it INFORMATIKA ÉS TUDOMÁNYELEMZÉS

BRAUN TIBOR • SCHUBERT ANDRÁS
SZERKESZTŐK

Szakértői bírálat (peer review) a tudományos kutatásban

Válogatott tanulmányok a téma szakirodalmából

MTA 1826 K

BUDAPEST • 1993

Braun Tibor és Schubert András
szerkesztők

SZAKÉRTŐI BÍRÁLAT (PEER REVIEW) A TUDOMÁNYOS KUTATÁSBAN

Válogatott tanulmányok a téma szakirodalmából

A MAGYAR TUDOMÁNYOS
AKADÉMIA
KÖNYVTÁRÁNAK
INFORMATIKAI ÉS
TUDOMÁNYELEMZÉSI
SOROZATA

7.

Sorozatszerkesztő:
Braun Tibor

Tartalomjegyzék

Braun Tibor · Schubert András
szerkesztők

SZAKÉRTŐI BÍRÁLAT (PEER REVIEW) A TUDOMÁNYOS KUTATÁSBAN

Válogatott tanulmányok a téma szakirodalmából

Magyar Tudományos Akadémia
Könyvtára
Budapest · 1993

Magyar Tudományos Akadémia Könyvtára
Library of the Hungarian Academy of Sciences

ISSN 0230-4619
ISBN 963 7302 867

Felelős kiadó: az MTA Könyvtára főigazgatója

Alak: B/5 – Terjedelem: 17,2 (A/5) ív
Megjelenés: 1993 – Példányszám: 500

Nyomdai előkésztés: W&T Consulting Kft.
Készült az MTA Könyvtára házi sokszorosító részlegében

Előszó

Láng István, akadémikus

A Magyar Tudományos Akadémia vezetői a hetvenes évek második felében felismerték, hogy mielőbb szükséges megteremteni a szellemi és műszaki feltételeket ahhoz, hogy a magyar tudományos kutatás publikációs tevékenységét és annak eredményességét mennyiségi és minőségi mutatókkal lehessen jellemezni. Ennek érdekében az MTA Könyvtára, és azon belül az Informatikai Igazgatóság felépítette a *Magyar Természettudományos Alapkutás Publikációs Adatbankját*, továbbá megteremtette a pénzügyi lehetőségeket az Institute for Scientific Information (USA, Philadelphia) által kiadott *Science Citation Index* gépi szakirodalom-figyelő szolgáltatás hazai adaptációjához. Közép-Kelet-Európában Magyarországon jött létre elsőnek ez az új, számítógépes rendszer, amely a szakirodalmi információigények magas szintű ellátása mellett, nagy hatással volt a tudományos kutatók publikációs stratégiájára. A neves nemzetközi folyóiratokban való publikálás igénye minőségi változást hozott és ez a váltás nálunk előbb következett be, mint a többi közép- és kelet-európai országban. Ez jól tükröződik a *Nature* 1993. január 14-i számában (361. kötet, 104. old., 1993) közölt ábrán, ahol ezen országok tudományos közleményeinek átlagos idézettségét mutatják be 1981 és 1990 között. A vizsgált időszakban a magyar publikációs mutatók meghaladják a többi országét. Ez az eredmény elsősorban a kellő időben felismert új publikációs stratégia megvalósításának következménye, de természetesen a jelentős és eredeti tudományos eredmények elérése is szükséges feltétel volt.

A nyolcvanas években a természettudományos kutatások értékelésénél, továbbá személyek és kutatócsoportok tevékenységének megítélésénél széles körben használták a publikációkra vonatkozó adatokat. Ez jelentősen hozzájárult az eredményesség objektív megítéléséhez. A módszer propagálói mindig hangsúlyozták, hogy a publikációs adat csupán egyike az értékelésnél használandó mutatóknak, mely számos egyéb adattal, információval kiegészítve adhat alapot a döntésekhez. Ennek ellenére túlzások, egyoldalú lekicsinylések, vagy ellenkezőleg, a publikációs adatoknak vélt kizárólagossága egyaránt előfordult.

Az utóbbi években világszerte felerősödött az igény a szakértői bírálat (peer review) metodikájának, legyen az egyéni vagy csoportos jellegű, jobb megismerésére és szakavatottabb használatára az elbírálások és minősítések során. Kétségkívül

vannak fontos tényezők, amelyre a tudománymetria alig tud választ adni, a szakértői bírálat viszont képes lehet a válaszadásra. Ilyen pl. egy pályázatnál az eredetiség megítélése.

Szinte minden olyan jelentésben, amelyet amerikai vagy nyugat-európai tudósok írtak a magyar tudomány jelenlegi helyzetéről, megtalálható az az ajánlás, hogy fordítsunk nagyobb figyelmet a jövőben a peer review módszer szakszerű alkalmazására és ahol lehet és indokolt, ott a tudománymetriai módszerekkel összekapcsolva hajtsuk végre az értékelési és elbírálási munkákat.

Ezt az igényt kívánja részben kielégíteni a jelen kiadvány, amelyben a peer review módszer alkalmazásáról, sajátos problémáiról találhatunk eredeti tanulmányokat. Angol nyelven adjuk közre ezeket a cikkeket. A magyar nyelvre való lefordítás egyrészt jelentős többletkiadást igényelt volna, de ettől függetlenül is úgy érezzük, hogy a magyar tudományos kutatók döntő többsége ma már jól olvas angolul és megérti a kutatások értékelésével foglalkozó módszertani cikkeket.

Kívánom, hogy hasznosítsák mindazokat a gondolatokat, amelyek ezekben a cikkekben találhatók, és amelyek valóban újak és tényleg hasznosíthatók számunkra. Nem a nulláról indulunk a szakértői bírálati módszer alkalmazásánál, de van még mit tanulnunk és elsajátítanunk olyanoktól, akik valóban magas szintű módszertani vizsgálatokat végeztek.

EUGENE GARFIELD:

Refereeing and Peer Review. Part 1. Opinion and Conjecture on the Effectiveness of Refereeing

Current Contents, August 4, 1986

Peer review is so much a part of the fabric of scholarly inquiry that it is often taken for granted. I have written many essays over the years that are directly or indirectly related to peer review. These include several on authorship[1-3] and editing,[4] faculty evaluation,[5] identifying Nobel-class science through citation analysis[6-9]—and even a few on various aspects of refereeing itself.[10-12] But I have never before discussed the intricacies of the system in detail. Since the subject is central to scholarly life, we have decided to devote a three-part essay in *Current Contents*® to it.

The first two parts will cover refereeing for publication. Part 1 examines how the refereeing system works and lists some of the common opinions about its advantages and disadvantages. Part 2 will cover scientific studies of refereeing and some proposed alternatives to the present system. Part 3 will follow later and will focus on the peer review of grant proposals. Note that I distinguish between a *referee* (one who evaluates an article before it is published) and a *reviewer* (one who evaluates already published material or, in the case of grant reviews, research-grant proposals). I generally use the term *referee* to mean one who advises editors on the publishability of a scholarly manuscript. The process by which this advice is solicited I usually call *refereeing*, but occasionally *reviewing* or *peer review* seems appropriate. The term *peer review* is also used to denote the evaluation of research proposals; more generally, it can refer to the professional review of patient records by special committees of physicians that many hospitals use to maintain high-quality patient care.

Refereeing: How It Came About and How It Works

Refereeing is meant to ensure that articles submitted for publication meet the accepted standards of their fields. Like editing, refereeing is a complex intellectual, political, and social process; it often involves a spectrum of activities that blend into one another in complex ways, in a fashion similar to the range of practices related to ghostwriting.[3] Among many who have expressed the idea, Peter Amiry, former editor, *Journal of the Operational Research Society*, wrote in an editorial that referees are an editor's insurance policy, providing a reservoir of knowledge that few editors could hope to match.[13]

The practice of refereeing manuscripts prior to publication is now well established, but it was not always so, state sociologists Harriet Zuckerman and Robert K. Merton, Columbia University, New York, in their classic 1971 study of patterns of evaluation in sci-

ence.[14] It evolved in response to the development of scholarly societies and the scientific journal. I summarized this and other work in an earlier essay on the changes in scientific communication over the past 300 years.[15]

According to David A. Kronick, professor of medical bibliography, University of Texas Health Science Center at San Antonio, "science in the seventeenth and eighteenth centuries...differed in many ways socially, intellectually, and economically from the science of the twentieth century."[16] Although associations and societies promoting scholarly activities had existed for hundreds of years,[17] (p. 46), the social role of "scientist," as well as conventions for doing research, had yet to emerge.[16] In fact, Kronick notes, "individuals did not begin to regard themselves as scientists rather than philosophers until the seventeenth century."[17] (p. 34)

The learned journal as we know it today also traces its origins to the seventeenth century, with the founding of the *Philosophical Transactions of the Royal Society of London* and the *Journal des Sçavans*, associated with the *Académie des Sciences* in Paris.[14] By the early eighteenth century, Kronick says, members of these and other scholarly societies sponsoring official or semiofficial publications began to realize that if scholars were to have confidence in the content of these journals, then material submitted for publication had to be critically evaluated before it was published.[16]

Societies thus began to take measures to preserve their credibility. Some adopted strict regulations governing publication that members had to comply with to retain their membership. And by the mid-eighteenth century, according to Kronick, some—such as the Royal Society of Medicine of Edinburgh, Scotland—had developed techniques of evaluating and approving manuscripts before publication that are almost indistinguishable from today's system of refereeing.[16] Kronick, incidentally, is the author of a recent book on the literature of the life sciences that includes a short section on the refereeing and the publication process in that branch of science.[18]

The procedures involved in refereeing a manuscript vary from journal to journal and from field to field, but there are certain general steps that virtually every paper has to go through before it is published. Among the first steps an editor takes, whether or not the journal is refereed, is to evaluate a submission's compatibility with the scope and style of the journal, according to Robert A. Day, consultant, ISI Press®, and former managing editor, American Society for Microbiology (ASM) journals.[19] Once this is done, an editor must then choose appropriate referees for a given manuscript.

Donald Christiansen, editor, *IEEE Spectrum*, conducted a survey of referee selection practices among 26 of the *IEEE Transactions* editors. Common sources from which referees are recruited include widely recognized experts, members of a journal's editorial board, professional acquaintances, previous referees, and scientists cited in the author's references.[20] Sometimes authors are asked to supply a list of suggested referees. A few journals are using manual and computer-assisted bibliographic retrieval methods to select referees. For example, Stevan Harnad, editor, *Behavioral and Brain Sciences* (*BBS*), reports that *BBS* staffers search a microcomputer file of the journal's referees that has been coded by areas of expertise. They also search the current biobehavioral literature through the *Science Citation Index*® and the *Social Sciences Citation Index*® for additional referee candidates.[21,22]

Usually two referees are chosen, according to Claude T. Bishop, director, Division of Biological Sciences, National

Research Council of Canada (NRCC), and editor-in-chief, NRCC Research Journals. "The merits of this system," he writes, "are that it usually provides at least one solid [report], that the two [referees] can be checked against each other, and that one referee may cover points that the other missed."[23] But Harnad notes that, for many journals, the "number of referees [selected for a manuscript] is an empirical matter requiring research."[21] *BBS* uses five to eight referees per paper. In Harnad's experience, such a sample is more likely to produce a balanced review.[24]

Along with the manuscript, referees generally receive a list of instructions and a form for comments and recommendations. Routinely, referees respond within a few weeks, recommending either publication or rejection or requesting modifications; they often include specific comments for both the author and the editor.

A paper is most likely to be accepted, according to Michael Gordon, research associate, Primary Communications Research Centre, University of Leicester, UK, when the referees agree that it meets three criteria.[25] (p. 6-8) First, it should be *sound*. The author(s) should have employed reliable research techniques, drawn valid conclusions, and committed no flaws of logic. It should also be *original*, in the sense that its findings have never before been published. Finally, it should be *significant*, meaning that it should contain some new perspective or observation of potential importance.[25] (p. 6-8) Of course, published articles meet these criteria in varying degrees.

Referees do not always agree with one another, and some authors take this as evidence that the system is unreliable or capricious. But disagreement is at the heart of scientific inquiry. Harnad says that "the current and vital ongoing aspect of science consists of an active and often heated interaction of data, ideas, and minds, in a process one might call 'creative disagreement.' "[26] Moreover, reviewer disagreements are not simply shrugged off; editors generally resolve each dispute on an individual basis. Gordon described some of the options open to editors for dealing with these conflicts.[25] (p. 20-5) When reviewer disagreements are mild, for example, editors may rely on their own judgment to resolve them—with, perhaps, some communication with the author.[25] (p. 21) When differences are profound, editors may reject the paper without further reviewing or they may send the manuscript out for review once again, together with the comments of the disputing referees. Editors may also ask the author to respond to the referees' observations. After the "arbitrating" referee(s) and the author have reported, editors should be in a better position to make a final judgment. When authors take exception to referees' comments and provide editors with a point-by-point refutation, editors often follow a procedure similar to the one just outlined for adjudicating disputes between referees.[25] (p. 22-5)

Research, Pseudo-Research, or Non-Research?

The results of our literature search for this essay support the view that refereeing is an issue clouded with subjectivity and emotionalism—at least for a vocal minority. The dominant vehicle of discussion in the debate about the effectiveness of refereeing has been editorials and correspondence. Some contain incisive discussions, but with little or no empirical evidence to support what amounts to a litany of opinion and anecdote. Indeed, in an endeavor such as science, which depends on dispassionate logic and systematic evidence for much of its credibility, the dearth of rigorous thinking and hard data in the correspondence of many who are critical of refereeing is remarkable. Of the relative-

ly few controlled studies that have been done, many suffer from such severe methodological shortcomings that their conclusions are questionable. More will be said about research on refereeing in Part 2.

Refereeing and other forms of peer review have been discussed at length, especially in the four decades since World War II, but discussion alone does not constitute science or scholarship. Since we are all affected by peer review, it is not surprising that so many of us have opinions on the subject. Yet the literature representing controlled studies of peer review is either pitifully small or disgracefully absent, while the body of anecdote and opinion is quite large. We carefully distinguish here between studies, experiments, experience, and opinions.

In researching this essay, we also found that most published opinion on refereeing is negative. But we suspect that this is due, ironically, to the widespread acceptance of and satisfaction with the current system of peer review: most scientists simply do not feel that refereeing needs defending, so positive opinions are relatively scarce. It should also be kept in mind that these opinions on refereeing are themselves unrefereed. Furthermore, the existence and ranking of hundreds of refereed journals is concrete evidence that they are the preferred medium of publication.

Flaws in the System?

In a note published in the *New England Journal of Medicine* (*NEJM*), John C. Bailar III and Kay Patterson, Harvard School of Public Health, Boston, Massachusetts, speculate that current opinion on refereeing seems divided among one or more of four paradigms.[27] Based on their own informal observations, the authors assert that many scientists seem to perceive the process as a sieve, sifting the wheat from the chaff. Many also liken the process to a smithy, in which "papers are pounded into new and better shapes between the hammer of peer review and the anvil of editorial standards." Some seem to view it as a switch, reflecting the widespread belief that a persistent author can eventually publish a manuscript somewhere (although refereeing may determine exactly where). Finally, some scholars seem to consider refereeing a capricious and essentially unpredictable process—a "shot in the dark."[27]

Stephen Lock, editor, *British Medical Journal*, feels that refereeing "favours unadventurous nibblings at the margin of truth rather than quantum leaps."[28] An example supporting his opinion is the reception given the early demonstration, via radioimmunoassay, of insulin-binding antibody by the late Solomon A. Berson and Rosalyn S. Yalow, Veterans Administration, New York. This work was fundamental to the development of the radioimmunoassay into a "powerful tool for determination of virtually any substance of biologic interest," according to Yalow.[29] Although Yalow would share the 1977 Nobel Prize with Roger Guillemin, Salk Institute, San Diego, and Andrew Schally, Veterans Administration Hospital, New Orleans, the initial research concerning radioiodine-labeled insulin was rejected both by *Science* and, at first, by the *Journal of Clinical Investigation* (*JCI*) as erroneous.[29]

Nevertheless, when the paper was revised to meet the objections of reviewers, it was published in the *JCI*.[30] A comparatively recent poll of the authors of manuscripts rejected by the *JCI*, conducted by editor Jean D. Wilson, Department of Internal Medicine, University of Texas Health Science Center at Dallas, found that 85 percent of the rejected papers were subsequently published elsewhere. And Wilson also re-

ported that "most of the authors of the [other] 15 percent...were convinced by the review process that [their papers] were either unoriginal or wrong."[31]

Delays in Publication

In addition to charges that referees make too many serious mistakes, complaints also focus on the delays in publication that many attribute to the refereeing process. While conceding the value of thorough, constructive reports by referees, Richard Shea, editor, *Transactions on Nuclear and Plasma Sciences*, is nevertheless concerned about the time lost during the refereeing process; he is quoted by Christiansen as saying that "the ultimate referee is the reader."[20] And as noted by Kronick, the historical significance of papers ultimately depends on this reader evaluation and readers' willingness to cite what impresses them.[32] But one of the reasons for the existence of the refereeing system is that readers of scientific articles have varying interests and backgrounds; they *must* be able to rely on a high degree of validity in what they read, especially if it is somewhat outside their field.

Real or perceived, delays in publication resulting from refereeing may be the most prevalent concern among scientists, who may have job security, promotions, or the need to establish priority for a discovery hanging in the balance. In a note in *NEJM*, Thomas P. Stossel, Massachusetts General Hospital, Boston, voices his concern that the commercial potential of many new discoveries, especially in biotechnology, is giving rise to new and particularly taxing demands for rapid publication.[33]

In an editorial, Lawrence D. Grouse offers several explanations, based on his experience as senior editor of *JAMA*, for the lag time between submission and publication: "Excellent manuscripts are often criticized by reviewers with vested interests or contrary views. Overcritical reviewers flay manuscripts for minor or supposed deficiencies.... Reviewers may also cynically delay the appearance of research competing with their own."[34] And in a 1979 editorial in the *Journal of Clinical Psychiatry*, associate editor Marc H. Hollender asked "why it takes three months or longer to review an article that takes three minutes to read and perhaps took less than three months to write.... Does it take the referee that long to come to a conclusion and to dictate comments? It is more likely that the article gathers dust among other low-priority items."[35] In short, if I may use an old, informal phrase, referees should either fish or cut bait.

Bias and Unethical Behavior

Of all the complaints about refereeing, however, some of the most bitter—though not the most prevalent—concern the issue of referee bias (although little *uncontested* empirical evidence exists to indicate that authors' affiliations and the reputations of their institutions affect a referee's evaluation). Assuming that some bias exists, however, historian of science Donald deB. Beaver, Williams College, Williamstown, Massachusetts, suggests that a preconceived suspicion of scientific "have-nots" may be explained in terms of the second part of the "Matthew effect."[36] This concept, introduced by Merton in 1968,[37] draws an analogy between the misallocation of scientific credit and a passage from the gospel of St. Matthew: "Unto every one that hath shall be given, and he shall have abundance: *but from him that hath not shall be taken away even that which he hath*" (emphasis ours). Presumably, contributions from unknown scholars

from unrecognized or little-known institutions are less likely to be accepted for publication than occasionally comparable contributions by scholars of great repute.

Some cases of questionable referee ethics have been documented. Perhaps the most publicized example, according to a 1984 article by free-lance medical writer Barbara Fox in *Medical Communications*, the journal of the American Medical Writers Association,[38] was one reported on by former *Science* staff writer William J. Broad.[39] It involved a paper submitted by Helena Wachslicht-Rodbard, NIH, Bethesda, Maryland, to *NEJM*. The paper was assigned to two referees, one of whom recommended acceptance, while the other—Vijay Soman of Yale University, who had similar research in progress—recommended rejection. Arnold Relman, editor, *NEJM*, informed Wachslicht-Rodbard that her paper had "engendered considerable differences of opinion among our referees"[39] and told her the manuscript was unacceptable unless revised.

