Is unreliability in peer review harmful?

Article *in* Behavioral and Brain Sciences · March 1991

DOI: 10.1017/S0140525X00065936

1 author:

Henry Roediger
Washington University in St. Louis
303 PUBLICATIONS **25,937** CITATIONS

negative results rarely get submitted to begin with. Hence, rejection of papers that meet objective evaluation criteria, but contain negative results, may be an outcome that rarely occurs in actual practice. The point is investigators seem to act as if they have identified a single or small set of measurable characteristics of the reviewers contributing to experimental effects, when the need to maintain the conviction regarding potency of these selected variables may be greater than the evidence to support it.

In short, we are still left with the questions: "What process do reviewers undertake when they perceive information and selectively weigh its importance against arbitrary evaluation criteria?" and, "Which reviewer variables are associated with different review outcomes?" Answers to these questions pertain to independent variables and, as such, would serve to broaden our understanding of what leads to high or low levels of reliability or differential outcomes by discipline subspecialty.

Examining process variables may not be the answer, but peer review, by virtue of being a classification system, involves a process or activity leading to an outcome or decision. To date, the study of process variables in peer review has been largely neglected. Kuhn (1962) reminds us that, in a preparadigmatic state, the "real" solution in any field cannot be negotiated by a representative panel of experts. Cicchetti outlines historical constraints operating within the system of peer review and invites us to break away with some concrete recommendations for change. An additional recommendation should be to examine process variables believed to be associated with levels of expected agreement. The "black box" remains as long as creative efforts to examine and improve the system of peer review are neglected.

Unreliability in peer review harmful?

Henry L. Roediger III

Department of Psychology, Rice University, Houston, TX 77251-1892
Electronic mail: *roediger@ricevm1.rice.edu*

Cicchetti's target article provides an excellent analysis of studies assessing the reliability of peer review in journal and conference submissions and grant proposals. Even the best studies show modest levels of reliability, a fact decried by many who see arbitrariness in the peer review system. The underlying assumption behind the gloom that studies of peer review cast is that the publication (or granting) process would somehow be more accurate and fairer if the reliabilities involved in peer review were improved, say to .70 or .80. To me, this state of affairs seems unlikely to occur under any realistic set of conditions. Furthermore, I remain unconvinced that it would even be desirable, in the long run, for the scientific enterprise, even though it might make life easier for editors and grant administrators. Below I will provide underpinnings for these opinions.

Cognitive psychologists have long been interested in the processes involved in judgment and decision making in complex realms (e.g., hiring decisions, picking stocks, making clinical diagnoses). The literature is replete with findings of poor reliability and validity of human judgments when people, even experts in a field, are faced with complex, multiattribute decisions (e.g., Kahneman et al. 1982; Nisbett & Ross 1980). Given this backdrop, a finding of high reliability in peer review judgments would come as a surprise.

One reason for unreliability in peer review that may not pertain to other areas of judgment concerns how reviewers are selected by editors (see Roediger 1987). I spent five years as editor (and another three as associate editor) of a journal referred to by Cicchetti as a "specific focus journal" (the *Journal of Experimental Psychology, Learning, Memory, and Cognition*). Although perhaps specific in some sense, the topics under consideration seemed broad enough to me: reading, attending, learning, remembering, decision making, judging, problem solving, categorizing, perceptual-motor skill learning, and other topics. As editor, I would skim each submission to assign reviewers. A common scenario would be as follows: The authors of the paper would be examining a particular theory or line of thought about some phenomenon, or they would be contrasting two or more viewpoints. Based on a series of several experiments, they would usually reach some conclusion on the phenomenon in question. As editor, I would try to pick reviewers who would come at the paper from different viewpoints. If the authors eventually concluded that their results supported Theory X, then usually I would have someone associated with Theory X as one reviewer, and someone associated with Theory Y (or some other approach) as another reviewer. If the paper had some fatal flaw (poor reasoning, improper methods, inappropriate statistics, inconsistent results across experiments), both reviewers would probably argue against publication. This is just what Cicchetti shows: Peer reviews are quite consistent on flawed papers.

