In: Albert, M., Schmidtchen, D., Voigt, S. (eds), Scientific Competition. Conferences on New Political Economy 25, Tübingen: Mohr Siebeck 2008.

What Should We Expect from Peer Review?

Comment by
MAX ALBERT and JÜRGEN MECKL

Seidl, Schmidt and Grösche (henceforth, SSG) report on an internet questionnaire study that asked economists about their experiences and satisfaction with the referee process in economics journals. In this comment, we put the paper in the context of scientific competition and ask whether we should be worried by its results.

1 Scientific competition and scientific quality standards

Scientific competition is mainly driven by the quest for status or reputation. Researchers earn status when their contributions are used – and not just cited – by other researchers (see Hull 1988, 283). Status-seeking researchers should use the products of previous research (contained mainly in research papers) if they believe that it will help them to produce output that will, in turn, be used as an input in future research.

In a nutshell, then, research is the production of papers by means of papers. In order to use the results of a paper, researchers must, of course, be aware of the paper and believe it to be relevant to the problems they are working on. Even if these conditions are satisfied, however, they will not use a paper if they consider it (i.e., the results and ideas contained in it) to be of too low a quality.

Scarcity of attention and quality concerns explain the peer review system. Journals try to collect high-quality papers that have something in common, either a specific topic or, in the case of general journals, a potential to interest even off-topic researchers. Some journals are more successful than others in publishing high-quality work, get more attention, and, in turn, attract more submissions of high-quality work. This positive-feedback effect leads to quality rankings among journals.

Ultimately, publishers compete, with the help of their journals, for the attention of the scientific community. There is a hierarchy of delegation, where all agents pursue their own interests. Publishers select and control editors, who, in turn, select and control referees. On each stage, there is competition and moral hazard. Moreover, some deviations from the application of quality standards, like promoting papers sympathetic to an editor's or referee's research, can be viewed as payment in kind for editorial or refereeing services. For this reason and others, we should expect some amount of personal, institutional, or regional favoritism, as well as inner-scientific partisanship, in the selection of papers for publication.

The explanations so far assume given quality standards used by researchers in selecting input for their own research. It is, however, not at all clear how decision-relevant quality standards can become established in the production process we have described above. Even if everybody expects everybody else to use only inputs that satisfy certain standards of high quality, this expectation is not self-fulfilling – unless using high-quality inputs increases one's chances to produce high-quality output. However, the last assumption is quite reasonable for the quality criteria used in science.[1]

Quality standards in science, then, are the outcome of an intertemporal coordination problem among self-interested and forward-looking researchers. As explained above, we expect these quality standards to spill over, if imperfectly, into the peer review process. With perfect coordination, there exists a single quality standard. However, there are several factors working against perfect coordination. First, new methodological arguments can shift the focal point of the coordination game. Second, at a given time, there may exist several candidates for a focal point. Third, researchers have different information about current debates and different reaction speeds. Thus, the coordination process is slow and subject to shocks, working – despite the forward-looking attitude of researchers – like an evolutionary process of short-sighted adaption to one's perception of the current trends, with several standards competing during adjustment. Fourth, quality standards in different research areas (which are defined by relatively low probabilities of use across boundaries) differ, which leads to grey areas where standards are uncertain or disputed. Fifth, quality in science has many dimensions, and weighing these dimensions may be a problem even if there is broad agreement about the dimensions themselves.

Thus, scientific competition involves competition between quality standards. There exists pressure in the direction of harmonizing the standards, but one should not expect perfect harmony, especially with respect to the fine points and when new ideas threaten existing standards. Moreover, quality may be difficult to detect. Even on the basis of common standards, different editors or referees may still come to different conclusions because they prefer different trade-offs between errors of the first and second kind.

2 Quality standards for peer review

SSG suggest that journals only archive papers and hand out quality signals. In their conclusions, they write that peer review in science is a tribunal with decisive influence on individual careers. They then list the claims made in

[1] See Albert (2006) for a model explaining quality standards in scientific competition along these lines.

favor of peer review, which, as far as the journal referee process is concerned, are: quality control, improvement of manuscripts, promotion of innovative research, fostering dissemination of new research, and serving as a means to rank researchers. They note that, due to the publication lag, journals no longer disseminate new research; instead, their main purpose is "to imprint a signal of quality on a scholar's research". This, in their view, requires excellent performance of peer review, especially validity, impartiality, and fairness.

