

magnitude of its contribution relative to the subarea in question. Of these, the first and third decisions are primarily those of the editor, informed to some degree by the comments of the referees. The second, the rooting out of ambiguities and error, is, I contend, the major role of the referees of articles.

Neither the referees, as some commentators apparently think should be the case (Laming, 1991, and Schönemann, 1991, being perhaps the most extreme examples, but also Cicchetti himself), nor the editor should attempt to decide whether or not a line of research is, to use Schönemann's phrase describing my area, "absurd and sterile." What is "absurd and sterile" to one group of scientists may well seem insightful and valuable to another, and sometimes these evaluations change substantially over time. Psychologists hardly need to be reminded of the very sharp shifts that have taken place within the past 50 years. Topics that filled the journals and were deemed highly significant at the time are often quite forgotten – I think of Clark Hull and also of verbal learning – and yet for all I know they may be resurrected by some new development. To go farther afield, what evaluation would hard-headed experimental physicists have given, at the time, to such mathematical preoccupations as the theory of finite groups, Minkowskian geometry, or the theory of numbers? Harmless perhaps, even absurd, but surely not very significant in the larger nineteenth-century scientific setting. Yet in the twentieth century they proved essential to the development of, respectively, particle physics, general relativity, and modern coding theory.

Implicit in these discussions of peer review is a concept of validity that I contend is largely imaginary. What can possibly be the criterion measure? People simply disagree about the importance of a body of work, and not always from ignorance (although often that is a major factor); and we all know that virtually everyone is, at best, mediocre in predicting the course of future developments, whether political, economic, or scientific. The checkered history of pronouncements about scientific and technological developments a decade or, worse, a century away should serve as a caution. Nor do objective measures, such as the number of now unread articles, provide much help, if for no other reason than that as ideas become incorporated into the body of "standard" science, the need to read in the distant past is limited largely to those motivated by personal pleasure and to historians of science. To be sure, much work is simply discarded as apparently misguided and fails to become a part of the present standard science, but do we seriously believe that by some added training of referees these blind alleys could have been spotted and such articles kept from publication? To rule something out just because a handful of people claim now to know it has no scientific value is a strategy designed to limit, not improve, science.

Why, then, do we typically use more than one referee on manuscripts, and do we have any reason to hope that their judgments can become highly reliable? It is certainly true that if referees are selected for their knowledge of the area of the manuscript and if each referee were able to detect any failing of the paper – inadequate scholarship, unclear writing, error of design, error of analysis, error of reasoning, or error of interpretation – then highly reliable judgments should occur. And presumably we would stop using more than one referee. But who among us claims to be able to catch all of the errors, our own included, that pass before our eyes? One hopes to benefit by exploiting the individual differences in referees to reduce the likelihood that grossly erroneous or misleading findings are published.

Thus, by its very nature, the process should result in low reliability in order to achieve a better filter against errors. This, of course, is the reason experienced editors (e.g., Bailar, Kiesler, and Roediger) select referees with different special abilities – to catch different types of errors. To be sure, referees fail to uncover all errors, but I do not see how trying to increase reliability is going to be of much value in this endeavor unless we

Reliability is neither to be expected nor desired in peer review

R. Duncan Luce

Department of Cognitive Science, University of California, Irvine, CA 92717
 Electronic mail: rdluce@uci.bitnet

Cicchetti's (1991a) target article and many of the accompanying commentaries seem to embody a more-or-less implicit assumption concerning the purpose(s) of the peer review process. Their assumption strikes me as very surprising. Because peer review is used with both publication and grant decisions, and because both involve to some degree the allocation of scarce resources (although the evidence indicates that publication resources are not that scarce), there seems to be a strong tendency to lump them together as a single issue. I wish to argue that they are substantially different problems, and that neither is the problem that Cicchetti and many of the commentators assume it to be. Among the commentators who do not accept Cicchetti's assumption are Bailar (1991), Eckberg (1991), Kiesler (1991), Kraemer (1991), Roediger (1991), and Stricker (1991); in varying degrees they hold views similar to mine.

