

the validity of a judgment of an article to be about .70. (We know from the data presented by Cicchetti that it is actually much lower.) They then showed that, given this assumption and the fact that only about 15% of submitted articles are published, almost as many papers that "truly" deserve to be published will be rejected as will be accepted. Given the real reliability of judgments, it is probable that more papers that "truly" deserve to be published are rejected than accepted. Even under the current system most sociologists believe that the bulk of the articles published in the leading journals are of poor quality and of little interest. As a result of low levels of consensus, these feelings are probably common in all scientific fields.

Additional evidence against the view that lower rejection rates would reduce quality are the findings of Garvey et al. (1970) that a significant portion of articles published in "core" social science journals had previously been rejected by one or more journals. I am not suggesting that journals in fields like sociology publish all or even a majority of articles submitted. I am suggesting, however, that they gradually increase the proportion of submissions published.

If low levels of peer review reliability are caused by a lack of consensus, is there anything we can do to improve the reliability? Cicchetti suggests increasing the number of reviewers. Because the selected reviewers are essentially a small sample from the population of eligible reviewers, the larger this sample is, the more likely it is that the sample statistic (the mean rating of the reviewers) will approximate the population statistic (the mean rating we would obtain if all eligible reviewers participated in the evaluation process). But this would not necessarily help us make a "better" decision about whether to publish the paper. Would we want to make publication contingent on the relative proportion of the population who would recommend publication? Following such a policy, innovative work that goes against current ways of thinking might not be published.

The situation for the distribution of grants is different. Here there is a limited amount of money to be distributed and the scientific community does not have the power to increase the size of this pool. It is therefore necessary to be able to give priority to submitted proposals. Because of the inherent lack of consensus on research-frontier science, it is *inevitable* that many worthwhile proposals will be rejected and that some proposals of little value will be funded. This was the major finding of my peer review study (Cole et al. 1981). The problem here is the failure to recognize lack of consensus as the reality we must deal with. If we recognize this, there are a number of steps we can take to reduce (but never eliminate) the impact of random factors on the allocation of grant funds. The most important step is for such funding agencies as the National Science Foundation to recognize publicly that many rejected proposals are as worthy of funding as many accepted proposals. If they were to do this, they could set up an appeals procedure in which appeals would be treated sympathetically instead of as the complaints of "cranks." If such an appeals system were functioning properly, a significant portion of appeals should result in the awarding of grants, even at the expense of reducing the amount of funds available for the next round of new proposals.

In summary, the data suggest that the reliability of peer review can be improved by increasing the number of reviewers, but that given the inherent lack of consensus in science, this will not help solve the problem. Lack of consensus must be recognized as a reality; we can then introduce policies to minimize its effect on the development of knowledge and the careers of individual scientists.

Unreliable peer review: Causes and cures of human misery

Andrew M. Colman

Department of Psychology, University of Leicester, Leicester LE1 7RH, England

According to John Ziman (1968), the referee involved in the process of peer review is "the linchpin about which the whole business of science is pivoted" (p. 111). But, as the same commentator pointed out, "the most vexed and contentious topic in the business of scientific communication is the role of the referees, their danger as censors of new ideas, the procedures for appeal against their decisions, and so on" (Ziman 1976, p. 104). Cicchetti has marshalled a considerable body of evidence that shows referees' evaluations of scientific documents to be lamentably unreliable, and the topic is more vexed and contentious than ever. I shall confine my commentary to two possible remedies, only one of which was discussed by Cicchetti and to what I see as the root cause of the problem.

Cicchetti summarized several arguments for and against blind review, which is designed to eliminate the effect of referee bias toward individual authors or institutions. The debate about blind review is somewhat scholastic, in my view, because there is little evidence to show that this kind of crude referee bias is a significant factor. Even Peters & Ceci's (1982) well-known data on the fate of published articles resubmitted with fictitious authors and institutional affiliations can best be explained in terms of random error without invoking referee bias, and Occam's razor bids us reject the bias hypothesis in favor of the simpler random error null hypothesis (Colman 1982b). One important point that is worth adding to Cicchetti's remarks about blind review is that a grant applicant's past record of research could with some justification be considered a significant factor in predicting the likely outcome of any new award that the applicant might receive and ought, perhaps, to be taken into account by the referees. Blind review entails the deliberate concealment of this potentially relevant information.