But the matter was far from over. Soman had photocopied Wachslicht-Rodbard's study and, without informing his coauthor, Philip Felig, vice chairman of the Department of Medicine at Yale, of what he had done, sent their article incorporating the plagiarized data to the *American Journal of Medicine*, of which Felig was an associate editor. By coincidence, the journal sent the article out for review to Wachslicht-Rodbard's superior, who showed it to her. It contained more than a dozen passages, verbatim, from her own manuscript; she wrote to Relman accusing Felig and Soman of plagiarism and conflict of interest in the refereeing of her paper. Relman agreed that it had been highly improper for Soman to agree to even read the paper, which was later published in the *NEJM* under Wachslicht-Rodbard's name.[40]

The abuse of anonymity is a long-standing matter of concern. In an article appearing in *New Scientist*, biochemist Robert Jones, Royal College of Surgeons, London, asserted that "the act of submission of a paper can place the author at the mercy of the malignant jealousy of an anonymous rival."[41] The belief seems to be that, from behind the walls of their fortress of anonymity, referees are free to hurl at authors volleys of invective that cannot be effectively countered. "Anonymity tends to bring out the worst in people," according to Heinz Fraenkel-Conrat, Department of Molecular Biology and Virus Laboratory, University of California, Berkeley, in a letter to the editors of *Nature*.[42] "I was recently asked to review, and advocated rejection of, a paper for a virological journal on the basis of factual comments which I would have been quite willing to sign. The editor sent me, out of courtesy, copies of his rejection letter together with the other referee's sarcastic poison-pen comments, also rejecting the paper. There was no justification for one civilised person insulting another in such a manner.... That outburst was solely the joy of releasing adrenalin with anonymous impunity."[42] While Fraenkel-Conrat's analysis may be correct in this situation, there is little evidence, other than anecdotal, that this is a widespread phenomenon. But it suggests fertile ground for study: do *ad hominem* comments—those leveled at authors, as distinct from strong opinions about the authors' text—occur more frequently in signed or in unsigned reviews?

In a "Guest Comment" published in *Physics Today*, F. Curtis Michel, professor of space physics and astronomy, Rice University, Houston, calls for referees to back up their comments. "Accountability is now all directed back at the author," he writes.[43] "If there is any dispute, it is entirely the authors' fault because they have 'failed to convince their peers.' Here, the word 'peer' has a

nice ring of fairness to it.... However,... when a group of colleagues is permitted to have [their] comments taken as some kind of gospel, [they] are no longer peers but quite definitely superiors insofar as power and influence go."[43] It is in answer to just this kind of criticism, Harnad reports, that *BBS* is conducting an internal, statistical study of, among other things, the relationships among anonymity, referees' ratings of manuscripts, and authors' ratings of the usefulness of referee reports.[24]

Another criticism of the system is of the "Newcomb variety." I have often referred to the career of Simon Newcomb, who proved conclusively—just months before the Wright Brothers took off from the sands of Kitty Hawk—that a flying machine was impossible.[44,45] Sometimes this type of rejection is the result of referees who are hostile to innovative ideas or to those that clash with their own.[41] We don't know how often thoughtful, conscientious scientists—in good faith and in keeping with currently accepted theory—rendered an opinion concerning the implausibility of a given idea or theory, only to see that theory become the basis of a dramatic paradigm shift. Still, referees and journal editors should not consider such rejection experience as sufficient reason for extending some kind of "publication *carte blanche*" to would-be authors who want to prove, for example, that perpetual-motion machines are possible. I continue to be in favor of refereeing that prevents the publication of intellectual atrocities, including papers with inadequate documentation. For those articles straddling the border between science and speculation, there exist publications such as *Speculations in Science and Technology*, which was started specifically as a forum for the publication of ideas lacking support "in established theoretical and experimental work," according to an article by founder William M. Honig, senior lecturer in the physical sciences and engineering, Western Australian Institute of Technology, Perth, in the *Sciences*.[46]

Refereeing and Garfield's Uncertainty Principle

It is easy to "prove" on the basis of anecdotal evidence that the refereeing system doesn't work. From the hundreds of published *Citation Classics*® commentaries—such as those written by Oscar Buneman, Stanford University, California,[47] and Hans Lineweaver, US Department of Agriculture, Washington, DC[48]—or in correspondence with their authors, we know that dozens of significant papers have been rejected by some journals for various reasons. Some of these reasons might be described as "N-I-H," that is, "not invented here." Nevertheless, much scientific quackery is exposed by careful, insightful, constructive refereeing, and this far outweighs the ideas that have allegedly been suppressed because of referees who would not give them a chance to see the light of day.

A scientist's appreciation of the collaborative, communal goal of refereeing—protecting science and the public from errors and inferior work—varies according to a host of factors, including the scientist's age, status, and temperament. Famous, tenured, or established researchers may be better able to weather the occasional rejection notice than scientists just starting their careers and trying to make their mark. No other activity is as fundamental to democratic scholarship as refereeing. From all this, I concluded that there is an Uncertainty Principle of Refereeing: The more we have of it, the less we like it—but the less we have of it, the more we miss it.

We sometimes trivialize what we take for granted. Refereeing has been around for so long that it's easy to forget that it wasn't always there. The present stage of

its evolution will be affected by social and technological factors such as funding and electronic publishing. But the public discourse of scholarship, both formal and informal, is essential to the very existence of science. In the modern era of big science—and by that I mean both large-scale projects and large numbers of projects, whether small or large—we must find ways to inculcate new research practitioners with the precepts and ideals that "naturally" were taught in the era of little science. We cannot allow squabbling over limited research funds to cloud the fundamental need to preserve the scientific *process* implied by refereeing. But we must recognize that the very size of the scientific enterprise may make it necessary to modify rigid application of the Ingelfinger rule[49] [promulgated by the late Franz J. Ingelfinger, former editor, *NEJM*, which states that papers submitted to *NEJM* must "have been neither published nor submitted elsewhere (including news media and controlled-circulation publications)"] or other precepts that may have been reasonable before the electronic revolution.

Indeed, the community of science may become even more relevant in the new communications age, and so we have to examine more carefully the consequences for intellectual property rights and methods of adjudicating disputes concerning priority of discovery. If much of this sounds Mertonian in tone it is no accident, since Robert K. Merton is one of the few scholars who has devoted great effort to the definition of the problems involved in research on refereeing. In fact, the work of Zuckerman and Merton will form a significant part of the discussion in Part 2 of this essay.

* * * * *

My thanks to Stephen A. Bonaduce and Terri Freedman for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** From citation amnesia to bibliographic plagiarism. *Essays of an information scientist.* Philadelphia: ISI Press, 1981. Vol. 4. p. 503-7.
2. ----------. More on the ethics of scientific publication: abuses of authorship attribution and citation amnesia undermine the reward system of science. *Ibid.*, 1983. Vol. 5. p. 621-6.
3. ----------. Ghostwriting—the spectrum from ghostwriter to reviewer to editor to coauthor. *Current Contents* (48):3-11, 2 December 1985. (Reprinted in: *Essays of an information scientist: ghostwriting and other essays.* Philadelphia: ISI Press, 1986. Vol. 8. p. 460-8.)
4. ----------. Alternative forms of scientific publishing: keeping up with the evolving system of scientific communication. *Op. cit.*, 1981. Vol. 4. p. 264-8.
5. ----------. How to use citation analysis for faculty evaluations, and when is it relevant? Parts 1 & 2. *Ibid.*, 1984. Vol. 6. p. 354-72.
6. ----------. The 1984 Nobel Prize in medicine is awarded to Niels K. Jerne, César Milstein, and Georges J.F. Köhler for their contributions to immunology. *Current Contents* (45):3-18, 11 November 1985. (Reprinted in: *Essays of an information scientist: ghostwriting and other essays.* Philadelphia: ISI Press, 1986. Vol. 8. p. 416-31.)
7. ----------. The 1984 Nobel Prize in physics goes to Carlo Rubbia and Simon van der Meer; R. Bruce Merrifield is awarded the chemistry prize. *Current Contents* (46):3-14, 18 November 1985. (Reprinted in: *Essays of an information scientist: ghostwriting and other essays.* Philadelphia: ISI Press, 1986. Vol. 8. p. 432-43.)
8. ----------. The 1984 Nobel Prizes in economics and literature are awarded to Sir Richard Stone for pioneering systems of national accounting and to Jaroslav Seifert, the national poet of Czechoslovakia. *Current Contents* (49):3-13, 9 December 1985. (Reprinted in: *Essays of an information scientist: ghostwriting and other essays.* Philadelphia: ISI Press, 1986. Vol. 8. p. 469-79.)
9. ----------. Do Nobel Prize winners write Citation Classics? *Current Contents* (23):3-8, 9 June 1986.
10. ----------. Publishing referees' names and comments could make a thankless and belated task a timely and rewarding activity. *Op. cit.*, 1977. Vol. 1. p. 435-7.
11. ----------. Anonymity in refereeing? Maybe—but anonymity in authorship? No! *Ibid.*, 1977. Vol. 2. p. 438-40.

12. ----------. Reducing the noise level in scientific communication: how services from ISI aid journal editors and publishers. *Ibid.*, 1980. Vol. 3. p. 187-8.
13. **Amiry P.** Refereeing for *JORS. J. Oper. Res. Soc.* 34:1025-6, 1983.
14. **Zuckerman H & Merton R K.** Patterns of evaluation in science: institutionalisation, structure and functions of the referee system. *Minerva* 9:66-100, 1971. [Reprinted as: Institutionalized patterns of evaluation in science. (Merton R K.) *The sociology of science.* Chicago, IL: University of Chicago Press, 1973. p. 460-96.]
15. **Garfield E.** Has scientific communication changed in 300 years? *Op. cit.*, 1981. Vol. 4. p. 394-400.
16. **Kronick D A.** Authorship and authority in the scientific periodicals of the seventeenth and eighteenth centuries. *Libr. Quart.* 48:255-75, 1978.
17. ----------. *A history of scientific & technical periodicals.* Metuchen, NJ: Scarecrow Press, 1976. 336 p.
18. ----------. *Literature of the life sciences: reading, writing, research.* Philadelphia: ISI Press, 1985. 219 p.
19. **Day R A.** *How to write and publish a scientific paper.* Philadelphia: ISI Press, 1983. p. 82.
20. **Christiansen D.** Peer review reviewed. *IEEE Spectrum* 18:21, 1981.
21. **Harnad S.** Personal communication. 25 June 1986.
22. ----------. Commentary on "Computer-assisted referee selection as a means of reducing potential editorial bias" by H. Russell Bernard. *Behav. Brain Sci.* 5:202, 1982.
23. **Bishop C T.** *How to edit a scientific journal.* Philadelphia: ISI Press, 1984. p. 53.
24. **Harnad S.** Commentary on "Peer review and the *Current Anthropology* experience" by C. Belshaw. *Behav. Brain Sci.* 5:201, 1982.
25. **Gordon M.** *Running a refereeing system.* Leicester, UK: Primary Communications Research Centre, University of Leicester, 1983. 56 p.
26. **Harnad S.** Creative disagreement. *Sciences* 19(7):18-20, 1979.
27. **Ballar J C & Patterson K.** Journal peer review: the need for a research agenda. *N. Engl. J. Med.* 312:654-7, 1985.
28. **Lock S.** Letter to P.B.S. Fowler. 4 December 1984. *Brit. Med. J.* 290:1560, 1985.
29. **Yalow R S.** Radioimmunoassay: a probe for the fine structure of biologic systems. *Science* 200:1236-45, 1978.
30. **Berson S A, Yalow R S, Bauman A, Rothschild M A & Newerly K.** Insulin-I^{131} metabolism in human subjects: demonstration of insulin binding globulin in the circulation of insulin treated subjects. *J. Clin. Invest.* 35:170-90, 1956.
31. **Wilson J D.** Peer review and publication. *J. Clin Invest.* 61:1697-701, 1978.
32. **Kronick D A.** Personal communication. 20 June 1986.
33. **Stossel T P.** Speed: an essay on biomedical communication. *N. Engl. J. Med.* 313:123-6, 1985.
34. **Grouse L D.** The Ingelfinger rule. *JAMA—J. Am. Med. Assn.* 245:375-6, 1981.
35. **Hollender M H.** Authors, editors and referees. *J. Clin. Psychiat.* 40:331, 1979.
36. **Beaver D D.** On the failure to detect previously published research. *Behav. Brain Sci.* 5:199-200, 1982.
37. **Merton R K.** The Matthew effect in science. *Science* 159:56-63, 1968. (Reprinted in: *The sociology of science.* Chicago, IL: University of Chicago Press, 1973. p. 439-59.)
38. **Fox B.** Peer review and the public's right to know: a look at the Ingelfinger Rule. *Med. Commun.* 12:33-7, 1984.
39. **Broad W J.** Imbroglio at Yale (I): emergence of a fraud. *Science* 210:38-41, 1980.
40. **Wachslicht-Rodbard H, Gross H A, Rodbard D, Ebert M H & Roth J.** Increased insulin binding to erythrocytes in anorexia nervosa. *N. Engl. J. Med.* 300:882-7, 1979.
41. **Jones R.** Rights, wrongs and referees. *New Sci.* 61:758-9, 1974.
42. **Fraenkel-Conrat H.** Letter to editor. (Is anonymity necessary?) *Nature* 248:8, 1974.
43. **Michel F C.** Solving the problem of refereeing. *Phys. Today* 35(12):9; 82, 1982.
44. **Garfield E.** Negative science and "The outlook for the flying machine." *Opt. cit.*, 1980. Vol. 3. p. 155-72.
45. **Newcomb S.** The outlook for the flying machine. *Independent* 55:2508-12, 1903.
46. **Honig W M.** Science's Miss Lonelyhearts. *Sciences* 24(3):24-7, 1984.
47. **Buneman O.** Citation Classic. Commentary on *Phys. Rev.* 115:503-17, 1959. *Current Contents/Engineering, Technology & Applied Sciences* 15(37):16, 10 September 1984.
48. **Lineweaver H.** Citation Classic. Commentary on *J. Amer. Chem. Soc.* 56:658-66, 1934. *Current Contents/Life Sciences* 28(11):19, 18 March 1985.
49. Definition of "sole contribution." *N. Engl. J. Med.* 281:676-7, 1969.

EUGENE GARFIELD:

Refereeing and Peer Review. Part 2.

The Research on Refereeing and Alternatives in the Present System

Current Contents, August 11, 1986

Continuing our discussion of refereeing, which focused on complaints about the system in Part 1,[1] we now examine the empirical research on the subject, the anecdotal literature supporting the current system, and some of the suggestions for improving it. Part 3 will appear at a later date and will discuss the peer review of grant proposals. Again we will review the considerable literature of opinion and conjecture, but we will give special attention to the large-scale study by sociologists Stephen Cole, State University of New York (SUNY), Stony Brook, and Jonathan R. Cole, Columbia University, New York,[2,3] as well as other papers[4] and special reports.[5]

Editors: The Author's Guardians

Each anecdote purporting to reveal some fault in the present system of refereeing seems to find a ready counterpart in the opinion of a supporter. For instance, many critics claim that some referees do not review manuscripts dispassionately. But editors say that they usually take great pains to ensure that referees are fair. In *Running a Refereeing System*, Michael Gordon, research associate, Primary Communications Research Centre, University of Leicester, UK, recommends the use of two or more referees to reduce the risk of an offhand, frivolous, or biased treatment of a manuscript.[6] (p.13-5) When referees *do* cause excessive delays, return unsupported or capricious reports, or otherwise display "questionable ethics," they tend to be retired from the system, according to Patricia Dehmer, Argonne National Laboratory, Illinois, and member, Publications Committee, American Physical Society (APS) in a "Guest Comment" in *Physics Today*.[7] Whether this is the case in other disciplines is not known.

Critics also suggest that referees sometimes take advantage of the privileged information they are privy to in the manuscripts they review. But Dehmer asserts that many APS editors try to ensure that referees are not working along lines precisely like those of the papers sent to them, to reduce the possibility of conflicts of interest. But this is contrary to the practice in biomedicine and elsewhere. Most editors try to match submissions with reviewers as closely as possible, in an attempt to have the manuscript reviewed by those presumed to be most qualified to judge it. In either case, according to Claude T. Bishop, director, Division of Biological Sciences, National Research Council of Canada (NRCC), and editor-in-chief, NRCC Research Journals, referees ought to disqualify themselves when there is the possibility of a conflict of interest, or when they feel they cannot be objective about the paper or its author. In some instances, however, they might propose simultaneous publication of their own paper and the review paper, or even approach the authors of the review paper and propose a collaboration.[8] (p. 50, 82) As a parallel approach, many editors honor author requests that a paper not be sent to a given referee.[7]

Authors Often Lack Knowledge of Publishing

Editors also point out that authors frequently do not understand the publication process. For instance, many authors charge that referees make up a closed, "elite" group. Yet the number of active referees for a journal can far exceed the number of active contributors.[9] According to *JAMA* editor George D. Lundberg, that journal's list of active referees contains over 3,000 names.[10] The *Journal of the Operational Research Society*, a relatively small journal, used 285 referees in 1982 alone.[11] And a careful study of nine years of materials from the archives of *Physical Review* and *Physical Review Letters* by sociologists Harriet Zuckerman and Robert K. Merton, Columbia University,[12] showed that authors of every rank participated in the refereeing process. Their main finding, which is based on referee reports for both published and rejected manuscripts and which refutes another widely held belief, is that there is no consistent relationship between referee acceptance or rejection of manuscripts and the relative standing of authors and referees.[12] In addition, informed authors know that it is not referees, but editors, who are ultimately responsible for rejecting a manuscript.

Bishop says that authors also show a lack of understanding when they point to differences of opinion among referees as evidence that the system is capricious and unreliable.[8] (p. 43-9) At the root of some of these reviewer disagreements, in Bishop's view, are differences in the algorithms and paradigms fundamental to every branch of science. For instance, referees less often disagree substantially in well-established fields. But in fields pressing at the frontiers of knowledge, significant differences of opinion among referees are bound to be more common. When editors are confronted with a decision between two equally plausible referee interpretations of a given manuscript, they often employ one of several options that range from publishing the paper without comment to publication of the controversial paper along with comments by referees, invited critics, and rebuttals by the authors.[8] (p. 43-9)

Authors also seem to assume that their submissions are, in general, carefully written and based on substantial amounts of work. "Not so," asserts J.W. Cornforth, Milstead Laboratory of Chemical Enzymology, Sittingbourne Research Centre, Kent, UK, who served as a referee for a dozen journals over a 30-year period.[13] "In my experience," Cornforth continues in his letter to the editors of *New Scientist*, "a regrettably high proportion [of manuscripts] show careless or misleading presentation and meager experimental work, and the majority need some modification. Referees—and, of course, editors—almost invariably improve a paper that passes through their hands; often, they are doing what the authors ought to have done."[13]

The Many Faces of Rejection

Authors should also be aware that the scientific value of a paper is not necessarily the only factor that enters into editors' decisions to publish or not; many manuscripts never make it past the screening process that eliminates papers that are incompatible with a journal's readership or have not been submitted in the required format.[14] Or a journal may reject a manuscript simply because it has recently published another, similar paper, or has one currently under consideration.[10] Rejection rates are also significantly affected by the existence of page charges, which support publication and thus allow for much lower rejection rates. This practice is widespread in physics and chemistry but not unknown even in psychology.

It is also important to realize that rejection rates vary. In their study of patterns of evaluation in science, Zuckerman and Merton compiled a table of the rejection rates for a sample of 83 journals in the sciences, the social sciences, and the humanities.[12] Linguistics, geology, and physics journals had the lowest rate of rejection, turning down only 20

to 25 percent of the papers submitted to them. Biology journals rejected about 30 percent of the papers they received. Journals in experimental and physiological psychology had a rejection rate of over 50 percent, while sociology journals were over 80 percent and history journals hovered at 90 percent. Stephen Lock, editor, *British Medical Journal* (*BMJ*), made an observation that has also been noted by others who have read the study. He wrote that "the more humanistically oriented the journal, the higher the rate of [rejection]; the more experimentally and observationally oriented, with an emphasis on rigour of observation and analysis, the lower the rate of rejection."[15] (p. 17)

Zuckerman and Merton also reported that the editorial staff's attitude concerning its own errors in judgment constitutes an often-overlooked factor influencing acceptance rates.[12] Although editors and referees want to avoid errors in judgment altogether, they recognize that they cannot be infallible; thus, since they must make mistakes, they tend to have preferences for the *kind* of mistakes they are willing to risk. The staffs of some journals—notably those prestigious journals with high rejection rates—seem more willing to reject "unorthodox" manuscripts that the wider community of scholars might eventually consider important, rather than to run the risk of publishing a substandard work. The staffs of low-rejection journals, on the other hand, apparently prefer to publish the occasional work that doesn't measure up, rather than reject a paper that later turns out to be significant.[12]

The Research

A research front consists of a group of current papers that, together, cite one or more of a cluster of older, core publications. Since I referred earlier[1] to the paucity of empirical research on refereeing and peer review and the abundance of anecdote and opinion on the subject, one may wonder how a research front of any size might be generated on this subject. But even a large anecdotal literature, through repeated citations of previous anecdotal literature, as well as reputable studies, can form a pseudo-research front. Only a careful analysis of the core and citing literature can determine the nature and extent of the research front—even when very useful core review papers can be found. Since the literature on peer review and refereeing is vast, at the end of Part 2 of this essay I have added a selected bibliography of publications not mentioned in the text.