Reliability of the referee process means that different referees come to the same conclusion. Validity of the referee process means that quality judgments report the true quality of the paper. For instance, stories about highly successful papers that were frequently rejected before their eventual publication are often viewed as anecdotal evidence of low validity. Typically, the degree of reliability and validity is measured in terms of correlations between different referee conclusions or between quality judgments and quality.

Validity is ill-defined, however, and reliability is not to be expected when several quality standards compete. Only with (almost) perfect coordination on quality standards, low validity and reliability must be due to imperfections in the peer review process. We do not believe that perfect coordination has been reached in economics.

Impartiality and fairness mean absence of favoritism and, instead, reliance on quality standards, which is of course possible even with competing standards. SSG report regional favoritism. However, consider the case of the Quarterly Journal of Economics. This journal rejects most papers without referee reports but (not mentioned by SSG) publishes many previous NBER working papers. Since NBER papers are already subject to quality control, selecting from them might lower the cost of refereeing without lowering quality. Hence, it may be a sign of efficiency if some journals tap such pools of high-quality papers. Due to the nature of the NBER, this leads to regional favoritism.

However, there is no prima facie case that such practices lead to an unfair and partisan publication system. A combination of journals with different biases can lead to a fair system. Moreover, in judging the quality signals produced by journals, it is easy to adjust for known biases: If you publish a paper at a journal biased against you, it just means that the quality of your paper is probably higher than the journal's average.

Editors usually want the referee to point out possible improvements of the paper. Within limits, this is reasonable since the editor would like to be the paper as good as possible and the referee can produce at least some relevant hints as a by-product of quality control at almost no additional cost. However, referees should not invest much in improving a paper. If they did, this would create incentives for an author to abuse referees as unpaid ghostwriters or conscripted audiences. It is unlikely that this would result in an efficient team effort. Hence, it is perfectly alright if bad papers get sloppy and short reports.

This sets incentives to authors to invest more into their papers (and seek for coauthors by themselves) and makes a better use of the scarce time of the referees, who can concentrate on good papers. It also implies, however, that editors and referees should not necessarily aim at the satisfaction of authors.

Nevertheless, trying to measure imperfections is certainly an excellent idea. The question is whether the data of SSG actually point to imperfections, and hence whether we should be worried about the relatively poor performance of top journals.

We both admit that, independently of each other, we started filling in SSG's questionnaire but gave up since, due to lack of time or access, we could not consult our files. We both were resolved to get to it later, but as these things go, we never did. In line with our experience, SSG admit that those who persevered may have answered the questions from memory, which, as they recognize, may lead to systematic biases. They argue, however, that authors' memory is probably what counts in submission decisions and with respect to author satisfaction. We agree. However, we cannot quite see why author satisfaction should be important, especially if, as SSG find, it depends strongly on the competence and care invested in the reports.

SSG note that top journals receive lower-than-average ratings for their referee processes. Even if this indicated lower-than-average quality, this need not be problematic. Authors submitting a paper face costs in terms of submission fees, rejection risks and decision times, which may be more or less mitigated by the quality of the reports. Top journals overwhelmed by submissions should offer worse terms to authors; they could do this by urging their referees not to waste time on any but the most excellent papers.

However, we can think of two plausible explanations for lower-than-average ratings for top journals' referee processes even if these journals offer average quality. First, authors may just expect more from higher-ranked journals and judge referee processes not in comparison with each other but in comparison with their expectations. Second, authors can make two errors when submitting their papers: aiming too high or aiming too low. If they prefer the first error to the second, papers will on average be submitted too high, which, in the worst case, leads to a rejection based on a single sloppy and short negative report. Papers then trickle down to lower-ranked journals until paper quality matches journal rank. If referee reports get more careful as the gap between journal rank and paper quality shrinks, the trickle-down effect implies that author satisfaction with the referee process increases with falling journal ranks, even if journal policies are all the same. In this context, signalling a large gap between journal rank and paper quality by a sloppy reports offers a further advantage: authors may adjust their self-assessment more quickly, which reduces the number of wasted submissions in the trickle-down process.

References

Albert, M. (2006), "Product Quality in Scientific Competition", *Papers on Strategic Interaction* 6-2006, Max Planck Institute of Economics, Jena.

Hull, D. L. (1988), *Science as a Process,* University of Chicago Press: Chicago and London.