The use of multiple (more than two) independent referees is not a remedy that appeals to me, although it has its supporters, including *Behavioral and Brain Sciences (BBS)*. My reservations about multiple refereeing are based partly on the findings of research in social psychology and partly on commonsense considerations. Experimental evidence suggests that the involvement of several referees would produce a well-documented phenomenon characterized by a decrease in individual effort, called "social loafing" (Latané et al. 1979), and would also encourage diffusion of responsibility (Darley & Latané 1968). Both of these phenomena are likely to undermine the general quality, and hence the reliability of referees' reports. People tend to apply themselves more diligently and to behave with greater social responsibility when they feel that their input is important and that their efforts are likely to be instrumental in influencing outcomes (Colman 1982a, Chapter 9), but in the peer review process this feeling of instrumentality is bound to be an inverse function of the number of referees.

Second, multiple refereeing tends to increase the nonproductive component of scientists' workloads. The volume of material that requires refereeing is already daunting: Some 40,000 scientific journals currently publish approximately two new articles per minute (Mahoney 1982). Refereeing manuscripts and grant applications is difficult, time-consuming, and generally unrewarding work. What is worse, conscientious refereeing is an ultimately self-defeating activity because it tends to generate ever-increasing workloads. Conscientious referees find their popularity with editors increasing and more and more manuscripts landing on their desks long after their own research has begun to suffer, until they cannot even cope with their refereeing work efficiently. It is clear that the reinforcement structure of science punishes virtuous behavior and rewards sloppy,

superficial, casual, thoughtless, insensitive, inefficient, and therefore unreliable refereeing. Any increase in the number of manuscripts and grant applications that scientists are called upon to referee as a result of the introduction of multiple refereeing is likely to exacerbate the malaise and eat further into the time they should be devoting to doing science. In summary, multiple refereeing seems both counterproductive and gratuitously labor intensive.

The most important safeguard, not even mentioned by Cicchetti, against bias and incompetence on the part of referees, would be an automatic author's right of reply to referees' criticisms. Under the peer review system in its conventional form, authors of scientific papers and grant applicants often find themselves in a Kafkaesque situation analogous to that of a person prosecuted and condemned in a court of law without any right of defense. Sometimes, scientific work is rejected on grounds that the authors believe, rightly or wrongly, to be demonstrably invalid. In my view, before reaching a final judgment, editors and those who award research grants, should routinely solicit the authors' responses to the referees' criticisms, and if necessary the referees' replies to the authors responses, until a clear resolution of the issue emerges. It may sometimes be necessary to submit the original manuscript together with the referees' reports, the authors' responses, and the referees' replies to a qualified independent arbiter before a fair decision can be reached. This procedure was implemented when I was editor of *Current Psychology: Research & Reviews*. I found it immensely helpful, and there is no doubt that at the very least it increased the face validity of the manuscript evaluation process. Although this is clearly no panacea, I feel sure that if it were generally implemented, it would make authors, editors, and even referees feel happier about the peer review process. I am reasonably optimistic that the reliability and validity of the process would correspondingly improve.

Evaluating scholarly works: How many reviewers? How much anonymity?

John D. Cone

School of Human Behavior, United States International University, 10455 Pomerado Road, San Diego, CA 92131

Cicchetti documents fairly convincingly that: Researchers agree on the "normative" criteria to apply in judging a paper's scholarly worthiness; they disagree on the application of these criteria to given manuscripts and on the publishability of given papers. Cicchetti also asserts the commonly held belief that "levels of interreferee agreement are substantially higher for journals in the physical sciences."