The 1983 ISI® research front entitled "Objectivity of reviewers in peer review" (#83-8291) consists of but 2 core papers and 12 citing papers. One core paper is the highly controversial 1982 study by Douglas P. Peters, University of North Dakota, Grand Forks, and Stephen J. Ceci, Cornell University, Ithaca, New York.[16] The other core paper is a 1982 editorial by Lock, entitled "Peer review weighed in the balance."[17] In it Lock discusses the conclusions drawn by Peters and Ceci and details some of the flaws in their study. In spite of these problems, however, Lock believes that Peters and Ceci have underscored some shortcomings within the system. Most of the recommendations Lock makes for improving refereeing—particularly double-blind review—are discussed in detail below.

Peters and Ceci

This controversial study involved the resubmission of 12 psychology articles—published by authors from prestigious and highly productive departments—to the journals that originally published them.[18] Peters and Ceci became interested in doing the study after reading about an informal experiment conducted by Los Angeles free-lance writer Chuck Ross.[19] He reports having submitted the untitled, untyped manuscript of Polish-born US literary author Jerzy Kosinski's novel *Steps*[20] under a pseudonym to publishers and literary agents to see if "unknown" authors re-

ceive fair consideration. Although the book had won the 1969 US National Book Award, Ross claimed that 14 publishers—including the book's original publisher—and 13 agents rejected it.[19]

In the Peters and Ceci study, the presentation of the data in the original papers was slightly altered. Fictitious names and institutions were substituted for the real ones, but the content of the articles was unchanged. Three of the resubmissions were detected as such; of the other nine, eight were rejected. The authors concluded that the rejections resulted from a systematic bias against unknown authors and institutions. In the commentary section published along with Peters and Ceci's article, however, many commentators pointed out a number of flaws in the study. For instance, according to anthropologist Sol Tax, University of Chicago, Illinois, and Robert A. Rubinstein, School of Public Health, University of Illinois Medical Center, Chicago, the names Peters and Ceci chose for their bogus institutions were far removed from the mainstream of psychology institutions. Thus, what the investigators really demonstrated, say Tax and Rubinstein, is a bias against materials originating outside *appropriate* institutions.[21] Nobel laureate Rosalyn S. Yalow, Veterans Administration, New York, commented, "How does one know that the data are not fabricated?... Those of us who publish establish some kind of a track record. If our papers stand the test of time, it can be expected that we have acquired expertise in scientific methodology.... The work of established investigators in good institutions is more likely to have had prior review from competent peers and associates even before reaching the journal."[22]

Garth J. Thomas, Center for Brain Research, University of Rochester, New York, suggests that referees and editors may have recognized the resubmitted articles as very like something they had seen before, but rather than raise the specter of plagiarism, they fell back on statistical criticisms to justify their negative comments.[23] Janice M. Beyer, School of Management, SUNY, Buffalo, writes that the most likely fate of any submitted article is to be unanimously rejected, as 80 to 90 percent are in the social sciences.[24]

In addition, psychologist Grover J. Whitehurst, SUNY, Stony Brook, notes that Peters and Ceci had no control group.[25] Richard M. Perloff, Department of Communication, Cleveland State University, Ohio, and Robert Perloff, Graduate School of Business, University of Pittsburgh, suggest that, among other controls, Peters and Ceci's study should have included resubmitting articles by authors from low-status institutions under by-lines with equally low-status affiliations, as well as resubmitting articles by high-status authors under equally high-status by-lines.[26] "Without such controls it is impossible to argue that the findings reflect the status bias [that Peters and Ceci] suggest," the Perloffs write.[26]

But *Is* There Bias?

Still, Tax and Rubinstein feel that a bias preventing competent work from being published is much more damaging than one that lets mediocre work slip through.[21] And anecdotal evidence of bias is so widespread that the possibility should not be dismissed by researchers. For instance, in another commentary on the Peters and Ceci article, Robert Rosenthal, Department of Psychology, Harvard, said that as a young member of the psychology faculty at the University of North Dakota, he was unable to publish 15 to 20 articles in mainstream journals in the 1960s. Within a few years of his move to Harvard, however, he says that most of these articles were published in the same journals that had previously rejected them.[27] He does not say, however, whether these were the identical articles, or if they had been substantially revised to meet the objections of reviewers or changed in any other way.

In a 1970 investigation of how attitudes might influence referee judgment, Leonard D. Goodstein and Karen

Lee Brazis, University of Cincinnati, Ohio, mailed abstracts of an empirical study on astrology to 282 members of the American Psychological Association.[28] They were asked to rate the design of the paper. Half were sent an abstract that reflected a conclusion confirming commonly held scientific attitudes toward astrology; the other half received an identical abstract, except that it included a conclusion that ran counter to scientific beliefs. The former was rated by most referees as better designed and having more significance for future research. The latter, which contradicted common wisdom, was rated as flawed.

When Zuckerman and Merton examined the selection of articles for the *Physical Review*, they found that papers by physicists of great repute affiliated with prestigious institutions were more likely to be exempted entirely from the refereeing process. Their papers were accepted and published more quickly than papers by lesser known physicists.[12] And in a large-scale study of papers submitted to physics journals, Gordon reported a strong bias in referees from major universities toward papers by authors who were also from large, well-known universities.[29]

Lock, however, found no evidence of referee bias in a study of 1,558 manuscripts submitted to *BMJ* between January and August 1979. The study was published in his book *A Difficult Balance: Editorial Peer Review in Medicine*.[15] Of the 246 external referees who were sent manuscripts by *BMJ*, 143 held academic positions, while the rest had nonacademic affiliations; yet the proportion of papers recommended for acceptance did not differ from one group to the other.[15] (p. 56-71) Moreover, regardless of the affiliations of both referee and author, Lock said that referees judged manuscripts "to an equal standard."[15] (p. 61)

Suggestions for Improvement

A few years ago, Norton D. Zinder, Rockefeller University, New York City, sent me the text of a talk he gave to the Society of Editors in 1969, when he was an associate editor of *Virology*.[30] Tongue partially in cheek, Zinder asked, "What would be so terrible if there were no refereeing of scientific papers?... As we now operate, with [the] restriction of publication by reviewing, the *number* of publications becomes a thing in itself.... If we were to cease refereeing papers,...there'd be little bar [to publication, and] quality might reassert its role, since there'd be less pressure to have long lists of publications."[30] The Perloffs write that the "caveat emptor approach [of having no refereeing system at all] might be viewed as a nod to the free market of ideas. Let millions of flowers bloom."[26] Some may feel that the continued growth of the literature may lend support to these views. However, others, including myself, believe that a few non-refereed publications can exist only because the refereed journals set the standards for all the others.

I believe that most scientists would agree that if something is indeed shown to be wrong with refereeing, an attempt should be made to repair the system, rather than to abandon it. Unfortunately, with little or no solid, systematic evidence of refereeing's deficiencies, most suggestions for improvement are as conjectural as the ills they are meant to cure. Among the most discussed options—one that is already prevalent among sociology journals—is that of double-blind refereeing, also called reciprocal anonymity, in which neither the authors nor referees are aware of the others' identities. There is precedent for author anonymity: David A. Kronick, professor, medical bibliography, University of Texas Health Science Center at San Antonio, notes that "maintaining the anonymity of the author was a standard practice in the prize essay competitions (a sort of early form of sponsored research) of eighteenth-century scientific societies, which had elaborate devices to maintain the anonymity of contributors."[31]

The rationale behind double-blind refereeing, as was pointed out in an ap-

propriately anonymous editorial in *Nature*, is that referees could still be frank about a manuscript's shortcomings without fear of ruining working relationships or being subjected to the anger of rejected authors.[32] Such a system would also, in the opinion of J. Scott Armstrong, Wharton School, University of Pennsylvania, Philadelphia, "reduce the prejudice against unknown authors from low-status institutions."[33]

Many justify the present system by citing what Marcel C. La Follette, editor, *Science, Technology, & Human Values*, calls the "crackpot avoidance" theory.[34] According to this idea, an author's record of achievement and the stamp of legitimacy provided by the author's institutional affiliation help referees evaluate manuscripts because they constitute presumptive "proof" that the research described was really done. La Follette says that accepting manuscripts without regard for the potential of misrepresentation or error is unwise, but she points out that a prestigious affiliation is no guarantee against fraud—in fact, it may even help the perpetrator evade detection.

According to John Moossy, editor-in-chief, and Yvonne R. Moossy, managing editor, *Journal of Neuropathology & Experimental Neurology*, a common objection to double-blind refereeing is a widespread conviction that experienced referees can identify authors despite the removal of the authors' names from their manuscripts.[35] In a study conducted to test this contention, they removed the names of authors and their departmental and institutional affiliations from 33 papers sent out for refereeing from May 1983 through April 1984. Each of the 67 referees, who filed a total of 85 reports, was asked to identify the authors and their departments or disciplines; 34 percent were able to make correct identifications. Eleven percent made incorrect identifications, and 55 percent would not even hazard a guess. Interestingly, only 9 referees objected to the double-blind procedure; a surprising number—24—had "no opinion," while 33 favored it, citing such reasons as greater objectivity and less risk of being swayed, either for good or ill, by the author's reputation.[35]

Another frequently proposed reform is "open refereeing." It is the exact opposite of double-blind refereeing: the referee's name is revealed to the author, who in turn is made known to the referee. Proponents argue that open refereeing might reduce the number of careless and superficial reports, on the presumption that referees will take more care with their reports if they have to sign their names to them. And in fact, I noted long ago that the time of the more qualified referees is of proportionately greater value; thus, they may sometimes be less than enthusiastic over the prospect of a manuscript to evaluate.[36] Anonymity is a dull spur to effort; "Aren't we all more likely to do something properly if our name is attached to it?" asks Ronald Mirman, Department of Physics, Long Island University, Brooklyn, New York, in a letter to the editor of the *American Journal of Physics*.[37]

Armstrong proposes that referees might designate a portion of their report to be signed and published along with the manuscript. He believes this would provide useful information to scientists because few readers can devote the kind of attention to a paper that a referee gives to it.[33] However, a number of problems might be encountered were referee anonymity abolished. For instance, the late Franz J. Ingelfinger, former editor, the *New England Journal of Medicine*, believed that "the referee who is several steps below the author on the status ladder" might be put in an uncomfortably vulnerable position and might even be unwilling to criticize candidly the manuscript in question.[38] Some reviewers might soften their objections to manuscripts, rather than jeopardize working relationships with the authors.[6 (p. 16)] Identifying referees would also enable authors to get in touch with them. This might foster a communication process that excludes the editor, or even exposes referees to verbal attacks.[31]

The Perloffs have another suggestion for promoting a greater sense of responsibility among referees. They argue that

paying referees would encourage them to perform their task more thoroughly and impartially.[26] Although they do not say how much referees should receive, they suggest that such fees could come from "authors' institutions, their research funding, or their personal resources."[26] They present no empirical evidence supporting their argument, but the notion of paying reviewers, like other ideas reported in this essay, could form the basis of an interesting study. In this case, the questions might be, "Do paid referees perform better than unpaid ones?" and "How much money does it take before a significant effect is noticed?"

Conclusion

It is difficult to draw substantive conclusions about how well the refereeing process functions. But Lock makes an interesting observation: the validating of experimental results and theoretical conclusions is ultimately not through the refereeing process but through the broader evaluation that articles receive over time at the hands of a larger, informed scientific community.[15] (p. 128) Of course, refereeing does not always detect fraud, plagiarism, errors, and muddy thinking. Still, it is probably impossible for most journals to switch to a system of in-house evaluation: despite its faults, real or imagined, refereeing is probably the most efficient and effective method for distinguishing the promising from the meretricious—at least, until it is *proven* otherwise.

In assessing refereeing's supposed flaws, one of the key issues seems to be delays in publication. Much of the accumulated anxiety about refereeing in many fields seems traceable to the tedious process that is often made out of what should be a straightforward decision. At the heart of many delays are referees who allow manuscripts to gather dust on their desks without informing editors that they cannot complete a review in a timely fashion.

As I see it, at the root of many of the alleged deficiencies of peer review are the attitudes of the scientific community itself. Were quality valued over quantity, and spurious "productivity" deplored instead of rewarded with tenure and promotions or research grants, then the incentive to publish shoddy or half-finished research would diminish. This might reduce the burden placed upon editors and reviewers because of the publish-or-perish syndrome. Unfortunately, we have not yet emerged from the stage of regarding the sheer number of publications as significant,[39] but there is a growing tendency to limit the number of papers to be listed on nominations for awards, grants, and so on.[40] And in fact, one of the often-stated goals of citation analysis is to encourage quality, high-impact work, rather than publication for the sake of pure output.

Of the myriad comments about refereeing, it is difficult to find one brief, all-encompassing statement that says it all. But John Ziman, Imperial College of Science and Technology, London, UK, and editor, *Science Progress*, has come close. In a commentary on Peters and Ceci, he wrote:

> Informed discourse on the primary communication system of science takes for granted the basic utility and reliability of the peer-review process, at least up to some modest practical level of human competence. The height of this level should not be exaggerated: It is not an indicator of permanent scientific worth. Acceptance for publication by a reputable journal implies no more than that the work is superficially sound, mildly interesting, and moderately original. The opinion that it should at least be taken into consideration by other scientists is only a preliminary assessment, likely to be contradicted and entirely superseded in the light of further study. Nevertheless, this weak and uneven standard of quality appears real enough to the authors, editors, and reviewers who tussle endlessly to establish and maintain it. Specific accusations of prejudice, inquiries concerning systematic bias, and demands for institutional reform have all been addressed to imperfection of performance around and about this hypothetical benchmark.[41]

The question of refereeing must be discussed in the larger context of peer review for funding research. In the next part of this essay, I hope to review the anecdotal as well as systematic information available. But refereeing and peer review are ethical and sociopolitical issues scientists must review periodically. Democratic institutions are dynamic. We want to retain the best of what we have had, but we must be willing to change that which no longer satisfies the needs of a changing world.

Postscript

Since it is a primary mission of ISI Press® to publish books on the process of scientific communication, it has published several such works mentioned in this essay. Several more, including Lock's *A Difficult Balance: Editorial Peer Review in Medicine*,[15] will be printed or reprinted by ISI Press in the fall. They are: *Medical Style and Format: an International Manual for Authors, Editors, and Publishers*[42] and *How to Write and Publish Papers in the Medical Sciences*,[43] by Edward J. Huth, editor, *Annals of Internal Medicine*; *How to Copyedit Scientific Books and Journals*,[44] by Maeve O'Connor, CIBA Foundation, London, UK; and *An Insider's Guide for Medical Authors and Editors*,[45] by Peter Morgan, scientific editor, *Canadian Medical Association Journal*. Incidentally, Lock's book contains a bibliography of over 200 references—some of which appear following the references in this essay in the selected bibliography. In a review[46] of Lock's book, Alfred Yankauer, editor, *American Journal of Public Health*, says it is "an invaluable reference for all those interested in the editorial process." In his review, he quotes a passage from Alexander Pope[47] that he feels "captured the essence" of Lock's views on refereeing and the editor's role. Yankauer suggests that for the word "critic," the reader should substitute "editor" or "referee/reviewer."[46]

> But you who seek to give and merit fame,
> And justly bear a Critic's noble name,
> Be sure yourself and your own reach to know,
> How far your genius, taste and learning go;
> Launch not beyond your depth, but be discreet,
> And mark that point where sense and dullness meet....
>
> Careless of censure, nor too fond of fame;
> Still pleas'd to praise, yet not afraid of blame;
> Averse alike to flatter or offend;
> Not free from faults, nor yet too vain to mend.
>
> Alexander Pope
> *An Essay on Criticism*

* * * * *

My thanks to Stephen A. Bonaduce and Terri Freedman for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** Refereeing and peer review. Part 1. *Current Contents* (31):3-11, 4 August 1986.
2. **Cole S, Rubin L & Cole J R.** *Peer review in the National Science Foundation: phase one of a study.* Washington, DC: National Academy of Sciences, 1978. 193 p.
3. **Cole J R & Cole S.** *Peer review in the National Science Foundation: phase two of a study.* Washington, DC: National Academy Press, 1981. 106 p.
4. **Russell A S, Thorn B D & Grace M.** Peer review: a simplified approach. *J. Rheumatol.* 10:479-81, 1983.
5. **Sanders H J.** Peer review. How well is it working? *Chem. Eng. News* 60(11):32-43, 1982.
6. **Gordon M.** *Running a refereeing system.* Leicester, UK: Primary Communications Research Centre, University of Leicester, 1983. 56 p.

7. **Dehmer P.** APS reviews refereeing procedures. *Phys. Today* 35(2):9; 95-7, 1982.
8. **Bishop C T.** *How to edit a scientific journal.* Philadelphia: ISI Press, 1984. 138 p.
9. **McCaffery M.** Peer review—or sneer review? *Can. Fam. Physician* 29:857, 1983.
10. **Landberg G D.** Appreciation to our peer reviewers. *JAMA—J. Am. Med. Assn.* 251:758; 817-23, 1984.
11. **Amiry P.** Refereeing for *JORS. J. Oper. Res. Soc.* 34:1025-6, 1983.
12. **Zuckerman H & Merton R K.** Patterns of evaluation in science: institutionalisation, structure and functions of the referee system. *Minerva* 9:66-100, 1971. [Reprinted as: Institutionalized patterns of evaluation in science. (Merton R K.) *The sociology of science.* Chicago, IL: University of Chicago Press, 1973. p. 460-96.]
13. **Cornforth J W.** Letter to editor. (Referees.) *New Sci.* 62:39, 1974.
14. **Day R A.** *How to write and publish a scientific paper.* Philadelphia: ISI Press, 1983. p. 82.
15. **Lock S.** *A difficult balance: editorial peer review in medicine.* London: Nuffield Provincial Hosptials Trust, 1985. 172 p.
16. **Peters D P & Ceci S J.** Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain Sci.* 5:187-95, 1982.
17. **Lock S.** Peer review weighed in the balance. *Brit. Med. J.* 285:1224-6, 1982.
18. **Peters D P & Ceci S J.** A manuscript masquerade. *Sciences* 20 (7):16-9; 35, 1980.
19. **Ross C.** Rejected. *New West* 4(4):39-43, 1979.
20. **Kosinski J.** *Steps.* New York: Random House, 1968. 147 p.
21. **Tax S & Rubinstein R A.** Responsibility in reviewing and research. *Behav. Brain Sci.* 5:238-40, 1982.
22. **Yalow R S.** Competency testing for reviewers and editors. *Behav. Brain Sci.* 5:244-5, 1982.
23. **Thomas G J.** Perhaps it was right to reject the resubmitted manuscripts. *Behav. Brain Sci.* 5:240, 1982.
24. **Beyer J M.** Explaining an unsurprising demonstration: high rejection rates and scarcity of space. *Behav. Brain Sci.* 5:202-3, 1982.
25. **Whitehurst G J.** The quandary of manuscript reviewing. *Behav. Brain Sci.* 5:241-2, 1982.
26. **Perloff R M & Perloff R.** Improving research on and policies for peer-review practices. *Behav. Brain Sci.* 5:232-3, 1982.
27. **Rosenthal R.** Reliability and bias in peer-review practices. *Behav. Brain Sci.* 5:235-6, 1982.
28. **Goodstein L D & Brazis K L.** Psychology of scientist: XXX. Credibility of psychologists: empirical study. *Psychol. Rep.* 27:835-8, 1970.
29. **Gordon M D.** The role of referees in scientific communication. (Hartley J, ed.) *The psychology of written communication.* New York: Nichols, 1980. p. 263-75.
30. **Zinder N D.** *Editing without reviewers; or the review process—a protection from what?* Unpublished speech presented to the Society of Editors, 19 May 1969. Cambridge, MA. 6 p.
31. **Kronick D A.** Personal communication. 19 June 1986.
32. In defence of the anonymous referee. *Nature* 249:601, 1974.
33. **Armstrong J S.** The ombudsman: is review by peers as fair as it appears? *Interfaces* 12(5):62-74, 1982.
34. **La Follette M C.** On fairness and peer review. *Sci. Technol. Hum. Val.* 8(4):3-5, 1983.
35. **Moosy I & Moosy Y R.** Anonymous authors, anonymous referees: an editorial exploration. *J. Neuropathol. Exp. Neurol.* 44:225-8, 1985.
36. **Garfield E.** Publishing referees' names and comments could make a thankless and belated task a timely and rewarding activity. *Essays of an information scientist.* Philadelphia: ISI Press, 1977. Vol. 1. p. 435-7.
37. **Mirman R.** Letter to editor. (For open refereeing.) *Amer. J. Phys.* 43:837, 1975.
38. **Ingelfinger F J.** Peer review in biomedical publication. *Amer. J. Med.* 56:686-92, 1974.
39. **Garfield E.** How to use citation analysis for faculty evaluations, and when is it relevant? Parts 1&2. *Op. cit.*, 1984. Vol. 6. p. 354-72.
40. **Angell M.** Publish or perish: a proposal. *Ann. Intern. Med.* 104(2):261-2, 1986.
41. **Zinman J.** Bias, incompetence, or bad management? *Behav. Brain Sci.* 5:245-6, 1982.
42. **Huth E J.** *Medical style and format: an international manual for authors, editors, and publishers.* Philadelphia: ISI Press. (In press.)
43. ----------. *How to write and publish papers in the medical sciences.* Philadelphia: ISI Press. (In press.)
44. **O'Connor M.** *How to copyedit scientific books and journals.* Philadelphia: ISI Press. (In press.)
45. **Morgan P.** *An insider's guide for medical authors and editors.* Philadelphia: ISI Press. (In press.)
46. **Yankauer A.** Review of "A difficult balance: editorial peer review in medicine" by S. Lock. *CBE Views* 9(2):51-2, 1986.
47. **Pope A.** *Pastoral poetry and an essay on criticism.* (Audra E & Williams A, eds.) London: Methuen, 1961. p. 244; 326.