But suppose the paper did not suffer from any obvious flaws. A typical (but not universal) pattern for such a paper supporting Theory X would be for another proponent of Theory X to evaluate the paper positively, whereas a "Theory Y reviewer" might recommend against publication, suggesting further research. As Cicchetti notes, the reviewers may not even disagree on their assessments of the facts, but rather of the weightings given to them. Of course, these "unreliable" judgments seem perfectly sensible to anyone editing a journal. Further, both reviewers are often right, in the sense that most papers (excluding the truly bad ones weeded out by peer review) have some merits and some demerits to which reviewers can point.

If this scenario is representative, then some unreliability in the peer review system may be occasioned by editors seeking the advice of experts with varying points of view on the topic at issue. This process may occasion unreliability of peer judgment, but probably provides better information to the editor and the authors. If this is one cause of reviewer unreliability, then one way to enhance reliability would be for editors to try to identify reviewers who had in the past consistently agreed or disagreed with the position argued by the author in the manuscript under review and to send the paper only to like-minded reviewers. I assume no one would seriously argue for this proposal, which shows the danger of emphasizing reviewer reliability at the cost of other considerations (such as providing a variety of perspectives).

Finally, consider the neglected issue of the validity of peer review. Can scientists really predict accurately which manuscripts or grant proposals will lead to surer progress in the field? Can *any* reviewer validly discriminate the top 20% of the papers or proposals from the next 20%, which is often the task in the behavioral sciences with their high rejection rates? Given that peer judgments are unreliable, asking questions about validity is even more hazardous, especially since there is likely to be disagreement about the criterion variable. For example, suppose that reviewers or editors were asked to predict the number of cumulative citations over a 10-year period for papers accepted for publication. Would the resulting correlations between predicted and actual citations even approach the modest .2 to .3 we have come to expect from peer review studies? I doubt it.

My skepticism about the outcome of such a study is based in part on informal observations of colleagues discussing controversial papers that have been published and have then shaped the direction of my field (cognitive psychology). Often, years later, one will still hear debates about the original paper, whether or not it should have been accepted, and whether the resulting approach has been worthwhile or a blind alley. If scientists cannot agree, even in retrospect, that heavily cited and important papers were indeed worthy, then what hope do we have of deciding such matters a priori? (See Roediger, 1987,

for an example.) This matter deserves more formal study, but accurate judgments of scientific importance are probably reliable only years after the fact of publication, with the wisdom of hindsight.

In summary, let us simply grant that the peer review system is inherently unreliable, to a great extent. Two reasonable people, both experts in their fields, can look at the same manuscript or grant proposal and reach quite different conclusions about its merit. But if scientists cannot really make valid judgments about such matters (which seems likely, too), then the unreliability may not actually be harmful. Perhaps the randomness introduced into the system is good for it, if even reliable judgments have little validity. If these conclusions are indeed facts, should we be depressed and give up peer review? I don't think so. After all, peer review does function well (a) to eliminate the real "bloopers," and (b) to provide expert opinion to authors, which is often helpful (in my experience). And there seems no reasonable alternative to peer review, no system that would work so well without engendering more problems than it solved.

My recommendation is that editors and grant administrators recognize fully the potential flaws in the peer review system and work around them. In cases of divided opinion, editors may use the heuristic of "when in doubt, accept" (cited by Cicchetti). My view is that, in most fields, the unreliability of peer review does little harm and may do good, assuming that several journals are appropriate outlets for a piece of work. If a paper is rejected by one, the negative reviews can be used as advice for improvement for resubmission elsewhere. Given several outlets, persistent authors, and unreliability in the peer review system, worthy papers will eventually see the light of day, even if not in the outlet of first choice, and at a slight delay.

The situation with regard to grant proposals is less optimistic, mainly because there are fewer sources of funds. A negative evaluation is more likely to mean that the work will not be carried out. Evaluating proposed research seems even more fraught with difficulty than evaluating completed work. One solution would be to follow the Canadian system in which (as I understand it) many researchers are given small seed grants at the beginning of their careers, and then the system rewards those who carry forward successful research programs. Perhaps in awarding grants we should place greater emphasis on the applicant's past record of research and less emphasis on the writing of a promissory note (in the form of a proposal) for future work. This recommendation assumes that greater reliability and validity can be exhibited by judges in evaluating research records than in evaluating research proposals, a topic that awaits future investigation.