For me, the overall purpose of the journal editorial process, including refereeing, is, first, to decide whether the article is suited to the areas covered by the journal; second, to determine whether it is an adequate presentation of its material and free from technical errors of various types; and third, to try to establish a priority rating that is based on an evaluation of the

mean to improve appreciably each of our abilities to spot error. It seems highly unlikely that editors alone will get very far in that effort. High reliability can also be achieved by using only clones of imperfect referees, but that hardly seems helpful. I do not mean to suggest that no effort should be made to improve refereeing, but if the measure of success is increased reliability then we are on a road to trouble, not improvement.

To me a better measure of the failure of the refereeing process would be the fraction of articles that passed the filter only to be found, after publication, to be deeply flawed in a fashion that, in principle, could have been spotted at the time of refereeing. If that fraction is large, then the filter is failing. Of course, this proportion must be pitted against the fraction of articles that are rejected as flawed, only later to be recognized as correct. These measures are probably far too difficult to collect to be practical, but it is they, not reliability, that seem to me to be relevant.

The issue of grant evaluations is a different, somewhat more vexing, matter. Of course, one aspect is, like that of journal refereeing, the exposure of error. That tends to be the easiest thing to do, especially since that is what we are highly attuned to, and it certainly dominates a lot of grant evaluation. Indeed, the staffs of funding agencies bemoan the way psychologists damn each other on grounds of purported methodological error – we are by a factor of two or more worse in this regard than are physical and biological scientists and even other social scientists. One difficulty in the error-seeking approach is that the format of grant applications typically does not lend itself to the detail needed for such evaluations to be successful. Nor, given the very tentative nature of such proposals, is the pursuit of methodological error necessarily a reasonable goal. Rather, I think we should be contemplating the allocation of scarce resources among scientists. We are engaged in placing bets on lines of research that we think have the best chance of achieving some not very clearly specified payoff. We are participating in that most difficult of social decisions, the setting of priorities in situations where the outcome is inherently highly uncertain. Here we really are stumbling over the issue of what criteria we are trying to satisfy and how to predict the degree to which a given project will meet them. Since we seem quite unable to agree upon a concept of validity and surely we are notoriously poor at making predictions, high reliability seems most unlikely unless it is induced artificially.

Because I have little or no faith in our judgments about specific topics, I think the most reasonable thing to do is to bet on people. They seem to establish track records, although we are all fully aware (often as a result of tenure decisions) of our failures to predict people's long-term performance. We seem to be able to a degree to dissociate our evaluations of scientific creativity, originality, and accuracy from our judgments about the long-term worth of specific projects. For example, I remain skeptical of the attempts to devise unified cognitive theories of the type developed by Newell (1990), but there was absolutely no doubt about his vast intellectual abilities and so of the importance of supporting his research. One can distinguish quality of mind and research ability without necessarily engaging in long-term evaluations of anticipated significance of specific enterprises.

Is not the major purpose of peer review in the case of grants to maintain fair and dispassionate judgments of the scientific qualifications and expectations of the proposers, judgments that are supposed to be uncontaminated by issues of bias, patronage, friendship, and the like? I contrast this with an attempt to decide what will or will not turn out in the long run to be scientifically significant, something we should by now accept that we are very poor at doing. Who in the nineteenth century would ever have anticipated the long-range impact of those who tinkered with electrical and magnetic phenomena?

Perhaps we should recognize that some ideal goals, lovely as they may seem in the abstract, are simply beyond our reach, and that it is better to focus on the possible. We may well be dealing

with an example where the best, especially a singularly ill-understood best, is the enemy of the good. For journals, I contend that the primary goal of refereeing is the expulsion of error and ambiguity, and the fact that no referee is close to perfect in carrying this out coupled with the fact that referees differ substantially in their expertise leads automatically, and desirably, to low reliabilities. For grants, the purpose is to establish priorities of funding and that should, I believe, be based primarily on the quality of mind and skills of the investigator(s), and the procedure should be deemed to be fair and unbiased. In neither case should there be an attempt to judge the significance of the subfield, although one may try to judge the significance of the work within the subfield. Evaluations of subfield significance are established in other forums – the intellectual structure used by funding agencies with the corresponding allocation of funds and the types of scientists hired by departments being two of the more important ones. To try to use peer review as a way to establish priorities seems to me doomed to failure; it should be strongly resisted so peer review can continue to serve its important functions: the search for error, the establishment of some order by quality within a subfield for publications, and the allocation of resources among investigators for grants.