Armstrong J S. Peer review of scientific papers. *J. Biol. Resp. Modif.* 3:10-4, 1984.
Beck C W. Trouble in the hedgerows. *J. Archaeol. Sci.* 12:405-9, 1985.
Crane D. The gatekeepers of science: some factors affecting the selection of articles for scientific journals. *Amer. Sociol.* 2:195-201, 1967.
Dixon G F, Schonfeld S A, Altman M & Whitcomb M E. The peer review and editorial process: a limited evaluation. *Amer. J. Med.* 74:494-5, 1983.
Fox T. *Crisis in communication.* London, UK: Athlone Press, 1965. 59 p.
Gardner M J, Altman D G, Jones D R & Machin D. Is the statistical assessment of papers submitted to the "British Medical Journal" effective? *Brit. Med. J.* 286:1485-8, 1983.
Harnad S. Rational disagreement in peer review. *Sci. Technol. Hum. Val.* 10(3):55-62, 1985.
----------. Review of "A difficult balance" by S. Lock. *Nature* (In press.)
----------, ed. *Peer commentary on peer review: a case study in scientific quality control.* Cambridge, UK: Cambridge University Press, 1982. 71 p. (Reprinted from: *Behav. Brain Sci.* 5:185-255, 1982.)
Juhasz S, Calvert E, Jackson T, Kronick D A & Shipman J. Acceptance and rejection of manuscripts. *IEEE Trans. Prof. Comm.* PC18:177-85, 1975.
Koshland D E. Memorandum to Universal Science Foundation. *Science* 229:921, 1985.
Light R J & Pillemer D B. *Summing up. The science of reviewing research.* Cambridge, MA: Harvard University Press, 1984. 191 p.
Lloyd J E. On watersheds and peers, publication, pimps and panache. (An editorial abstract.) *Fla. Entomol.* 68:134-9, 1985.
Maddox J. Privacy and the peer-review system. *Nature* 312:497, 1984.
Mahoney M J. Open exchange and epistemic progress. *Amer. Psychol.* 40:29-39, 1985.
Meadows A J. The problem of refereeing. *Scientia* 112:787-94, 1977.
Miller A C & Serzan S L. Criteria for identifying a refereed journal. *J. Higher Educ.* 55:673-99, 1984.
Morgan P P. When reviewers disagree. *Can. Med. Assn. J.* 129:1172-3, 1983.
----------. Anonymity in medical journals. *Can. Med. Assn. J.* 131:1007-8, 1984.
----------. Author, editor and reviewer: how manuscripts become journal articles. *Can. Med. Assn. J.* 124:664-6, 1981.
Patterson K & Ballar J C. A review of journal peer review. (Warren K S, ed.) *Selectivity in information systems: survival of the fittest.* New York: Praeger, 1985. p. 64-82.
Shils E. The confidentiality and anonymity of assessment. *Minerva* 13:135-51, 1975.
Silver S. Ethical questions in the peer review system. *ASM News.* 46:302-6, 1980.
Smith B M & Gough P B. Editors speak out on refereeing. *Phi Delta Kappan* 65:637-9, 1984.
Stossel T P. Reviewer status and review quality: experience of the *Journal of Clinical Investigation. N. Engl. J. Med.* 312:658-9, 1985.
Strasburger V C. Righting medical writing. *JAMA—J. Am. Med. Assn.* 254:1789-90, 1985.
Suppe R J & Zirkel P A. The importance of refereed publications: a national survey. *Phi Delta Kappan* 64:739-40, 1983.
Whitehurst G J. Interrater agreement for journal manuscript reviews. *Amer. Psychol.* 39:22-8, 1984.
----------. Interrater agreement for reviews for *Developmental Review. Develop. Rev.* 3:73-8, 1983.
----------. On lies, damned lies, and statistics: measuring interrater agreement. *Amer. Psychol.* 40:568-9, 1985.

EUGENE GARFIELD:

Refereeing and Peer Review. Part 3. How the Peer Review of Research-Grant Proposals Works and What Scientists Say About It

Current Contents, January 26, 1987

This essay is the third in a series on refereeing and peer review in science. The first part examined the anecdotal evidence and other literature and opinions about refereeing, the evaluation of scholarly articles before publication.[1] The second discussed research on refereeing and suggestions for improving the system.[2] This part focuses on the workings of the peer-review system of evaluating research-grant proposals, as employed by major US federal funding agencies such as the National Science Foundation (NSF) and the National Institutes of Health (NIH); the fourth section will cover the research on the grant-review system and proposed alternatives to it.

The emphasis in Parts 3 and 4 is on peer review in the physical, chemical, and biological sciences, since those are the fields examined by the major studies sponsored by the NSF and the NIH. However, it should be noted that the social sciences and the arts and humanities also have funding mechanisms that incorporate peer review and that funding for science, the arts, and the humanities also comes from numerous private sources that have their own methods of determining the level of support they wish to provide.

The Science-Government Connection

The principle, if not the full-fledged system, of peer review developed along with the scholarly societies and learned journals that were founded in the seventeenth and eighteenth centuries.[3,4] But until this century, it remained a matter of interest and concern only within the scientific community. In the US, however, during the 1940s, science and government began to establish a close working relationship that went beyond the advisory role scientists had previously played in affairs of state. For instance, according to Jay A. Levy,[5] University of California School of Medicine, San Francisco, with the passage of the Public Health Service Act in 1944,[6] the US Surgeon General was authorized to "make grants-in-aid to universities, hospitals, laboratories, and other public or private institutions and to individuals for...research."[5,7] And in the late 1940s, according to Rustum Roy, director, Science, Technology and Society Program, Pennsylvania State University, University Park, the Office of Naval Research (ONR) was "the first systematically organized government source of research funds for universities."[8] At that time, "peer review began as an informal 'seeking of a second opinion' by the grants manager, who mailed a copy of a proposal on the periphery of his competence to a colleague and followed up with a phone call."[8]

The close ties that evolved during World War II between the government and the scientific community were formalized in 1950 by the creation of the NSF.[9] According to former NSF director, psychologist Richard C. Atkinson, chancellor, University of California, San Diego, and physicist William A. Blanpied, currently international studies specialist at the NSF, the "science-government contract [was an attempt] to bring science into the political system while at the same time preserving its autonomy."[9] But Roy claims that in the process, each agen-

cy using some form of informal peer review "enshrined" its version, "without any thought, examination, or analysis,...[as] 'THE peer review' system."[8]

As Atkinson and Blanpied note, however, it was primarily through peer review that scientists convinced the government that the public interest would best be served "if scientists...retained decisive influence over how public funds were spent to support scientific activities."[9] The assumption was—and is—that since few in public office have the expertise to determine, from a technical standpoint, which projects are most deserving, the task of evaluating research proposals should fall to scientists. The Public Health Service Act allowed for the provision of funds only to projects approved by the National Advisory Health Council or the National Advisory Cancer Council—the precursors of the current NIH system of National Advisory Committees.[10] This method, with peer review as its cornerstone, served as a model for other government agencies, such as the NSF.

Atkinson and Blanpied point out, as have many others, that in recent years, despite increases in levels of NSF funding, the total funding of science, in terms of real dollars, has declined.[9] There has not been a similar decline in funds or in buying power at the NIH, according to William F. Raub,[10] deputy director, but Raub does not dispute Levy's observation that 95 percent of competing applications recommended for funding received support in the mid-1960s, while only 30 to 40 percent receive funds today.[5] "Funding has gone up, but the number of those asking for funds has gone up even faster," Raub explains.[10] And of those whose applications are okayed, most receive support at levels reduced from the amounts originally recommended.[5] Incidentally, an interview[11] with Raub on the subject of misconduct in science was recently published in *The Scientist*.[12,13]

Still, until the early 1980s, scholars and the institutions with which they were affiliated abided by the consensus that peer review was the best method to ensure the fair distribution of the ever-smaller federal pie.[9] But in 1983 and 1984, 15 universities bypassed the peer-review system and obtained more than $100 million in special authorizations and appropriations for new facilities directly from the US Congress; some even hired professional lobbyists to assist them.[9,14]

Peer Review: Love It or Leave It?

John Silber, president, Boston University, in a comment reported in *Science News*, justified his institution's abandonment of accepted channels by charging that the peer-review system is an old-boy network that preferentially funds some 20 institutions.[14] In calling for reforms to the peer-review system, Robert L. Sinsheimer, chancellor, University of California, Santa Cruz, said that peer review perpetuates the status quo.[15] And according to a recent report in *Chemical & Engineering News*, Senator Dennis DeConcini from Arizona claimed that "50 percent of all federal R&D [research and development] funding was put into the hands of 16 eastern and West Coast universities" in the 1984 fiscal year.[16] (DeConcini represents an area that includes the University of Arizona, which secured $25 million from the Senate Appropriations Committee.) Whether or not there is research to support the claims of Silber, Sinsheimer, and others is a question that will be discussed in Part 4, but it is interesting to note that Columbia University—which can by no means be labeled a "have-not" institution—was among those that took shortcuts with the system. Atkinson and Blanpied note that Columbia officials secured $8 million in US Department of Energy (DOE) funds for the construction of a chemical research laboratory.[9]

The tactics of those who have bypassed the peer-review system have predictably elicited a strong, negative response from those who have remained within the system. Roland W. Schmitt, chairman, National Science Board (the policy-making arm of the NSF), claims that, without peer review, US science is on "the fast track to mediocrity."[14] Senator Jeff Bingaman of New Mexico worries that bypassing peer review may weaken the morale of scientists who have worked hard to develop meritorious proposals, only to find themselves politically out-

maneuvered.[16] The practice may also divert scarce resources from research projects that the scientific community considers to be of higher priority. As Atkinson and Blanpied write, "At issue is not whether meritorious research will be carried out in facilities obtained through pork-barrel tactics. Rather, [such] tactics violate the understanding that available resources are to be allocated in the best overall interests of science—and the public—rather than in the interests of individual claimants."[9] Roy, however, pointedly wonders who will define what the "best overall interests of science" are.[17]

How Peer Review Works

The principal agencies that support basic scientific research in the US are the NSF, the NIH, the Veterans Administration (VA), the Department of Defense, the DOE, the Department of Agriculture, and the National Aeronautics and Space Administration (NASA),[9] although funds are also provided through such organizations as the American Chemical Society, which administers the Petroleum Research Fund.[18] Various congressional committees and the US Office of Management and Budget determine the amount of money each government agency has available to disburse.[9] In general, how that money is spent depends largely on peer review: area experts judge proposals on their scientific and technical merit and make recommendations accordingly. But each agency or organization charged with dispensing funds for scholarship operates with a somewhat different set of procedures.

The NIH, which accounts for most US basic research-grant funding in terms of total dollars per year, makes use of a two-tiered system called "dual review."[19] (p. 41) At the first level, a panel of experts in a given field, called the Initial Review Group (IRG), evaluates a research application for its scientific merit. The IRG also comments on the applicant's performance on any previous grants and recommends a funding priority for the application, as well as the amount and duration of the grant.[18] At the second level, the respective National Advisory Council/Board of each bureau, institute, or division of the NIH reviews the recommendations made by the IRG and makes its own, final judgment concerning the application's relevance to the NIH's various programs and priorities.[19] (p. 42-3)

At the NSF, applications in chemistry, physics, and mathematics are mailed to 3 to 10 independent experts selected by a program officer.[18,20] (p. 7) These experts, referred to as mail reviewers, individually evaluate applications for their scientific quality through written comments and boxes checked off on a multiple-choice form.[20] (p. 7) Applications in the earth, biological, and social and behavioral sciences are usually reviewed by a combination of mail reviewers and panel reviewers.[18,20] (p. 7) Panels consist of scientists selected by program directors; the size of panels varies from section to section, but each meets in Washington, DC, three times a year to evaluate proposals.[20] (p. 8) Like the NIH's IRG, NSF reviewers report on the applicant's track record, as well as on the relevance of the work and on the capability of the applicant's institution to provide technical support. Based on these comments, the program officer makes a recommendation to higher-ranking officials, who in turn make the final decision.[18,19] (p. 22-4) [20] (p. 3-11) One significant difference between the NIH and the NSF procedures is that NSF officers have considerable discretion to modify or even disregard peer-review recommendations, whereas at the NIH, all recommendations by the IRGs are followed very closely.[10]

The procedures of other agencies and organizations are, for the most part, variations on either the NSF or the NIH models. For instance, the VA, like the NIH, conducts an initial review of a grant application through a discipline-based review board that makes a recommendation to VA administrators; they constitute a second level of review.[19] (p. 25-6) Unlike the NSF or the NIH, however, the VA attempts to provide some funding for all approved proposals. When NASA receives unsolicited research proposals, it operates in much the same fashion as the NSF, with *ad hoc* reviewers who make recommendations based on scientific merit. Review procedures at the ONR are also sim-

ilar to the NSF's, in that Navy scientists may evaluate proposals themselves or have them reviewed through the mail or by a panel of experts convened for the purpose.[18] Naval officers have much more to say in the decision-making process, however, than do their counterparts at the NSF or the NIH.

Peer Review Outside the US

The peer-review systems of the UK, France, and the Federal Republic of Germany (FRG) provide a perspective on complaints about the US system and point the way toward possible improvements. In the UK, according to a two-volume compendium of source materials researched and published by the NSF, general support for all university programs and operations is provided through a dual system.[21] The first element, the University Grants Committee (UGC), provides general support for all functions of British universities in the form of annual block grants; about a quarter of this money goes to the direct support of research. Although there is general agreement between the universities and the UGC on how this money should be spent, the universities have wide latitude in the use of these grants. On rare occasions, however, the UGC makes a grant for a specific purpose and suggests the most effective use for the funds.

The second element of the UK system is provided in the form of research grants for specific university activities. These monies are administered by the five publicly funded research councils—the Science and Engineering Research Council, the Medical Research Council, the Natural Environment Research Council, the Agricultural Research Council, and the Economic and Social Sciences Research Council.[21] The role played by the UK's scientific community in the disbursement of funds from the UGC and the research councils is similar to the peer-review process in the US, but much of UK scientists' advice is provided through informal channels.

The US and the UK carry out more than half of their key basic research in universities, whereas national laboratories and independent institutions produce most of the work in other countries. In France, for instance, the single most important funding agency is the National Center for Scientific Research (CNRS) in Paris, which in 1979 directly employed about 8,500 scientists and 14,000 support personnel.[21] R&D priorities are developed at the level of national policy by the Secretary of State for Research, and the size of each CNRS laboratory, its budget, and the number of new positions in the system are all determined by the government. All university faculty are civil servants, paid directly by the Ministry of the Universities. The most prominent scientific advisory group is the Advisory Committee for Research in Science and Technology (CCRST), also known as the Committee of Sages. Made up of 16 members, the CCRST advises the Secretary of State for Research on a wide range of scientific issues.

Several large French government R&D agencies, however, pursue courses that are essentially independent of CNRS.[21] The Ministry of Defense, for example, which accounts for one-third of all government financing of research, relies heavily on its own facilities and establishes its own priorities, although it does have extensive contact with industry and academia. Other largely independent agencies include the National Institute for Health and Medical Research, the National Institute for Agricultural Research, the National Center for Space Studies, the National Center for Telecommunications Studies, and the National Center for Exploitation of the Oceans. The CCRST has no influence with the technical ministries, which have their own advisory groups of scientists.

Research funds in the FRG are provided by state and federal governments, private foundations, and industry.[21] For basic research, most of the funding is supplied by the federal government's Ministry of Research and Technology (BMFT) and the Ministry for Education and Science. However, state governments also contribute significantly—especially to the privately operated Max Planck Institutes, a system of research institutions set up outside the university system to support outstanding scientists in key fields. The BMFT also provides the principal support for applied re-

search. The money from these agencies is funneled into grants by the German Research Society (DFG), a scholarly society that operates beyond the boundaries of formal government. For scientists affiliated with German universities, which are state-owned, DFG grants supplement a certain minimum level of funding. DFG support goes out mainly in the form of small, individual project grants that run from one to three years. Grant applications are evaluated by peers who are elected to their positions by the entire scientific community.[21]

In summary, the three countries briefly discussed here, as well as others, provide a relatively stable level of operating support to their universities, as well as to a parallel basic-research system separate from the university system.[21] And although many governments provide some funds on a competitive, peer-reviewed basis to scientists working both within and outside of the university systems, such support is relatively small compared with the baseline funding. Since research support in other countries is not limited to individual projects for short periods of time, foreign scientists, unlike their US counterparts, do not have to cope with the distractions of securing grant money and the disruptions suffered when grants are reduced or discontinued.[21] Atkinson and Blanpied claim that systems outside the US are less effective in encouraging competition among the most innovative ideas, and that other nations' faculty members, who are virtually or even literally employees of their respective governments, "cannot claim the same degree of autonomy they can in the United States."[9] Roy says that "not one study has ever been supported to test this" claim[17] or to compare the review systems of various government agencies with other methods of allocating funds.[22]

Criticisms of Peer Review

With research projects, jobs, and even careers at stake each time the review process renders a verdict, it is not surprising that the effectiveness and fairness of the system are matters of great concern—especially since the mid-1970s, when government support began declining.[18] Making matters worse, the NIH and the NSF, the major grant-giving agencies, do not have the money to fund every application they approve. In a remark reported in *Science*, NIH director James B. Wyngaarden said that the issue of distinguishing between "shades of excellence" was among those that most concerned scientists. The distinctions between one excellent proposal and another are often so fine that judgments concerning relative quality cannot be rendered on an objective basis, leaving those whose top-rated proposals are rejected "angry and frustrated."[23] Many scientists also feel it is inappropriate to rank disparate proposals that have little in common with one another.[5] And Roy goes even further, charging that "there is no theoretical or empirical justification to support the contention that 'good' research can be predicted on the basis of the 'evaluation of' proposed ideas contained in an essay."[8]

Another, almost universal, concern about peer review—found even among reviewers and agency administrators—is the time, effort, and money it takes to complete the paperwork involved in applying for and evaluating proposals. To ensure approval of a grant application, physiologist Daniel H. Osmond, University of Toronto, Canada, writes, "many have sacrificed 1-3 months of productive research.... The entire year is dominated by thoughts of preparation, and of the tragic consequences of refusal.... We must do quick experiments, write them up fast, and publish, publish, publish.... Innovative, time-consuming work must be done on the side with unbudgeted dollars to diminish the risk of rejection by an overcautious grants committee when the work eventually surfaces."[24]

Rosalyn S. Yalow, VA, New York, and the 1977 Nobel laureate in physiology or medicine, adds that grant proposals are "inherently dishonest," since "few established investigators whose contributions are highly original and imaginative can spell out... detailed plans for a three- or a five-year period."[25] If the investigator can do so, she continues, then "he does not expect to make a discovery; in fact, that mind-set can keep him from recognizing a discovery."[26] And Osmond adds that the constant pressure of applying for grant renewals can cause scien-

tists, both consciously and unconsciously, to groom research results to fit the expectations of the funding agency, rather than allowing the work its own head.[24]

Other complaints about peer review closely resemble those made concerning the refereeing of manuscripts prior to publication, which were discussed in detail in Part 2 of this essay.[2] Just as authors complain of referee bias and old-boy networks that conspire to keep new, challenging ideas out of print, so too do applicants for research grants charge that young or new scientists with little or no track record don't get a fair shake in competition with older, more established scientists and that grant-review committees are hesitant to risk funds on innovative or speculative proposals.[18,27] And, like authors, grant applicants also fear that those who review their work may end up stealing from it as well.[24]

Summarizing the workings of the complex peer-review systems in the US and some parts of Europe is not a simple task. Equally difficult is the job of condensing the dissatisfactions with peer review, which are mainly reflected in anecdotal complaints about the current US system. The final part of this series will focus on research findings concerning grant-review systems and suggestions for improvements.