Some indices of the reliability of peer review

Robert Rosenthal

Department of Psychology, Harvard University, Cambridge, MA 02138

Cicchetti has performed an important service to the several sciences by summarizing what is known about the reliability of peer review. Given the impact of *Behavioral and Brain Sciences* target articles, it is likely that his paper will encourage further research and further thinking about the reliability of peer review. Its impact may also extend to the encouragement of the use of various indices of reliability of judgments. It is therefore of special importance to be clear about several issues relevant to the choice of indices of reliability. The purpose of this commentary is to suggest some friendly amendments to the evaluations of several indices of reliability referred to or used in the target article.

Three more-information-efficient indices. Three of these indices of reliability are very information-efficient in the sense that they use all the information available and give a single, unequivocal, focused, single *df*, easy to interpret index of magnitude of relationship (Rosenthal 1987; Rosenthal & Rosnow 1985; Rosenthal & Rubin 1982). These are the Pearson R, the intraclass correlation, and Cohen's (1960) *kappa* applied to the 2 × 2 table. Especially for that case of the intra-class *r* in which each rater judges all stimuli, all three of these indices are equivalent to product-moment correlations. Indeed, Fisher developed the intraclass R to be able to apply Pearson R to twin-data in which it would be arbitrary to designate either twin as the X or the Y. Fisher originally dealt with this situation by listing each twin pair twice, once as XY and once as YX (Snedecor & Cochran 1967). Cohen's *kappa* in the 2 × 2 case is equivalent to the Pearson R in its 0,1 incarnation, an R sometimes referred to as the *phi* coefficient. In short, these three indices all tell essentially the same story, so it seems inconsistent to label the intraclass R as appropriate (Cicchetti, sect. 3.3) and the Pearson R, from which the intraclass is derived, as inappropriate (sect. 3.4). The Pearson R "ignores the extent to which given pairs of reviewers disagree on any single evaluation" precisely to the same degree that the intraclass R (Model II) does. If it is desired that absolute differences in raters' judgments be considered, intra-class R Model I can be used.

Incidentally, it should be noted that the equations given for intraclass R Models I and II are *not* standard. [Corrected in printed version, *Ed.*] The definitional equation (Guilford 1954; Snedecor & Cochran 1980) for Model I is:

$$R_I = \frac{MSS - MSE}{MSS + (r-1)MSE} \text{ (Model I)} \qquad (1)$$

where MSE pools raters and residual mean squares, whereas for Model II it is:

$$R_I = \frac{MSS - MS(RS)}{MSS + (R-1)MS(RS)} \qquad (2)$$

where MS(RS) is the residual mean square only.

Three less-information-efficient indices. Three of these indices are usually less information-efficient, sometimes very much so: rates of agreement (sect. 4.7), χ^2 (sect. 4.7), and *kappa* for tables larger than 2 × 2 in which *kappa* has not been weighted to become effectively a focused, single *df*, effect-size estimate. Rates of agreement suffer from the problem that nearly perfect agreement can occur with actual R near zero (Rosenthal 1984; 1987). χ^2 suffers from its being a product of R^2 and *N* so that it is driven up not only by increases in reliability but by increases in sample size as well (Rosenthal & Rosnow 1984). *Kappa* on *df* > 1 suffers from the same problem as any other diffuse or omnibus procedure, namely, that whatever its size, we cannot tell where the agreements or disagreements arise unless *kappa* approaches unity so that there are no disagreements (see Fleiss 1981).

An example. Because of the valuable information provided in Cicchetti's Note 6 we essentially had the raw data for the *Journal of Abnormal Psychology* set of 1,313 articles and the ratings of two referees for each article. Each referee could use 4 levels of evaluation, so the data could be cast into a 4 × 4 table of agreement. The product moment R using linear contrast scores of −3, −1, +1, +3 for the 4 levels of evaluation was .189. The corresponding *kappa* was .108. When the 4 × 4 table was condensed to a 2 × 2 table, the product moment R was identical to *kappa;* both were .145, illustrating both the loss of information in going from 4 levels to 2 and the equivalence of R and *kappa* for a 2 × 2 table (*df*=1).

The same data of Note 6 can be used to address an additional issue. In section 4.7, agreement rates had been used to assess the question of whether reviewers agree more on decisions to reject than on those to accept manuscripts. Table 5 of the target article shows agreement levels of 44% on decisions to accept and 70% on decisions to reject for the data on the *Journal of Abnormal Psychology*. Using *kappa* or Pearson R, however,