers are influenced by whether they like the results, so reviews should focus on methods and should be carried out before results are known where possible. And transparency allows clients and the public to act as a further layer of review, by making it possible for them to see how findings were produced.

Making quality control work

Several principles follow from the evidence:

- **Use several layers.** No single check is reliable. Combine design review, data checks, peer review, replication and openness to later scrutiny.
- **Review designs, not just results.** Criticism is most useful when it can still change a study, and reviews of design are less influenced by whether the results are welcome.
- **Check the data, not just the report.** Many errors live in the data and the code, where reviewers of written reports cannot see them.
- **Make work open to checking.** Sharing methods, materials and, where possible, data allows errors to be found after publication.
- **Correct openly.** Errors will occur. Organisations that acknowledge and correct them promptly earn more trust than those that do not.

Conclusion

Peer review is a central institution of modern knowledge, but it is younger, weaker and more variable than its reputation suggests. Experiments show that reviewers miss many errors, often disagree and are subject to bias. The reliability of knowledge therefore depends on several layers of checking: preregistration and design review before a study, peer review and data checks before publication, and replication, correction and open discussion afterwards. For organisations that produce research outside academic journals, the same principles apply. Quality comes not from any single gate but from a system of checks, each designed to catch what the others miss, and from a willingness to let others examine how findings were made.

References

Baldwin, M. (2018). Scientific autonomy, public accountability, and the rise of "peer review" in the Cold War United States. *Isis*, *109*(3), 538–558. https://doi.org/10.1086/700070

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE*, *5*(12), e14331. https://doi.org/10.1371/journal.pone.0014331

Chambers, C. D. (2013). Registered reports: A new publishing initiative at *Cortex*. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Csiszar, A. (2016). Peer review: Troubled from the start. *Nature*, *532*(7599), 306–308. https://doi.org/10.1038/532306a

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033. https://doi.org/10.1073/pnas.1212247109

Godlee, F., Gale, C. R., & Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial. *JAMA*, *280*(3), 237–240. https://doi.org/10.1001/jama.280.3.237

International Organization for Standardization. (2019). *Market, opinion and social research, including insights and data analytics: Vocabulary and service requirements* (ISO Standard No. 20252:2019). ISO.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jefferson, T., Alderson, P., Wager, E., & Davidoff, F. (2002). Effects of editorial peer review: A systematic review. *JAMA*, *287*(21), 2784–2786. https://doi.org/10.1001/jama.287.21.2784

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS ONE*, *11*(11), e0166387. https://doi.org/10.1371/journal.pone.0166387

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2–17. https://doi.org/10.1002/asi.22784

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161–175. https://doi.org/10.1007/BF01173636

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences, 5*(2), 187–195. https://doi.org/10.1017/S0140525X00011183

Rennie, D. (2016). Let's make peer review scientific. *Nature, 535*(7610), 31–33. https://doi.org/10.1038/535031a

Ross-Hellauer, T. (2017). What is open peer review? A systematic review. *F1000Research, 6*, 588. https://doi.org/10.12688/f1000research.11369.2

Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L., & Smith, R. (2008). What errors do peer reviewers detect, and does training improve their ability to detect them? *Journal of the Royal Society of Medicine, 101*(10), 507–514. https://doi.org/10.1258/jrsm.2008.080062

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine, 99*(4), 178–182. https://doi.org/10.1177/014107680609900414

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences, 114*(48), 12708–12713. https://doi.org/10.1073/pnas.1707323114

Zuckerman, H., & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva, 9*(1), 66–100. https://doi.org/10.1007/BF01553188