* * * * *

My thanks to Stephen A. Bonaduce and Terri Freedman for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** Refereeing and peer review. Part 1. Opinion and conjecture on the effectiveness of refereeing. *Current Contents* (31):3-11, 4 August 1986.
2. ————. Refereeing and peer review. Part 2. The research on refereeing and alternatives to the present system. *Current Contents* (32):3-12, 11 August 1986.
3. **Zuckerman H & Merton R K.** Patterns of evaluation in science: institutionalisation, structure and functions of the referee system. *Minerva* 9:66-100, 1971. [Reprinted as: Institutionalized patterns of evaluation in science. (Merton R K.) *The sociology of science.* Chicago, IL: University of Chicago Press, 1973. p. 460-96.]
4. **Kronick D A.** Authorship and authority in the scientific periodicals of the seventeenth and eighteenth centuries. *Libr. Quart.* 48:255-75, 1978.
5. **Levy J A.** Peer review: the continual need for reassessment. *Cancer Invest.* 2:311-20, 1984.
6. Public Health Service Act. (PL 410, 1 July 1944), *United States Statutes at Large*, 58, p. 682-711.
7. **National Institutes of Health.** *1983 NIH almanac.* Bethesda, MD: NIH, 1983. p. 5. NIH Publ. No. 83-5.
8. **Roy R.** Funding science: the *real* defects of peer review and an alternative to it. *Sci. Technol. Hum. Val.* 10(3):73-81, 1985.
9. **Atkinson R C & Blanpied W A.** Peer review and the public interest. *Issues Sci. Technol.* 1(4):101-14, 1985.
10. **Raub W F.** Personal communication. 12 December 1986.
11. **Powledge T M.** NIH's Raub on misconduct. *The Scientist* 15 December 1986. p. 18-9.
12. **Garfield E.** Introducing *The Scientist*: at last, a newspaper for the science professional. *Current Contents* (29):3-6, 21 July 1986.
13. ————. *The Scientist*: how it all began. *Current Contents* (33):3-6, 18 August 1986.
14. **Mathewson J.** More controversy over peer review. *Sci. News* 128:71, 1985.
15. **Sinsheimer R L.** Letter to editor. (Peer review and the public interest.) *Issues Sci. Technol.* 2(1):9-10, 1985.
16. **Long J.** Funding bill stirs academic research issue. *Chem. Eng. News* 64(24):12, 1986.
17. **Roy R.** Personal communication. 29 November 1986.
18. **Sanders H J.** Peer review. How well is it working? *Chem. Eng. News* 60(11):32-43, 1982.
19. **Kirschstein R L, Akers R P, Brooks G T, Fretts C A, Gary N D, Goldwater W H, Green J G, Solowey M, Kaufman A A, Raub W F, Russell G F, Riseberg R J, Schlaffino S S & Wilson K S.** *Grants peer review: report to the director, NIH. Phase I.* Washington, DC: NIH, 1976. 226 p.
20. **Cole S, Rubin L & Cole J R.** *Peer review in the National Science Foundation: phase one of a study.* Washington, DC: National Academy of Sciences, 1978. 193 p.
21. **Committee on Science and Public Policy of the National Academy of Sciences.** Research in Europe and the United States. (National Science Foundation) *The 5-year outlook on science and technology, 1981.* Washington, DC: NSF, 1981. Vol. 1. p. 255-84.
22. **Roy R.** Alternatives to review by peers: a contribution to the theory of scientific choice. *Minerva* 22:316-26, 1984.
23. **Culliton B J.** Fine-tuning peer review. *Science* 226:1401, 1984.
24. **Osmond D H.** Malice's Wonderland: research funding and peer review. *J. Neurobiol.* 14:95-112, 1983.
25. **Yalow R S.** Peer review: some suggestions. *Chem. Eng. News* 57(40):5, 1979.
26. ————. Peer review and scientific revolutions. *Biol. Psychiat.* 21:1-2, 1986.
27. **Horrobin D F.** Referees and research administrators: barriers to scientific research? *Brit. Med. J.* 2:216-8, 1974.

EUGENE GARFIELD:

Refereeing and Peer Review. Part 4.

Research on the Peer Review of Grant Proposals and Suggestions for Improvement

Current Contents, February 2, 1987

This is the conclusion to a four-part series on refereeing and peer review in science. The first two parts discussed the refereeing of scholarly articles prior to publication.[1,2] The third part focused on the mechanics of the peer-review system for the evaluation of research-grant proposals at the National Science Foundation (NSF) and the National Institutes of Health (NIH) and opinions about those systems.[3] This part examines the research on peer review and some proposed alternatives and improvements.

The COSPUP Study

One of the best-known and most thorough studies of peer review was conducted by sociologists Stephen Cole and Leonard Rubin, State University of New York (SUNY), Stony Brook, and Jonathan R. Cole, Columbia University, New York. At the request of the Committee on Science and Public Policy (COSPUP) of the US National Academy of Sciences (NAS), the Coles and Rubin examined the peer-review system of the NSF. Phase one of the study, a retrospective statistical analysis of "how peer review works in the day-to-day operation of the Foundation [NSF],"[4] (p. 17) was started in 1974 and completed in 1978. Phase two, coauthored by the Coles and COSPUP, reported the results of experiments designed to address the question of whether program officers influence the peer-review process through their selection of reviewers. It was started in 1978 and published in 1981.[5]

In phase one, the authors interviewed 70 scientists involved in all stages of the peer-review process, including current and former NSF program directors, advisory- and review-panel members, NSF section and division heads, and the director and associate director of the NSF.[4] (p. 18) To determine the most decisive factors in securing a grant, they collected data on 1,200 applicants, half of whom had been successful. In some cases, the authors examined not only the proposal but also the reviewers' comments, correspondence, and all paperwork connected with the funding decision.

Phase two was carried out in two stages. First, the Coles submitted 150 proposals previously reviewed by the NSF to a set of surrogate program directors. Half of the surrogates received proposals that had been edited in an attempt to conceal the authors' identities; the other half received copies that were exactly as they had been submitted to the NSF. The surrogate directors were asked to name a set of possible referees for the proposals, and the Coles once again attempted to conceal the identities of half the authors. None of the participants knew how the proposals had been rated by the NSF. The Coles asked them not only to evaluate the proposals but also, when applicable, to try to identify the authors. Reviewers of "blinded" proposals were also asked whether the removal of title pages, lists of references, budgets, and other identifying comments made the proposal more difficult to evaluate.[5] (p. 6-19)

The COSPUP Findings

The main conclusion of phase one is that peer review in the NSF functions fairly.[4] (p. viii-ix) The authors found a high correlation between high reviewer ratings and favorable funding decisions. They also found that an applicant's age and track record had little effect on the chances of getting a grant and that reviewers from major, "high-status" institutions treated proposals from researchers at prestigious institutions no differently than proposals from workers at less-prestigious institutions.

On the whole, the results of phase two corroborate the findings of phase one: the Coles found no evidence of bias on the part of program officers in their selection of reviewers and no evidence that external criteria such as gender, age, and race had any influence on reviewer decisions.[5] (p. 4) In the matter of blinded proposals, the Coles found it difficult to conceal authorship: "To omit all possible identifiers, in addition to the name(s) of the author(s) of the proposal, made the proposal almost unreadable," said Jonathan Cole.[6] This was reflected by the opinions of the COSPUP reviewers, who felt that the blinding process severely compromised the integrity of the proposals. Nevertheless, proposals that received high ratings by NSF reviewers generally received high ratings from COSPUP reviewers as well.

However, "there was a great deal of well-considered variance in opinions among equally qualified reviewers," in the words of Jonathan Cole.[6] "Thus, if we work with a small number of reviewers and a high variance in opinion, the outcome of an evaluation will depend greatly on the people selected to review the proposal.... This is not to imply that the process is 'unfair,' but that there is a substantial level of reviewer disagreement on rational grounds, e.g., quality of past work, priority given by a particular reviewer to the subject of the proposal, the assessment of the methods to be used, etc."[6]

The Coles concluded that perhaps 25 to 30 percent of NSF funding decisions would be reversed if applications were evaluated by another, equally qualified group of reviewers. In both their phase-two monograph[5] and a paper they published in *Science* with statistician Gary A. Simon, SUNY, Stony Brook,[7] the Coles acknowledge that some scholars, taking note of this, will feel that the complicated system of peer review "does not buy you much."[5] (p. 43) Jonathan Cole points out, however, that "there is apt to be a great deal of disagreement on the contents of proposals...at the cutting edge of scientific inquiry,...and therefore, we should not be wholly surprised at the proportion of reversals."[6] Such reversals probably indicate that no "single, agreed-upon dogma"[7] is dominant in the fields studied, and, in fact, one of the most surprising results of the COSPUP study was that the level of consensus among reviewers was no higher in physics than in the social sciences.[4-7]

Phase one of the COSPUP study has been cited in over 74 papers since it appeared in 1978; phase two has been cited in 17. The *Science* article has been cited 49 times through 1986 and is one of four papers forming the core of a research front entitled "Alternatives to, arbitration in, and other aspects of peer review of scientific journals and research proposals" (#85-4243). The other three core papers include the classic study of patterns of evaluation in science by Harriet Zuckerman and Robert K. Merton, Columbia University;[8] a controversial study of bias in the journal refereeing process by Douglas P. Peters, University of North Dakota, Grand Forks, and Stephen J. Ceci, Cornell University, Ithaca, New York;[9] and a paper on the rate of agreement between reviewers by psychologist Grover J. Whitehurst, SUNY, Stony Brook.[10] All three papers were mentioned in Parts 1 and 2 of this essay.[1,2]

According to Jonathan Cole, the key policy implication of the COSPUP study was

that "the lower the number of reviewers used to evaluate the proposal, the greater the chance for...reversals."[6] As a result, the NSF now requires a certain minimum number of reviewers for every proposal it receives. Science journalist Tineke Boddé lists a number of other changes in the NSF system that have been made more recently.[11] For instance, the entire process has been streamlined, with a limit of 15 pages per proposal and a policy requiring a decision within nine months. Specific guidelines on conflicts of interest have been established, verbatim copies of all reviewer comments have been made available, and a system has been set up to reconsider declined proposals. Under certain circumstances, some proposals are now exempt from peer review, and program officers can extend existing grants without further review if they feel outstanding progress has been made.[11]

Peer Review in the NIH

Fourteen scientists and administrators from various agencies within the NIH were appointed to the NIH Grants Peer Review Study Team by then-acting director, Ronald W. Lamont-Havers. Chaired by Ruth L. Kirschstein, director, National Institute of General Medical Sciences, the team was charged with evaluating the NIH's peer-review system and with making, where applicable, recommendations for improvement.[12] In making its assessment, the team printed an open solicitation in the *Federal Register*[13] and mailed a memorandum to 30,000 individuals, asking for written comments on the peer-review system (1,500 replies were received). The team also held open public hearings for the scientific and lay communities. The team members considered everything they read and heard, according to William F. Raub, team member and deputy director, NIH, but the project was an informal survey and, ultimately, the recommendations the team made were based on a consensus of its members' best judgments.[14]

Virtually every recommendation made by the study team has been implemented.[14] Among these was the suggestion that guidelines on conflicts of interest and a formal appeals system for the reconsideration of rejected proposals be established. In addition, as part of the appeals procedure, the team suggested that specific criteria be established for reevaluating proposals and that an independent ombudsman be appointed to adjudicate disputes between the NIH and applicants. A change in NIH procedure that was recently instituted is the creation of two programs allowing the life of a grant to be extended for up to 10 years under certain very limited circumstances.[15]

In connection with the report by the NIH study team, Jonathan Cole suggests that a fruitful area for research would be a rigorous comparison of the NIH study-section approach to peer review with the individual approach used by the NSF. He says that "panels can evaluate the *relative* strength of a set of proposals, but, in fact, each panel member, while voting on all, actually only reads a few. This leads potentially to an artificial consensus, where a couple of strong characters on the panel dominate the decision-making process."[6]

Studies of Scholars' Attitudes Toward Peer Review

Sociologist Gilbert W. Gillespie, Cornell University; Daryl E. Chubin, director, Technology and Science Policy Program, Georgia Institute of Technology, Atlanta; and physician George M. Kurzon studied the factors that help shape applicants' attitudes toward the system.[16] The authors expected to find that those who experienced success in obtaining funding would tend to be satisfied with the *status quo* and that those who failed to obtain funding would tend to blame the system.

Gillespie and colleagues sent a three-page, 19-item questionnaire to 719 researchers whose proposals had been approved or rejected by the National Cancer Institute of the NIH in 1980 and 1981. The questionnaire stated that those who did not return the survey would be assumed to be satisfied with peer review, so the authors find it noteworthy that 336 (47 percent) responded—although they do not presume that satisfaction with the system was the only reason for nonresponse. It is also interesting, the authors said, that 205 (61 percent) of the responses came from scholars whose proposals had been funded, since they expected a heavier response from scholars who had been denied funding.[16] Because the questionnaire was sent to researchers who had recently submitted proposals for review, it could not measure the attitudes of those whose discontent with the system had led them to give up submitting proposals.

As the authors expected, previous success in obtaining funding was found to be inversely proportional to a desire to change the system. Gillespie and colleagues also found that those who have been unsuccessful until very recently in obtaining funding tended to support the current process, while those who had been successful in the past but who had recently been denied funding tended to favor modifications to the system. The authors also concluded that several complaints about the peer-review system reflected a surprising ignorance of the procedures governing the operation of the system. For instance, those who believe that cronyism or old-boy networks control the process fail to take into account the limited time that an individual may serve in a review group and the NIH's strict requirements concerning the makeup of such groups, which ensure a balanced cross section of scientists that changes constantly.[16] Jonathan Cole points out, however, that the choice of a given individual reviewer from among a number of roughly comparable candidates "can be a function of social and intellectual ties with study-section members."[6]

Flaws in the System?

There may be instances in which peer review operates with unintended blind spots or unsuspected inefficiency. Alan L. Porter, School of Industrial and Systems Engineering, and Frederick A. Rossini, School of Social Sciences, Georgia Institute of Technology, studied the fate of proposals that "fall between the cracks" of the NSF's disciplinary programs.[17] After analyzing 257 reviews received by 38 approved, cross-disciplinary proposals in five different subject areas, they found that reviewer decisions were more favorable when the proposal fell within the reviewer's own discipline. In discussing this finding, the authors found it reasonable "for a reviewer of proposed research to favor that which is more familiar.... In such a case, one is apt to understand better what is planned; one may know the researchers personally or by reputation, and hence appreciate their expertise; and one can feel more secure in making strong recommendations."[17] Porter and Rossini conclude that interdisciplinary research proposals should not be reviewed in the same way as disciplinary projects.

A study by Anthony S. Russell, associate professor of medicine, and Michael Grace, both at the University of Alberta, Edmonton, and Bonnie D. Thorn, director of finance and administration, Arthritis Society, Toronto, Canada, supports the widespread belief that the peer-review process is unnecessarily long and complex.[18] Russell and colleagues examined 113 grant applications to the Arthritis Society, a national voluntary health organization, to determine whether there were any substantial differences between the initial assessment each proposal received in-house and the detailed, out-of-house review that followed. They found that in-depth reviews had little impact on the original rating, implying that review procedures that operate in a similar, two-tiered fashion could be greatly streamlined.[18] And in fact, in an analysis of nearly 1,400 reviews of about 200 NSF proposals, David Klahr, Carnegie-Mellon University,

Pittsburgh, Pennsylvania, found that independent mail reviewers had little impact on the final rating given to a proposal by panel reviewers.[19]

Suggestions for Further Change and Improvement

As I mentioned earlier, both the NSF and the NIH have instituted changes in their review procedures over the last few years. Nevertheless, there are plenty of suggestions for changing the system. Unfortunately, since so little empirical data exist, most of these suggestions are little more than remedies for *perceived* ills. It is hard to know which ones are worth implementing without further research.

One suggested change concerns the time and effort consumed by writing proposals and filling out forms. Typical of many scientists' feelings is a remark attributed to Nobel laureate biochemist Albert Szent-Györgyi (1893-1986). In an article published in *Chemical & Engineering News*, science journalist Howard J. Sanders reports that Szent-Györgyi once remarked that writing grant proposals filled his "scientific life with agony."[20] Rosalyn S. Yalow, Veterans Administration, New York, the 1977 Nobel laureate in physiology or medicine, suggests that researchers of demonstrated ability should not have to go through the process of making a formal application year after year for the renewal of funding.[21] Instead, they should receive a constant level of funding that is renewable every three years, subject to review of their progress.[22]

Rustum Roy, director, Science, Technology and Society Program, Pennsylvania State University, University Park, also wants research funding to be based on an investigator's performance.[23] But in a departure from other scholars' suggestions, he proposes a formula, "based on three kinds of post-hoc peer review,"[24] on which to base grants to individuals, university departments (or research units of a similar size), and institutions.[23,25,26] Roy claims his peer-review formula would eliminate the subjective elements of allocating grant money and does not tie funds to specific projects; instead, money would be administered at the departmental level and would be distributed based on a researcher's past performance, rather than on future promise (with allowances to be made for new or young investigators without track records).[26] Henry R. Hirsch, Department of Physiology and Biophysics, College of Medicine, University of Kentucky, Lexington, also proposed that all active faculty members ought to receive funding, varying to reflect the administration's judgment concerning "the costs and merits of different kinds of research."[27]

Roy's proposal met with considerable individual criticism. In a number of letters written in reply to his original editorial in *Science*,[23] various scientists expressed misgivings about jettisoning the "informed judgment"[28] and the concern with quality that they feel are intrinsic to peer review in its present form.[29,30] But in his reply, Roy says these objections assume that peer review "is in some mysterious way linked with the progress of science" and that the process can accurately predict the quality of research not yet performed. Roy states that both claims are totally unsupported.[31]

Another funding alternative to peer review, supported by a "small but vocal number of scientists," as Sanders puts it,[20] involves block grants, a system common throughout Europe,[3] in which funds are awarded to a research institution for allocation as it sees fit. The money would not go directly to an individual; instead, distribution would be determined by department heads or administrative officials. But Sanders notes that most US scientists strongly oppose a block-grant system, in the belief that a department head or administrative official or committee is less qualified to decide how to allocate research funds than an expert peer-review group.[20] Moreover, according to Joshua Lederberg, president, Rockefeller University, a block-grant system

would merely substitute "the politics of the institutions for the politics of the review committees."[32] And the people who make the funding decisions not only won't be anonymous to those in need of funds, they will have to live and work with them daily, and thus, as Sanders writes, "are less apt to make their choices impartially."[20]

Some scientists also question the underlying assumption of the present peer-review system: that only experts from an applicant's field or a closely allied discipline are qualified to judge that research proposal. David Apirion, Department of Microbiology and Immunology, Washington University School of Medicine, St. Louis, Missouri, suggests the creation of a class of professional, salaried science reviewers to replace peer review.[33] As Apirion puts it, "In all other branches of human creative enterprise [such] as literature, music, sculpture, etc., the producers of new works as well as the performers of new and old works are often judged by a special class of persons, reviewers and critics, who are seldom actively involved in the expansion of the particular discipline that they are entrusted to judge and evaluate."[33]

Pressures on the Peer-Review System

Several authors made observations concerning peer review that bear emphasizing. Yalow pointed out that there is a certain deadening effect—or dishonesty—inherent in trying to explain or justify research that has yet to be done; if your project is so low-risk that you already know what you expect to find, Yalow asks, then how original or important can it be?[21,22] Daniel H. Osmond, University of Toronto, notes that there may be a certain amount of pressure, once funding is approved, to "groom" research results to fit the expectations of the granting agency.[34] Perhaps the biggest problem with peer review, however, isn't really a problem with peer review at all, but rather with the amount of funding available. In the "golden years" of the 1950s and 1960s, money for research was relatively plentiful and granting agencies generous; now, with money tight and with so many applicants, even deserving projects are sometimes denied funding.[35] As Lederberg says, "When there's not enough [money] to go around, some people are inevitably hurt—sometimes arbitrarily and unfairly."[32] Frustration with such decisions carries over to the system by which the decisions are rendered.