It would be of some interest to know more about interreferee agreement on judgments about manuscripts submitted to physical science journals. Conducting such studies would require care in controlling certain likely confounding factors, however. For example, in comparing agreement for relatively focused journals (e.g., *Nuclear Physics*, *Condensed Matter*) with relatively more diffuse ones (e.g., *General Physics*, *Particles and Fields*), the number of reviewers would need to be held constant. The common belief that reviewing is more reliable in the physical sciences may stem from the greater use of the single initial reviewer system in the physical sciences. It might be that such a system yields higher acceptance rates. This is because higher acceptance rates might be prevalent when less critical reviewing is undertaken. The basis for this reasoning is the assumption that reviewing is at least partially under audience control. If so, the mere presence of another reviewer could lead to more critical reviews and, in turn, to higher rates of rejection. If audience control is a factor, the "partial anonymity of the

reviewer case" should lead to greater rejection rates than the "total anonymity case." It would be interesting to investigate this prediction.

The well-designed study would vary both the number of reviewers and the level of anonymity and use acceptance rates and interreviewer reliability as its dependent variables. My prediction would be for lowest acceptance and highest agreement rates for the multiple reviewers subjected to only partial anonymity, because reviewers who know that others are performing the same task and that agreement is to be checked will tend to be more conscientious. The increased vigilance associated with such reviewing will turn up more concerns about aspects of the submission and lead to a greater probability of rejecting it. Related data on this issue are available in the direct observation assessment literature, where it has been shown that observers who know they are being checked for agreement tend to be more reliable and to record more of the behavior being observed (e.g., Romanczyk et al. 1973).

Cicchetti provides no evidence for his assertion that "manuscripts requiring more than one reviewer tend to be those that are problematic." It could be that using multiple reviewers merely turns up more problems. This being the case, the use of more than one reviewer should be associated with lower rates of acceptance, as Cicchetti's Table 3 indeed reveals.

An undiscussed variable in the Cicchetti review is submission rate. Journals with fewer submissions might be expected to have higher rates of acceptance, a supposition given some support by the data in Table 3. In behavioral psychology the proliferation of journals has led to correspondingly fewer submissions to any one journal. Associated rates of acceptance have therefore gone up. Research on reviewer reliability needs to take this into account. A journal with relatively fewer submissions (e.g., the *Nuclear Physics* section of *Physical Review*) will tend to have higher acceptance rates than one with two or three times the submissions (e.g., *General Physics*, *Condensed Matter*). Acceptance rates or judgments and their reliability should be compared for journals with equivalent submission rates; this would help control for any tendency toward leniency just to keep the pages filled.

Another variable worthy of study is the acceptance/rejection base rate of a particular journal and the reviewers' knowledge thereof. While these can be adequately controlled with appropriate statistics (e.g., $kappa$, R_i) in the computation of agreement, the reviewers' judgments themselves may be partly determined by their knowledge of base-rate acceptance levels for the particular journal. The base-rate problem has long been studied in clinical decision making in psychology; it is well established that clinicians' "hit" rates for particular diagnoses vary with the base rates of the diagnoses in the population. If agreement with the editor's ultimate decision is viewed as a "hit," and something reviewers strive to accomplish, base rates would need to be controlled when comparing acceptance and, possibly, reviewer agreements across journals.

Finally, while I am generally sympathetic to Cicchetti's observations and recommendations and found his review a good stimulus for some of my own verbal behavior, I did puzzle over his summary of Mahoney's studies. He asserts that the best available evidence shows that reviewers apply subjective criteria in judging scholarly submissions. As support for this assertion he points to the fact that manuscripts were "accepted or rejected on the basis of whether the findings were positive, negative, or mixed, rather than on the basis of their worthiness." It is not clear what is subjective about this. Indeed, basing decisions on outcome should be one of the more *objective* approaches to the process. Moreover, contrary to Cicchetti, it should have a *positive* influence on the reliability and validity of peer review. After all, at least in the behavioral sciences, it is not obvious that there is all that much that is worthy about a study that fails to reject the null hypothesis.