Obviously, the process of peer review grinds on in spite of such troubling issues. There was a consensus of views expressed by scientists interviewed for Sanders's wide-ranging special report. In spite of all the complaints and all the faults hinted at, peer review is still considered the best method by which society places its bets on the most fruitful research.[20] Yet the credibility of peer review in the eyes of both the public and the scientific community is threatened by the activities of those who lobby Congress directly for funds. Richard C. Atkinson, former director, NSF, and currently chancellor, University of California, San Diego, and physicist William A. Blanpied, currently international studies specialist at the NSF, warn that the abandonment of peer review might reduce science to just another special-interest group, with funds being allocated based on political acumen rather than on a consensus of what best serves the advancement of scientific knowledge.[35] To prevent more institutions from joining those that have already abandoned the system, further changes in peer review may be necessary. But we should not confuse the forest with the trees. Without a strong peer-review system, albeit constantly reexamined, science might become tentative and inefficient.

* * * * *

My thanks to Stephen A. Bonaduce and Terri Freedman for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** Refereeing and peer review. Part 1. Opinion and conjecture on the effectiveness of refereeing. *Current Contents* (31):3-11, 4 August 1986.
2. ----------. Refereeing and peer review. Part 2. The research on refereeing and alternatives to the present system. *Current Contents* (32):3-12, 11 August 1986.
3. ----------. Refereeing and peer review. Part 3. How the peer review of research-grant proposals works and what scientists say about it. *Current Contents* (4):3-8, 26 January 1987.
4. **Cole S, Rubin L & Cole J R.** *Peer review in the National Science Foundation: phase one of a study.* Washington, DC: National Academy of Sciences, 1978. 193 p.
5. **Cole J R, Cole S & the Committee on Science and Public Policy, National Academy of Sciences.** *Peer review in the National Science Foundation: phase two of a study.* Washington, DC: National Academy Press, 1981. 106 p.
6. **Cole J R.** Personal communication. 13 November 1986.
7. **Cole S, Cole J R & Simon G A.** Chance and consensus in peer review. *Science* 214:881-6, 1981.
8. **Zuckerman H & Merton R K.** Patterns of evaluation in science: institutionalisation, structure and functions of the referee system. *Minerva* 9:66-100, 1971. [Reprinted as: Institutionalized patterns of evaluation in science. (Merton R K.) *The sociology of science.* Chicago, IL: University of Chicago Press, 1973. p. 460-96.]
9. **Peters D P & Ceci S J.** Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain Sci.* 5:187-95, 1982.
10. **Whitehurst G J.** Interrater agreement for journal manuscript reviews. *Amer. Psychol.* 39:22-8, 1984.
11. **Boddé T.** Evaluation of peer review draws mixed reactions. *BioScience* 32:10-2, 1982.
12. **Kirschstein R L, Akers R P, Brooks G T, Fretts C A, Gary N D, Goldwater W H, Green J G, Solowey M, Kaufman A A, Raub W F, Russell G F, Riseberg R J, Schiaffino S S & Wilson K S.** *Grants peer review: report to the director, NIH. Phase I.* Washington, DC: NIH, 1976. 226 p.
13. NIH grants peer review study team: establishment. (FR Doc. 75-23368). *Fed. Reg.* 40:40870, 1975.
14. **Raub W F.** Personal communication. 12 December 1986.
15. **Culliton B J.** NIH starts new grants program. *Science* 232:566, 1986.
16. **Gillespie G W, Chubin D E & Kurzon G M.** Experience with NIH peer review: researchers' cynicism and desire for change. *Sci. Technol. Hum. Val.* 10(3):44-54, 1985.
17. **Porter A L & Rossini F A.** Peer review of interdisciplinary research proposals. *Sci. Technol. Hum. Val.* 10(3):33-8, 1985.
18. **Russell A S, Thorn B D & Grace M.** Peer review: a simplified approach. *J. Rheumatol.* 10:479-81, 1983.
19. **Klahr D.** Insiders, outsiders, and efficiency in a National Science Foundation panel. *Amer. Psychol.* 40:148-54, 1985.
20. **Sanders H J.** Peer review. How well is it working? *Chem. Eng. News* 60(11):32-43, 1982.
21. **Yalow R S.** Peer review and scientific revolutions. *Biol. Psychiat.* 21:1-2, 1986.
22. ----------. Peer review: some suggestions. *Chem. Eng. News* 57(40):5, 1979.
23. **Roy R.** An alternative funding mechanism. *Science* 211:1377, 1981.
24. ----------. Personal communication. 29 November 1986.
25. ----------. Funding science: the *real* defects of peer review and an alternative to it. *Sci. Technol. Hum. Val.* 10(3):73-81, 1985.
26. ----------. Alternatives to review by peers: a contribution to the theory of scientific choice. *Minerva* 22:316-26, 1984.
27. **Hirsch H R.** *A proposal for per capita distribution of research funds with administrative flexibility.* 1983. 5 p. (Unpublished paper.)
28. **Liebman J C.** Letter to editor. (Alternative to peer review?) *Science* 212:1336, 1981.
29. **McCreery R L.** Letter to editor. (Alternative to peer review?) *Science* 212:1336, 1981.
30. **Kalt M R.** Letter to editor. (Alternative to peer review?) *Science* 212:1336-7, 1981.
31. **Roy R.** Letter to editor. (Alternative to peer review?) *Science* 212:1338-9, 1981.
32. **Lederberg J.** Personal communication. 17 August 1986.
33. **Apirion D.** Letter to editor. (Research funding and the peer-review system.) *Fed. Proc.* 38:2649-50, 1979.
34. **Osmond D H.** Malice's Wonderland: research funding and peer review. *J. Neurobiol.* 14:95-112, 1983.
35. **Atkinson R C & Blanpied W A.** Peer review and the public interest. *Issues Sci. Technol.* 1(4):101-14, 1985.

DOMENIC V. CICCHETTI:

The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation

Behavioral and Brain Science, **14** (1991) 119-186

Abstract: The reliability of peer review of scientific documents and the evaluative criteria scientists use to judge the work of their peers are critically reexamined with special attention to the consistently low levels of reliability that have been reported. Referees of grant proposals agree much more about what is *unworthy* of support than about what does have scientific value. In the case of manuscript submissions this seems to depend on whether a discipline (or subfield) is general and diffuse (e.g., cross-disciplinary physics, general fields of medicine, cultural anthropology, social psychology) or specific and focused (e.g., nuclear physics, medical specialty areas, physical anthropology, and behavioral neuroscience). In the former there is also much more agreement on rejection than acceptance, but in the latter both the wide differential in manuscript rejection rates and the high correlation between referee recommendations and editorial decisions suggests that reviewers and editors agree more on acceptance than on rejection. Several suggestions are made for improving the reliability and quality of peer review. Further research is needed, especially in the physical sciences.

Keywords: cross-disciplinary comparisons; evaluation; grant review; manuscript reviews; peer review; quality control; reliability

1. Objectives

This paper will analyze the peer-review process in the evaluation of manuscript submissions and grant applications. First, we will discuss research designs and statistical procedures, and then we will critically review the major studies of peer review across disciplines, providing some reasons and remedies for the low reliability of manuscript and grant reviews as well as some suggestions for future research.

2. Theoretical Issues

Gottfredson (1978, p. 920) has stressed the importance of peer evaluation in scientific activity from a Kuhnian standpoint (Kuhn 1962). Until the beginning of the nineteenth century, scientific theory was thought to be an approximation of what Laudan (1984, p. 83) referred to as "absolute truth," "certainty," or "infallible knowledge." By the twentieth century, this view of scientific theory was replaced by the more modest goal of developing theories that were, again in Laudan's (1984, p. 83) terminology "plausible," "probable," or "well-tested." Laudan (p. 83) notes that this paradigm shift "represents one of the great watersheds in the history of scientific philosophy: the abandonment of the quest for certainty." Kuhn's ideas about paradigm development and paradigm "shifts," although undergoing reevaluation and reinterpretation almost since their inception (e.g., Boehme 1977; Boehme et al. 1976; Gholson & Barker 1985; Lakatos 1972; Laudan 1984; Mulkay 1977; Price 1963), continue to play a central role in our understanding of the evaluation of scientific work by the community of fellow scientists or "peers" (e.g., Mahoney 1985).

In a classic work, Robert Merton argued that the social system governing both the actions and the mobility of scientists is very fair and objective, supporting a *normative model* of science (Merton 1973): A scientist's work is judged for scientific merit on the basis of universal scholarly standards rather than by specific biases such as friendship, author affiliation, or nepotism (e.g., see Lindsey 1978, p. 55). As Lindsey (1978, p. 3) reminds us, however, Merton and his students (e.g., Cole & Cole 1973) were focusing mainly on the physical sciences. The normative model does not appear to hold well for either the behavioral or the medical sciences. In fact, as we shall discuss later, there are data to suggest that the model is not entirely appropriate for the physical sciences, either.

Peer review is a system of decision making by referees, editors, and research program directors in evaluating the quality of scientific research. It is here that Merton's normative model applies to the attributes that are used in evaluating papers submitted to professional organizations, manuscripts submitted to scientific journals, and research proposals submitted to funding agencies. These attributes can be derived from either objective judgments (e.g., experimental design) or subjective ones (e.g., importance). The attributes themselves must be distinguished from the criteria (or norms) used to judge them. Thus, scientists might use the criterion "brief," "to

the point," or "excessively verbose" to judge the succinctness (attribute) of a given manuscript. In peer review, referees evaluate the attributes of scientific documents according to sets of specific criteria. Then editors or granting officials apply additional evaluation criteria to reviewers' reports to decide whether or not to accept a manuscript or fund a proposal.

A number of evaluation criteria are relevant to the review of manuscripts as well as grant proposals. For example, reviewers are usually expected to use criteria to assess (1) the relevance and completeness of the review of the research literature; (2) the author's level of originality or imaginativeness; (3) the adequacy of the research methodology; (4) the data-analytic strategies; (5) the importance (usefulness) of actual or expected findings; and (6) the clarity and organization of the information the author presents.

Other attributes are specific either to manuscript or grant review: Reviewers and editors must judge a manuscript's level of interest to the readership of the journal, whether its length is justified, and how much space is available in the journal. Grant reviewers and program directors must judge: the applicant's prior scientific contributions (or "track record"); the adequacy of the institutional setting in which the research would be undertaken; the appropriateness of the budget request relative to the objectives stated in the proposal; and the availability of funds.[1]

3. Empirical Issues: Methodology and data-analytic strategies

3.1. Research designs used in peer-review studies. A wide spectrum of research designs has been used in studying peer review, including:

1. Qualitative or semiquantitative studies of reviews of selected journal manuscripts (Ingelfinger 1974; McCartney 1978; Patterson 1969; Smigel & Ross 1970)
2. Quantitative studies of hypothetical reviews of manuscripts (requiring referees to evaluate or rank order the value of normative attributes "as if" they were applying them to actual submitted manuscripts; e.g., Kerr et al. 1977; Lindsey 1978; Rowney & Zenisek 1980)
3. Quantitative naturalistic studies of the reliability of referee reports on scientific documents, including papers submitted to professional societies (e.g., Cicchetti & Conn 1976; Conn 1974), journal manuscripts (Cicchetti 1980; Cicchetti & Conn 1978; Cicchetti & Eron 1979; Hargens & Herting 1990a; Ingelfinger 1974; Lock 1985; Orr & Kassab 1965; Scott 1974; Smigel & Ross 1970; and Whitehurst 1983; 1984), and grant proposals (Cole & Cole 1981; 1985; Cole et al. 1978; 1981)
4. Quasi-experimental studies (Peter & Ceci 1982)
5. Experimental studies of the reliability of the peer-review process (Armstrong 1980; 1982a; Goodstein & Brazis 1970; Mahoney 1977; 1978; and Mahoney et al. 1978).

The distinction between quasi-experimental and experimental studies is based on the extent to which alternative interpretations of a given result can be ruled out. We agree with Peters & Ceci (1982, p. 246) that "the quasi-experimental design . . . is, in general, insufficient to rule out alternative explanations unequivocally," but we also agree with the same authors (p. 247) "that quasi-experimental designs can provide causal inferences when used along with convergent and cogent reasoning and analysis."

In a broader sense, when conclusions drawn from quasi-experimental and experimental studies are consistent with those from less well controlled studies (such as the first three research designs just described), one can be more confident that the missing controls did not materially influence the direction or quality of the results. This point will be reemphasized later when we compare conclusions from peer-review studies differing widely in how well potentially relevant variables were controlled.

3.2. Types of reliability assessments. One purpose of this paper is to examine the reliability of the peer-review process. Accordingly, it is important to analyze how reliability has been measured and what statistical approaches have been used. Depending on the specific research question, any of several types of reliability measures could be appropriate: internal consistency, interreferee agreement, or even stability across time. The most common measure is interreferee agreement at a single point in time.

Interreferee reliability is defined as the extent to which two or more independent reviews of the same scientific document agree. To choose an appropriate statistic for assessing levels of interreferee agreement, it must be known whether or not the same referees evaluated the documents under study and whether or not the same number of referees evaluated a given document. The statistic should also assess how much referee agreement is influenced by chance alone (e.g., Watkins 1979). Finally, the scale of measurement by which the data are expressed needs to be identified.

3.3. Appropriate statistics. Which reliability statistics are appropriate will vary according to whether the reviewers evaluate papers for presentation at scientific meetings, manuscripts submitted to professional journals, or grant proposals submitted for research funding.

Papers submitted to scientific meetings are sometimes all evaluated by the same two referees, since the scientific documents are usually rather brief (e.g., extended abstracts). Here, either the unweighted kappa statistic (Cohen 1960) or the weighted one (Cohen 1968) would be appropriate.[2] The choice would depend on the evaluative scale available to the referees. A nominal dichotomous scale such as "accept" or "reject" would require unweighted kappa, whereas an ordinal or rank-ordered evaluative scale, such as one ranging from "reject" to "accept only if time and space are available" to "accept unconditionally" would require the weighted kappa statistic. When the same three or more referees all independently evaluate the same set of papers, then the intraclass correlation coefficient (R_i), Model II would be appropriate (e.g., see Bartko 1966; 1974; 1976; Bartko & Carpenter 1976; Cicchetti et al. 1976; Cicchetti & Conn 1976; Fleiss 1981).[3]

Manuscripts submitted to professional journals are evaluated by different sets of reviewers, since it is obviously not feasible for the same two or more reviewers to undertake all the assessments. A statistic of choice here would be Model I of the R_i.[4] (For more information about

mathematical relationships between kappa and Models of R_i, see (a) Fleiss 1975 for the *nominal-dichotomous* case; and (b) Fleiss & Cohen 1973; and Krippendorff 1970 for the *ordinal* case.)

For the peer review of grant proposals, some granting agencies have used different sets of reviewers with the same number throughout (e.g., the American Heart Association, as described in Wiener et al. 1977): This is analogous to manuscript review; other granting agencies (e.g., National Science Foundation [NSF] as described in Cole & Cole 1981), however, not only use different sets of reviewers for each evaluated document, but the number of reviews varies from one proposal to the next. This design calls for R_f Model III based on the average number of reviews per proposal (e.g., see Bartko & Carpenter 1976; Cicchetti & Showalter 1988).[5]

It should also be mentioned that Gilmore (1979, based on an earlier approach reported in Garner & McGill 1956) has described yet another statistic for assessing the reliability of peer review: It fits the case of a dichotomous decision (e.g., "accept" or "reject"), with two ratings per document and does not distinguish between ratings all made by the same pair of referees and those made by different pairs of referees. Gilmore notes that the statistic is "very similar to the percentage of explained variance." The statistic therefore has some conceptual similarity to Lambda (due to Goodman & Kruskal 1954) a statistic that, with minor modifications, has been shown by Fleiss (1975) to be mathematically equivalent to kappa in the dichotomous case.

3.4. Inappropriate statistics. Two additional statistical tests have been applied, on occasion, to assess levels of interrater reliability for manuscript review. Both tests suffer from major defects. The first is the standard Pearsonian product moment correlation (R). This statistic assesses the extent to which two independent sets of ratings (e.g., manuscript or grant reviews) covary in the same order, but it ignores the extent to which given pairs of reviewers disagree on any single evaluation (e.g., see Bartko 1966; 1974; 1976; Bartko & Carpenter 1976; Kazdin 1982; Robinson 1957). In the specific context of journal manuscript reviews, Hendrick (1976; 1977) was able to demonstrate artifactually inflated levels of reviewer agreement when the Pearson *r*, rather than R_i, was used to make the reliability assessment.

Recently, Whitehurst (1983; 1984) reintroduced another statistic for assessing levels of referee consensus. The statistic was developed by Finn (1970) and can be symbolized by R_f. The mathematical difference between R_f and R_i (or kappa) statistics derives from an underlying assumption about chance agreement levels between any set of raters. Statistics such as R_f use levels of chance agreement that assume that "every judgment has the same probability of occurring under the hypothesis that the judges have no understanding of the scale applied and their ratings are purely random" (e.g., Lawlis & Lu 1972, pp. 17–18). In the specific context of manuscript review, this would mean that the recommendation to accept, reject, or resubmit a specific article would occur equally frequently, by chance alone. Given the known high rejection rates of many journals (often in excess of 80%), this definition of chance agreement cannot be valid (e.g., see Cicchetti 1985). Consistent with this argument, it has recently been shown that R_f (but *not* R_i or kappa) would fail to distinguish chance reviewer agreement from substantially higher levels (i.e., see again Cicchetti 1985).[6]

4. Empirical Issues: Major studies in peer review

4.1. Evaluative criteria: Scientists judge their value. Five studies are briefly considered here. Each bears on how scientists place weight on the various evaluative criteria we have mentioned. All five studies examined (1) the "importance" of the study to the field and (2) the perceived adequacy of the "research design" on their rating lists; otherwise, they were quite different. Two used "as if" designs for major behavioral science manuscripts and depended on mail responses, but the journals they studied were not the same ones; response rates also varied widely (50% in Wolff 1970, and 82% in Lindsey 1978). Two other studies used actual manuscript reviews, but again not the same journals (i.e., *Journal of Personality and Social Psychology* in Scott 1974, and the *Journal of Abnormal Psychology* in a study by Cicchetti & Eron 1979). The fifth study (Cicchetti & Conn 1976; Conn 1974) used only three referees who made "blind" assessments (author's identity unknown) of extended abstracts sent to a major professional medical society (The American Association for the Study of Liver Disease). The five studies also differed in data-analytic techniques. The "as if" studies asked referees to rank order the set of evaluative criteria as if they were being used for recommending the acceptance or rejection of a hypothetical manuscript. The remaining three studies used the size of the correlation between the ranking of a given evaluative criterion and the judged level of scientific merit of the document. Despite the extreme heterogeneity of these studies, all five indicated that the level of perceived "importance" of the contribution to the field and the perceived level of adequacy of the "research design" were the two most important evaluative criteria referees use for judging the merit of a given scientific document.

Although we are not aware of comparable studies on the peer review of grant proposals, information derived from a study (Weiner et al. 1977) of the reliability of reviews of grants submitted to the American Heart Association (AHA, New York State Affiliate) merits brief discussion. Primary reviewers (2 were assigned to each proposal) were given a set of 10 criteria to use in evaluating each grant. Each criterion received an *a priori* weight ranging from a low of 1 to a maximum of 2.5. Four criteria received the maximum weight. Three of them pertained to importance and research design issues. They were: (1) "The value of the expected data in increasing knowledge in a scientific field or in advancing the diagnosis and therapy of vascular disease"; (2) "Methodology: Is it valid and feasible?"; and (3) Research plan: (a) overall rationale; (b) quality of individual experiments, controls. (For further details, see Wiener et al. 1977, p. 307.)

4.2. Reliability of evaluative criteria. How well do pairs of referees agree in evaluating the relevance of criteria as they apply them to the same scientific documents? Available data for both manuscripts and abstracts (once again derived from several sources) are presented in Table 1 and indicate levels of interreviewer agreement. These

Table 1. *Levels of interreferee agreement (intraclass correlations) on various criteria applied to the evaluation of manuscripts and extended abstracts*

A. For manuscripts submitted to the "Journal of Abnormal Psychology" (1973–78)

Evaluative Criterion	Number of Manuscripts	Level of Interreferee Agreement
Importance	661	.23
Design	610	.32
Data Analysis	611	.22
Style and Organization	666	.22
Literature Review	660	.26
Reader Interest	663	.19
Succinctness	658	.30

B. For manuscripts submitted to the "Journal of Personality and Social Psychology" (Scott 1974)

Importance	312	.28
Design and Analysis	574	.19
Reader Interest	312	.07
Style and Organization	574	.25
Succinctness	574	.31
Literature Review	458	.37

C. For manuscripts submitted to the "British Medical Journal" (1979)

Importance	707	.33
Scientific Reliability	707	.22
Originality	707	.21
Suitability	707	.22

D. For abstracts submitted to "American Association for the Study of Liver Disease" (Cicchetti & Conn 1976)

Evaluative Criterion	No. of Abstracts	Levels of Interreferee Agreement			Composite Agreement
		A vs. B	*A vs. C*	*B vs. C*	
Importance	77	.22	.15	.31	.24
Design and Execution	77	.28	.21	.34	.29
Originality	77	.37	.21	.32	.30

Note: With the exception of "Reader Interest," Section B, all values are statistically significant at or beyond the .05 level.

range from essentially 0 (R = .07 for level of "reader interest") to "highs" of .37 for both "originality" and "literature reviews," which, according to guidelines representing levels of practical significance, would be considered *poor* (e.g., Cicchetti & Sparrow 1981; Fleiss 1981).

4.3. Reliability of manuscript reviews: Behavioral science. It can be seen from the data presented in Table 2A that the levels of chance-corrected interreviewer agreement (kappa or R_i values) range between .19 (*Journal of Abnormal Psychology* – Cicchetti & Eron 1979) and .54 (*American Psychologist* – Cicchetti 1980; Scarr & Weber 1978). It should be noted that the reviews for the *American Psychologist* were based on a very small number of manuscripts (N = 87), and that the level of peer-reviewer reliability could not be successfully replicated in a follow-up peer-review study (Cicchetti, unpublished), in which the R_i value dropped from .54 (fair agreement, see Cicchetti & Sparrow 1981) to .38 (poor agreement, Cicchetti & Sparrow op. cit., p. 133).

4.4. Reliability of manuscript and abstract reviews: Medicine. The data based on peer-reviewer chance-corrected reliability levels for medical journals (Table 2B) are very similar to those just presented for peer reviews of behavioral science manuscripts, namely a range between .31 (*Physiological Zoology* – Hargens & Herting 1990a) to .37 (a major medical subspecialty journal – Cicchetti & Conn 1978).

With respect to peer review of abstracts submitted to professional meetings, Cicchetti and Conn (1976) reported very similar levels of chance-corrected agreement for ratings of overall scientific merit (i.e., between .16 and .33, with corresponding p values between .10 and .01).

4.5. Reliability of manuscript reviews: Physical sciences. As far as we are aware, no formal studies of the reliability of peer review have been undertaken for manuscript or abstract submissions to journals in the physical sciences, yet there is a prevailing belief that levels of interreferee agreement are substantially higher for journals in the physical sciences than in other areas studied. This conclu-

Table 2. *Levels of reviewer agreement in the evaluation of the scientific merit of submitted manuscripts*

A. Behavioral Science

Journal	No. of Reviews	R_i or Kappa Value	Sources
"Social Problems" (1958–61)	193	.40[b]	Smigel & Ross (1970)
"Journal of Personality and Social Psychology"	286	.26[a]	Scott (1974)
"Sociometry"	140	.21[a]	Hendrick (1976)
"Personality and Social Psychology Bulletin"	177	.21[a]	Hendrick (1977)
"Journal of Abnormal Psychology" (1973–8)	1319	.19[a]	Cicchetti & Eron (1979; and unpublished)
"American Psychologist" (1977–8)	87	.54[a]	Cicchetti (1980); Scarr & Weber (1978)
"American Psychologist" (1978–9)	72	.38[a]	Cicchetti (unpublished)
"Journal of Educational Psychology" (1978–80)	325	.34[a]	Marsh & Ball (1981)
"Developmental Review"	72	.44[a]	Whitehurst (1983; 1984)
"American Sociological Review"	22	.29[a]	Hargens & Herting (1990)
"Law & Society Review"	251	.23[a]	Hargens & Herting (1990)

B. Medicine

Journal	No. of Reviews	R_i or Kappa Value	Sources
2 Untitled Biomedical Journals	1572	.34[b]	Orr & Kassab (1965)
"New England Journal of Medicine"	496	.26[b]	Ingelfinger (1974)
A Major Medical Subspeciality Journal	866	.37[a]	Cicchetti & Conn (1978)
"British Medical Journal"	707	.31[b]	Lock (1985)
"Physiological Zoology"	209	.31[a]	Hargens & Herting (1990)

Note: [a]Intraclass R values; [b]Kappa values; The criteria of Cicchetti & Sparrow (1981); Fleiss (1981); in which kappa or R_i values < .40=POOR; .40–.59=FAIR; .60–.74=GOOD; and .75–1.00=EXCELLENT. Note that levels of observed agreement (where available) ranged between 68.30% and 77.00% and the levels of chance-corrected agreement were all significant at or beyond the .05 level. Note also that the R_i value of .54 for reviews of the manuscripts submitted to the "American Psychologist" dropped to .38 on replication.

sion seems to be based on a statement made some years ago about one of the most prestigious journals in the physical sciences:

> We have found, for example, that in a sample of 172 papers evaluated by two referees for the *Physical Review* (in the period 1948–56), agreement was very high. In only five cases did the referees fully disagree, with one recommending acceptance and the other, rejection. For the rest, the recommended decision was the same, with two-thirds of these involving minor differences in the character of proposed revisions (Zuckerman & Merton 1971, p. 67).

Unfortunately, this brief analysis provides no answers to some very basic questions: (1) What type of rating system was used by the referees? (2) Given the high acceptance rates of *Physical Review*, how much agreement between reviewers would one expect on the basis of chance alone? (3) What is meant by "*minor* differences in the character of proposed revisions"? and (4) How representative a subset is this sample of all the manuscripts submitted at that time?

The question of representativeness seems the most important. Commenting recently on this issue, Hargens (1988) and Hargens and Herting (1990b, p. 17) note the following:

> One reason that studies of referee reliability are relatively rare for physical-science journals is that such journals often use the single initial referee system. Thus, data on pairs of referee assessments of all submissions are unavailable for these journals. Those manuscripts that do receive at least two independent referee evaluations under this system are an unrepresentative subset of all manuscripts. Thus, nonexperimental data on referee agreement for these journals, such as the evidence reported by Zuckerman and Merton, should be reviewed with caution.

Hargens is right in his conclusions, especially with respect to the structure of the journal *Physical Review* during the early study period (1948–56) from which the Zuckerman & Merton (1971) data were derived. From that time until 1969, the *Physical Review* did not allocate separate sections to physics specialty areas or subfields.

Beginning in 1970 however, and continuing to the present, the *Physical Review* allocated its total pages to four distinct subfields: general physics, condensed matter, nuclear physics, and particles and fields. Data, deriving from the *Physical Review* and *Physical Review Letters, Annual Report 1986*, indicate that although the overall acceptance rate of *Physical Review* for 1986[7] (75%) remained consistent with previous years (an average of 77% between 1969 and 1986), the percentage of manuscripts accepted in the four subfields varied rather widely. These data indicate that the acceptance rates were 81% for nuclear physics and 78% for condensed matter,

but only 70% and 69% for general physics and particles and fields, respectively. The nonparametric Jonckheere (1970) test of trend (Leach 1979) showed a highly significant trend, producing a Z value of 21.41 ($p < .00001$). This is interesting in its own right because it is consistent with the known higher manuscript-rejection rates for more general disciplines compared to more specific ones, the latter being thought of as "more experimentally and observationally oriented, with an emphasis on rigour of observation and analysis" (Zuckerman & Merton 1971, p. 77).

What further implications do such data have? It seemed plausible that even within the *Physical Review* journal, as the subfields become more and more general, there should be progressively less dependence on the deliberations of a single reviewer for any given manuscript. Would the pattern of acceptance rates across the four subfields *covary* with the tendency to rely on more than a single reviewer? The data in Table 3 indicate just that. Thus, the fit is quite remarkable, with the rank ordering between acceptance rates and the use of more than one reviewer proceeding about as one might predict, this despite the fact that the acceptance rates are based on 1986 data and the variation in number of reviewers per manuscript is based on 1987 data. The trend for variation among the decreasing percentage of manuscripts using a single reviewer, subfield by subfield, is also statistically significant (Jonckheere 1970, $Z = 6.87$, $p < .00001$).

Since manuscripts requiring more than one reviewer tend to be those that are problematic, these data indicate that even within the same physics journal the single initial referee system is not uniformly applied, but, rather, varies as a function of the subfield, with more general subfields having higher rejection rates and also requiring more reviewers before manuscripts are finally accepted for publication. We would predict that if the editors of *Physical Review* were willing to undertake a reviewer reliability study of manuscripts submitted in the four subfields, one would find appreciably higher levels of agreement for nuclear physics and condensed matter than for particles and fields and general physics. These recent findings are also of great theoretical importance, since they allow one's reasoning to come "full circle" to the conclusion that Merton's normative model is not even wholly appropriate for the physical sciences. Another way of putting this is that physics itself appears to share many of the same problems facing the general journals in both behavioral science and medicine.

There are other data deriving from physics that are consistent with those just presented. Qualitative statements made about manuscripts submitted to the *Physical Review Letters* also suggest that some of the problems about the applicability of Merton's (1973) normative model may not be unique to medical and behavioral science. According to the editors of *Physical Review Letters:* "The referees, representative of the readers, are severe judges of the papers. Only about 45% of the 2,300 papers submitted each year are accepted for publication" (Adair & Trigg 1979, p. 475).

The editors continue in their Statement of Policy for the journal:

> For the majority of the papers the comments of the two referees are sufficiently equivocal so that the editor cannot decide, with confidence, on the disposition of the paper. Further information is sought from the authors, from further communication with the original referees, from other referees, and/or from the Divisional Associate Editors. The editors initiate an average of five written communications per paper to referees, authors, and Associate Editors to gather the information which allows them to come to a conclusion concerning the disposition of the paper. Even then, for most papers, accepted or rejected, the evidence is not

Table 3. *The parallel relationship between acceptance rates for manuscripts submitted to "Physical Review" and the use of one or more reviewers*

A. 1986 Data (N=5264 Total Manuscripts [MS])

Subfield	No. MS Received	No. MS Accepted	% Accepted
C. nuclear physics	540	440	81%
B. condensed matter	2281	1786	78%
A. general physics	1325	931	70%
D. particles & fields	1118	775	69%
Across all Subfields	5264	3932	75%

B. 1987 Data (N=933 Accepted MS)

Subfield	No. MS With 1 Reviewer	No. MS With 2+ Reviewers	Total	% MS With 1 Reviewer
C. nuclear physics	79	12	91	87%
B. condensed matter	347	93	440	79%
A. general physics	168	53	221	76%
D. particles & fields	122	59	181	67%
Across all Subfields	716	217	933	77%

completely conclusive and the editors must judge as best they can the inconclusive evidence which bears on the subjective acceptance criteria (Adair & Trigg 1979, p. 476).

Consistent with Adair's assessment, Lazarus (1982, p. 219) notes that with respect to levels of interreviewer agreement for manuscripts submitted to the *Physical Review Letters*, "in only 10–15% of cases do two referees agree on acceptance or rejection the first time around – and this with the authors' and institutional identities known!"

Adair (1982) has expressed optimism that this situation will improve. Formal studies of the reliability of peer review for manuscripts submitted to physical science journals, especially in the more general areas, must be conducted, however, so that our conclusions can be based on more quantitative results than have been available thus far. Since the *Physical Review Letters* has been considered one of the two most prestigious publications in the field (Beyer 1978; Lodahl 1970), and, similar to the general journals in behavioral science and medicine, it does use the two-initial-referee system, a more quantitative assessment of peer-review practices should be of more than passing interest to an important segment of the scientific community. If such a study were undertaken, we would predict that levels of referee consensus for *Physical Review Letters* would be of the same relatively low order of magnitude (typically below R_1 of .40) characterizing general journals in many other disciplines.

The 1985–86 rejection rates of *Physical Review Letters* (consistent with the ordering of those for the *Physical Review*) are the highest for the general subfields of general physics (74%, or 631 manuscripts [MS] rejected/854 MS received) and cross disciplinary physics (68%, or 71/106); the rejection rate was lowest for the much more specific subfield, atoms and molecules (52%, or 243/470). Moreover, these data are consistent with journal rejection rates in psychology (Summary Report of Journal Operations 1988) in which general focus journals have the highest rejection rates, for example, the *Journal of Applied Psychology* (93%), *Psychological Review* (89%), and the *Journal of Experimental Psychology (JEP): General* (81%). At the same time, the more specific focus journals have the lowest rejection rates, for example, *JEP: Learning, Memory, and Cognition* (58%), the *Journal of Comparative Psychology* (39%), and *Behavioral Neuroscience* (also 39%). These data are also consistent with those reported by Lock (1985) for medical journals. Similarly, Hargens (1988, p. 139) notes that "cultural anthropology journals have higher rejection rates than physical anthropology journals, and rates for journals in social, abnormal, and educational psychology exceed those in experimental, comparative, and physiological psychology." During the early 1980s, the general focus (cultural) journal, *American Anthropology*, had a rejection rate of 85%, while the *American Journal of Physical Anthropology* evidenced a sharply contrasting rejection rate of only 22% (Hargens 1988, p. 150).

Our work and that of Hargens and Herting (1990b), support the argument that while manuscripts submitted to the journals studied in the behavioral and medical areas seem routinely to receive at least two independent reviews, this option is used in physics and related fields only when a manuscript seems problematic. In contrast to the experience of *Physical Review* and other physics journals (e.g., Abt 1988), fewer than 1 in 4 manuscripts (22% of 2274 manuscripts) submitted to the general *Journal of Abnormal Psychology* in 1973 received reviews based on the deliberations of a single referee. Moreover, the overwhelming majority of them (52/59 or 88%) were rejected.

Since the only comprehensive study of peer review of grant proposals was undertaken by Cole et al. (1981), this area is completely open to further research. Roy (1985) reminds us that there are five systems of grant review which are so different that criticisms aimed at one of them are not applicable to the others. For example, although all five systems use mail reviewers, they differ in terms of: (a) who selects the reviewers (i.e., program managers or peers unknown to the program managers); (b) the specific method of grant evaluation (average of referees' ratings, or the decision of an independent panel of peers); and (c) whether or not peer reviews are followed by a panel site visit. One interesting research question accordingly concerns how such differences might influence both the reliability and validity of grant reviews.

4.6. Reliability of grant reviews. The major source of data on the reliability of grant reviews is NSF grant submissions in three areas of study (chemical dynamics, economics, and solid state physics) as analyzed by Cole & Cole (1981, pp. 71–79). Three sets of reviews were considered: (1) NSF "open" (nonblind) reviews, (2) the Committee on Science and Public Policy of the National Academy of Sciences (COSPUP) "open" reviews, and (3) COSPUP "blind" reviews. Commenting on the interreferee reliability estimates from these data, Cole and Cole (1985, p. 38) wrote, "We have treated the reviewer variances as rough indicators of disagreement among reviewers."

In order to derive *direct* indicators of disagreement among reviewers, we first identified the problem of assessing grant-review reliability as a case of a more general problem in which: (1) there are two or more independent examiners per subject or object being evaluated; (2) both the number and actual examiners can vary from subject to subject (here, submitted manuscripts), and (3) the data derive from a continuous, dimensional, or quasi-dimensional scale of measurement. In their description of a computer program for analyzing such data, Cicchetti and Showalter (1988, pp. 717–18) noted that

> "an area of inquiry to which this design would apply is the assessment of the reliability of the peer review of grant applications. Thus, there may be two independent reviewers for grant A and four different independent reviewers for grant B. The range of possible ratings may be between, say, 10 (lowest score possible) and 50 (highest score possible), such as the evaluation schema used by referees in the peer review of National Science Foundation (NSF) grants (e.g., Cole & Cole 1981)."

As mentioned in section 3.3, the statistic of choice would be the intraclass correlation coefficient (R_1, Model III), based on the average number of reviews per grant proposal, as discussed in both Bartko & Carpenter (1976) and in Cicchetti & Showalter (1988).

These results are presented (for the first time) in Table 4 and once again indicate rather low levels of chance-corrected agreement. These range from .18 for COSPUP

Table 4. *NSF and COSPUP reviews: Summary of interreferee consensus levels*

Area of Study	No. of Proposals	No. of Reviews	Mean No. Reviews per proposal	R_i
A. NSF open reviews				
chemical dynamics	50	242	4.84	.25
economics	42	155	3.69	.37
solid state physics	50	192	3.84	.32
B. COSPUP open reviews				
chemical dynamics	50	213	4.26	.32
economics	49	181	3.69	.36
solid state physics	49	189	3.86	.34
C. COSPUP blind reviews				
chemical dynamics	50	212	4.24	.18
economics	49	198	4.04	.37
solid state physics	50	203	4.06	.33

Note: All R_i values are statistically significant at beyond the .005 level.

blind reviews of grants submitted in the area of chemical dynamics to .37 for NSF open and COSPUP blind reviews of grants submitted in the field of economics.

Similarly, the data on the reliability of peer review of AHA grants (the final calculated priority score) are expressed by Wiener et al. (1977, p. 309, Table 1) in terms of an intraclass correlation coefficient of .37 ($p < .001$).

4.7. Statistical meaning of low levels of reviewer agreement. The available data are clear. Quite low levels of chance-corrected interreviewer agreement are obtained in every area of scientific inquiry, from abstract, manuscript, and grant reviews. What does this mean from a biostatistical point of view? First, it must be understood that R_i (or kappa) statistics are omnibus indexes, meaning that they reflect only the overall level of chance-corrected agreement. "Overall" means reviewer agreement averaged over all possible rating categories (e.g., "accept," "resubmit," "reject" for manuscripts, or "high," "medium," "low" approval, or disapproval for grants). It has been shown (e.g., by Cicchetti 1985, in the context of journal peer review, and by Cicchetti 1988, more generally) that the overall level of agreement is nothing more than a weighted average of agreement on all possible rating categories (see also, Fleiss 1981; Spitzer & Fleiss 1974). It has also been demonstrated (again, Cicchetti 1985; 1988) that low levels of R_i or kappa can be produced not only by low levels of overall agreement, but also by large discrepancies in agreement on the various rating categories available to reviewers. We are referring specifically to wide discrepancies in reviewer agreement levels on approval (acceptance) categories as compared to rejection (disapproval) categories.

Some of the available literature on the reliability of peer review (e.g., Cicchetti 1985; Ingelfinger 1974) suggests, indirectly, that reviewer agreement on decisions to reject manuscripts is appreciably higher than agreement on acceptance. Is this true in general? Based on the available data, is there an analogue for the peer review of grant proposals?

To address these questions more specifically, one must develop rational criteria for dichotomizing reviewer agreement levels as "accept" or "reject." The analogue for grant reviews would be to dichotomize reviewer agreement between proposals receiving high ratings and those with low ratings. In the case of the *Journal of Abnormal Psychology*, 86% of those 203 manuscripts receiving ratings of either "accept" or "accept subject to revision" by both reviewers were accepted for publication. Analogously, of those 803 manuscripts receiving a rating of "resubmit" by both reviewers, or "reject" by one and "resubmit" by the other, or "reject" by both, 95% were rejected by the editor. This provides a rationale for combining "accept" or "accept subject to revision" into an "accept" category and "resubmit" and "reject" into a "reject" category. With respect to grant reviews, Cole et al. (1978) note that of those NSF applicants receiving evaluations of "very good" to "excellent" (40–50), 92% were *awarded* grants. Conversely, of those applicants with grades ranging from "poor" (10–19), "fair" (20–29), and "good" (30–39), 86% of them were *denied* grants. This provided a rationale for dichotomizing on the basis of peer-review scores of 40–50 (high probability of approval) and 10–39 (high probability of disapproval). The number of individual NSF and COSPUP reviews for any given grant varied between 1 and 8. Since the mean number of ratings for NSF open and COSPUP open reviews was quite similar (e.g., see Table 4), however, it seemed reasonable to use these more robust scores in our analyses.

The results based on these dichotomies are presented in Table 5 for manuscript reviews and in Table 6 for grant reviews (again, reported here for the first time). When one considers the manuscripts judged acceptable by one reviewer and then compares them to the corresponding set of manuscripts considered acceptable by a second reviewer, the agreement levels for "accept" vary between 44% and 66%. When the same set of analyses is performed on manuscripts classified in the "reject" category, however, the agreement levels vary between 70% and 78%. Direct comparisons between the proportion of reviewer agreement on accept versus reject recommendations produces chi square(d) values with corresponding p levels ranging between .10 and < .00001. As expected,

Table 5. *Relationships among chance-corrected reviewer agreement levels (R_i) and agreement levels for the acceptance and rejection of manuscripts submitted to major journals in behavioral and medical science*

Journal	R_i	Acceptance	Rejection	Combined	X_2	p
"Journal Abnormal Psychology"	.14	44% (462)	70% (857)	61% (1319)	83.99	.00001
Untitled Medical Specialty Journal	.26	50% (289)	76% (577)	67% (866)	57.895	.00001
"Developmental Review"	.27	52% (25)	74% (47)	67% (72)	3.413	.06
"American Psychologist"	.45	66% (62)	78% (97)	74% (159)	2.7315	.10

Note: R_i values are all statistically significant at beyond the .01 level.

Table 6. *Relationships among NSF and COSPUP chance-corrected agreement levels (R_i) and agreement levels on high=(40-50) and low=(10-39) rated grant proposals in 3 areas of research specialization*

Area of Specialization	R_i	Agreement on Proposals With: High Ratings	Low Ratings	All Proposals	X^2	p
Combined:	.32	54% (52)	76% (98)	68% (150)	5.69	.01
chemical dynamics	.16 (NS)	41% (17)	70% (33)	60% (50)	2.71	.10
solid state physics	.34	60% (20)	73% (30)	68% (50)	0.46	NS
economics	.44	60% (15)	83% (35)	76% (50)	1.88	NS

Note: The p values of R_i are at greater than the .10 level for chemical dynamics; at beyond the .01 level for solid state physics; and at beyond .001 for economics.

the lower chi square(d) values tend to be associated with those journals processing relatively small numbers of manuscripts. These results parallel both those reported earlier by Lock (1985) and those completed more recently for the *American Sociological Review*, *Physiological Zoology*, and *Law and Society Review* (Hargens & Herting 1990a).

The results in Table 6 for NSF and COSPUP grant reviews closely parallel those just reported for manuscript reviews. Specifically, reviewer agreement levels for proposals with low NSF and COSPUP ratings (10–39) were consistently higher (70%–83%) than agreement levels for those with high (40–50) ratings (41%–60% agreement). Though the numbers were too small to be statistically significant for a given specialty area, the combined Table 6 (row 1) data indicate significantly more interexaminer agreement on the 98 proposals with low ratings (76%) than on the 52 proposals with high ratings (54%).

Thus, on the basis of available data, it becomes clear for the first time that reviewers are indeed substantially more in agreement on which scientific documents to reject than on which to accept. Consistent with these data, it is noteworthy that editorial decisions for general journals (e.g., *Journal of Abnormal Psychology*) seem to give considerably more weight to referee consensus on rejection than to referee consensus on acceptance. Of the 203 manuscripts (of 1,316 submitted) for which independent reviewers *both* recommended *acceptance*, 28 (or 14%) were *rejected*. In comparison, only 28 (or 5%) of the 600 manuscripts that the reviewers agreed should be rejected were in fact accepted. Here, chi square(d), 1 df = 17.94 ($p < .00001$). In other words, the editor was almost three times more likely to reject a manuscript that reviewers agreed should be accepted than to accept a manuscript that the reviewers agreed should be rejected.[8] In attempting to interpret this important phenomenon further, however, one must consider a number of other issues.

5. Issues of Interpretation

5.1. Caveat #1: "One swallow does not a summer make." Although the findings are consistent across the types of scientific document analyzed (manuscripts, abstracts, grant proposals) and across areas of investigation (behavioral science: psychology, sociology; medicine: general and specialty areas; economics; and physical sciences: chemical dynamics, solid state physics), one must keep in mind that the documents investigated do not represent a broad cross-section of existing materials. Rather, the studies of peer review have been focused on a relatively small number of prestigious journals, professional organizations, and grant-reviewing agencies. More research is clearly needed to test the generality of the findings to date. For example, as correctly noted by Lock (1985), the journals investigated tend to be general ones that all share very high rejection rates. With such a focus in mind, reviewers may be more interested in determining what is wrong with a particular submission than in documenting some of its more positive attributes. It has been conjectured by workers in the field that journals that have much higher acceptance rates, such as *Physical Review* (between 73% and 81%, between 1969 and 1986) may display the reverse phenomenon, or "when in doubt, accept" (Zuckerman & Merton 1971, p. 78).

With respect to NSF grants, Cole and colleagues reported that at the time their study was undertaken, approximately one out of every two applicants was eventually funded. Given the current relative scarcity of NSF (and other) funds, what impact will this have on

future performance of NSF reviewing? Our data suggest that the difference in agreement levels on proposals with low and high ratings might even intensify. In contrast to this, Wiener et al. (1977) provided some suggestive data (variances and standard deviations) to show that reviewer agreement levels were highest on AHA grants receiving top grades, worst for grants receiving the lowest grades, and intermediate for grants receiving intermediate grades. An obvious question would concern what percentage of AHA grants were in fact funded during this 1974–75 period of grant submission and review. Again, further research and additional analyses of existing data on the peer review of grant proposals are urgently needed to help clarify these important issues.

5.2. Caveat #2: Field studies of peer review lack necessary controls for proper interpretation. Because varying sets or numbers of reviewers examine different manuscripts or grant proposals, it is never possible in what we have called *naturalistic* research designs (sect. 3.1) to determine how much of the unreliability results from differences in the characteristics of the reviewers themselves: for example, level of experience; harshness or leniency as critics; the quality or technical difficulty of the manuscripts, abstracts, or proposals under review; blindness or openness of reviews; or some attribute that may be masked in the reasons the reviewer offers for recommending rejection. Such attributes include the following: theoretical biases; biases against statistically nonsignificant results; and the prestige of the author or institution. To make matters even more complicated, the unreliability of peer review may in fact involve some complex interaction among some or all of these or still other uncontrolled variables.

6. Clarifying issues of interpretation

6.1. Quasi-experimental and experimental studies of peer review. To study directly the influence of prestige of the author's affiliation on the reliability of peer review, Peters and Ceci (1982) resubmitted 12 articles that had already been published in prestigious psychology journals (between one and one-half and three years earlier) by authors from highly regarded and well-published American psychology departments. The authors' names and affiliations were fictionalized, the latter being made much less prestigious (e.g., "Tri-Valley Center for Human Potential"). Only 3 of the 12 resubmissions were recognized as having been published previously. All but 2 of the 18 referees and editors recommended rejection of the resubmitted publications.

One weakness of this study was the authors' contention that the findings provided evidence of reviewer bias in favor of high-status authors or high-status affiliations. A plausible alternative explanation has been offered by critics, namely, that the results provide evidence of reviewer bias against low-status authors and/or institutions. As Peters and Ceci appropriately respond however, "While we do not know for certain, which of the two forms of bias is more likely, neither is desirable." (Peters & Ceci 1982, p. 247). Consistent with Peters & Ceci's findings, a large number of authors using research designs other than quasi-experimental ones have reported a relationship between author affiliation and the likelihood of publication in major journals (e.g., see Berelson 1960; Beyer 1978; Cleary & Edwards 1960; Crane 1967; Goodrich 1945; Kraus 1950; Pfeffer et al. 1977; Yotopoulos 1961).

A second criticism of the Peters & Ceci study is that it lacked an appropriate control group consisting of previously rejected manuscripts resubmitted for further review. Smigel and Ross (1970) tested just that: They resubmitted an "accidental" sample of eight rejected manuscripts that had remained in their editorial files to a new set of reviewers under a new editor of *Social Problems*. Of these, seven were rejected by both editorial referees and one was conditionally accepted by one referee with no opinion given by the second. Whatever interpretation one chooses to make of these findings (since neither study included proper controls), the results are consistent with the data presented in Tables 5 and 6, namely, that reviewers have much less difficulty in agreeing on rejection than on acceptance.

In one of the best controlled studies of peer review (89% response rate, random assignment to experimental conditions) Mahoney (1977) invited 75 guest reviewers of the *Journal of Applied Behavior Analysis* to review manuscripts that all tested the same dominant behavior modification hypothesis. The manuscripts had identical Introduction and Methodology sections, but varied systematically in whether the Results and Discussion sections were (i) not provided at all, or the findings were described as either (ii) " positive," (iii) "negative," or (iv) "mixed."

Referees were asked to judge the manuscript on the basis of overall scientific merit (publishability) and to apply normative criteria, including ratings of topical relevance, methodology, and data presentation. The referees of the manuscripts reporting positive results usually recommended acceptance with moderate revisions. The referees who received papers showing mixed results consistently opted for rejection. Those who read manuscripts giving negative results typically recommended rejection or major revisions. Referees evaluating manuscripts that reported no results at all gave more positive recommendations than those whose manuscripts had a Results section.

For both the positive and the negative manuscripts there was an R of .94 between ratings of perceived adequacy of "methodology" and potential publishability; there was a corresponding R of .56 between the perceived adequacy of "data presentation" and publishability.

In another set of analyses, marked discrepancies were found between what referees predicted as their expected levels of interrater reliability on the various evaluative criteria and what turned out to be their actual levels of interrater reliability: The *predicted* reliability (R_i) levels for the criteria (e.g., adequacy of methodology, extent of overall scientific contribution) varied within a narrow range of .69 to .74. The *actual* levels of R_i ranged between −.07 (below chance expectancy) and +.30. In fact, Mahoney's finding of an R_i of only .03 between referee ratings of methodologic adequacy, coupled with an R of .94 between perceived adequacy of the methodology and publishability is entirely consistent with the findings of two naturalistic studies discussed earlier (Cicchetti & Eron 1979; Scott 1974) and also with the results of two

experimental studies (Abramowitz et al. 1975, and Cicchetti & Conn 1976).

The bias against manuscripts reporting negative findings is consistent with the earlier work of Bozarth and Roberts (1972); Hunt (1975); Kerr et al. (1977); Reid et al. (1981); Rowney and Zenisek (1980); Smart (1964); and Sterling (1959). The related issue of bias against replication studies is still being debated in more recent literature (e.g., Bernstein 1984; Casrud 1984; Furchtgott 1984; Garber 1984; Heskin 1984; Sommer & Sommer 1984). With few exceptions (e.g., Rourke & Costa 1979), the apparent bias against replication studies is very strong (on the part of both reviewers and editors). With respect to the testing of major theories or hypotheses in a given field of scientific inquiry, one would be most concerned about the literature being glutted with Type I errors, that is, rejecting the null hypothesis (that there are no statistically significant differences) when the hypothesis is true (e.g., see Greenwald 1975; and most recently, Soper et al. 1988). A successful strategy has been simply to build the replication study into the first part of the research design, followed by the main study. Although referees and editors, in our experience, seem willing to accept replication studies embedded in an overall research design, they are quite unwilling to accept them alone. (For recent empirical data underscoring the vital need for replications in the examination of dominant theories or hypotheses, see again, Soper et al. 1988.)

Finally, in a qualitative evaluation of reviewers' comments, Mahoney noted the wide variability in responses. When examining the comments in isolation, he noted, "one would hardly think that very similar or even identical manuscripts were being evaluated" (Mahoney 1977, p. 171).

In conclusion, the results of Mahoney's experiment indicate a strong reviewer bias against both negative and mixed results, with an opposite bias in favor of manuscripts reporting positive results. Mahoney describes this phenomenon as confirmatory bias or the tendency to evaluate positively those results that are consistent with one's own beliefs and to evaluate negatively those that are inconsistent with them. (See also Beck 1976; Goodstein & Brazis 1970; and, most recently, Greenwald et al., 1986, for a critical discussion of the broader corpus of literature in which confirmatory bias and other theoretical biases are seen as obstructing scientific progress.)

In a second experimental study by Mahoney et al. (1978), 68 volunteer referees for two behavioristic psychology journals were sent experimental manuscripts that were identical in content, except that half the referees were randomly assigned manuscripts in which the alleged authors supported their arguments by citing their "in press" publications. The remaining referees received manuscripts in which "self-citation" was not used by the fictitious author. In addition, half the manuscripts in each group were given a prestigious author affiliation, while the remainder were described as having come from a "relatively unknown college." Referees were again asked to rate the manuscript using various evaluative criteria and to provide a summary recommendation concerning the article's publishability potential ("accept," "accept with minor revisions," "accept with major revisions," or "reject"). Statistically significant results ($p < .05$) indicated that articles in which the fictitious author provided self-citations were rated as more innovative and publishable than those in which no self-references were cited; institutional prestige, whether high or low, bore no significant relationship to either the reviewers' evaluation of the manuscript's normative attributes or to the reviewers' summary recommendations. Mahoney and colleagues note what may have been an unintended flaw in the design of the study however, namely, "the fact that none of the four institutions was known to specialize in behavioristic psychology so that – from the reviewer's perspective – there may have been little perceived variation in 'relevant' prestige" (Mahoney et al. 1978, p. 70). Despite this possible shortcoming, Mahoney's experimental research on peer review can still be appropriately described by the double entendre "rare," but "well done."

How do the Mahoney studies help us understand the low levels of reviewer agreement in the evaluation of scientific merit? Earlier (sect. 5.2), we noted that the low levels of reviewer agreement were difficult to interpret because we could not determine how much of the unreliability was due to differences in such important variables as the reviewers themselves (e.g., harsh vs. lenient critic), the manuscripts rated (e.g., some manuscripts were technically or otherwise more difficult to review than others), or the availability of author identity and affiliations (some journals use blind reviews, others do not). Because such variables were controlled in the Mahoney experiments, the low levels of reliability that were reported earlier are easier to accept now as probably nonartifactual.

In summary, on the basis of the best controlled studies of the peer-review process to date, we are forced to conclude that referees do at times apply subjective criteria, which cannot be described as "fair," "careful," "tactful," or "constructive," despite the fact that such traits are widely accepted as desirable characteristics of referees (e.g., Gordon 1977; Hall 1979; Jones 1974; Lindsey 1978; Merton 1973). The clearest instance of this phenomenon was that manuscripts were likely to be accepted or rejected on the basis of whether the findings were positive, negative or mixed, rather than on the basis of their worthiness. Such subjective considerations, when they affect one reviewer, or both, may have a negative influence on both the reliability and validity of the peer-review process. Somewhat paradoxically, the consistent application of the same biased criterion (say, a preference for positive findings) to a given set of manuscripts would inflate the reliability of the peer-review process, while potentially compromising its validity (i.e., falsely assuming that positive results are always more worthy of publication than negative ones).

6.2. Further reasons for the low reliability of peer reviews. As we have seen, the list of subjective criteria detected by the better controlled manuscript-review studies includes the extent of "confirmatory bias," "self-citation" bias, and "prestige of author and affiliation" bias. Although many will argue that better research emanates from more prestigious institutions, the categorical acceptance of such research, coupled with a summary rejection of research produced at less prestigious institutions, will build an inevitable bias into the peer-review process.

Although comparable quasi-experimental or experi-

mental studies of peer review of grant proposals do not appear to have been undertaken, there are some less direct data that bear on the subject. Mittroff and Chubin (1979) describe a report by Hensler (1976) that notes that both NSF reviewers and applicants feel that, all things being equal, applicants have a better chance of being funded if they are affiliated with a better known institution, are well established and well known, or are submitting a "mainstream" rather than a more innovative proposal. In a more comprehensive survey, however, Cole and Cole (1981) report little effect if any on NSF funding associated with the following: previous publication record, institutional affiliation, or the applicant's age. The lack of a substantial relation between track record and the probability of being funded is described by Cole and Cole (1981, p. 2) as "surprising, since one of the stated evaluation criteria used by the NSF in evaluating proposals is the ability of the scientists to conduct the research proposed." What Cole and Cole find to be the major determining factor in whether or not a given NSF grant is funded is the score (perceived merit) given to the grant by the reviewers. In commenting negatively on this phenomenon, one peer-review expert describes an alternative system of peer review "that involves not a promise in an essay (i.e., proposal), but uses a track record of performance in research" (Roy 1985, p. 73; see also, Chubin, 1982, in support of this general strategy). Other factors contribute to the unreliability of the peer-review process in a much more subtle or enigmatic manner (e.g., Cicchetti 1982; Smigel & Ross 1970).

6.3. "Enigmatic" issues and their influence on the reliability of peer review. In examining the content of referee comments and their relation to specific recommendations to the editor, Smigel and Ross (1970) identified two types of problem cases. In one, the referees agreed on either acceptance, resubmission, or rejection, but for entirely different and sometimes even conflicting reasons. If the editor were to focus solely on final reviewer recommendations (i.e., ignore the content of the reviews), then the conclusion to accept, require revision and resubmission, or reject would at times be based on illusory reliability.

The reverse phenomenon, an even more subtle one, occurs when referees are basically in agreement about the content of their reviews, but differ considerably in their recommendations to the editor. Specifically, one referee may opt for acceptance because he believes his criticisms are minor ones. The second referee, citing the same criticisms, feels they are major, and hence opts for rejection. On which referee does the editor rely? Understandably, no one has yet been able to resolve such difficult problems. As a result, we are left with the apparent paradox of instances in which conscientious and well-qualified reviewers and editors will offer essentially the same evaluation of a given manuscript, while drawing very different conclusions about its publishability.

Evidence suggests that this same phenomenon faces program directors in the peer review of grant proposals. One NSF program director noted that some of his reviewers never rate a grant proposal as "excellent," no matter how meritorious they perceive it to be. Directors learn not to "downgrade" an applicant on this basis, since one reviewer's rating of excellent for a given proposal may have the same meaning as another reviewer's "very good" (i.e., see Cole & Cole 1981).

7. Improving the reliability of peer review

7.1. Rationale. Somewhat paradoxically, disagreement among reviewers can sometimes serve a useful purpose. Thus, one referee may detect a flaw in reasoning that a second referee has failed to uncover (e.g., Bailar & Patterson, 1985, in the context of journal peer reviews; Cole & Cole, 1981, in the context of NSF peer reviews; Harnad, 1979; 1983, in the context of "creative" disagreement in open peer commentary). But whereas a valid case can be made for the potential informativeness of this kind of reviewer "unreliability," it is not really inconsistent with a concurrent desire to strengthen both the reliability and the validity of the peer-review process, as espoused, for example, by Harnad (1985).

Yet, even adopting this desideratum, Mahoney (1977; 1985) warns that one should not seek to improve reliability in peer review at the enormous expense of increasing the extent of referee bias or prejudice. Thus, training referees to agree by simply sharing the same biases or prejudices against various types of scientific documents would be quite "counterprogressive" (Mahoney 1985, p. 2). We would strongly agree. How to deal with this important issue then?

7.2. The role of multiple reviewers. To improve the reliability of peer review, a minimum of three independent referees has been recommended (e.g., Glenn 1976; Newman 1966). The procedure is already used by *Behavioral and Brain Sciences* (*BBS*), which sends a given manuscript to anywhere from five to eight reviewers (sometimes even more) explicitly chosen to represent the manuscript's specialty, as well as other specialties on which it impinges, and to include investigators likely to be favorable, critical, and neutral. Moreover, *BBS*'s decision to accept or reject hardly amounts to a "majority vote," referees' recommendations being weighted by their backgrounds, alignments and, above all, their *reasons* (Harnad 1983; 1985).

There are several arguments for consulting more than two referees: (1) The number of manuscripts that receive split reviews (therefore usually requiring a third review anyway) can be quite substantial: about 25% of manuscript submissions to the *Journal of Abnormal Psychology* over a six-year period (Cicchetti & Eron 1979 and additional unpublished data). (2) Existing pools of referees are large enough to make this option viable for behavioral science, medicine, and the physical sciences (e.g., see Lindsey 1978, p. 107). (3) Concerning issues of validity, the likelihood that an important feature of an article (or grant proposal, e.g., detection of a fatal design flaw) will be missed decreases as the number of independent reviews increases. (4) Consistent with argument (3), it is a well-known statistical fact that the reliability of ratings does increase as the number of raters is increased (Hargens & Herting 1990b; Nunnally 1978).

7.3. Using author anonymity or blind review. The main argument in favor of blind review for journal submissions

is the contention of some authors that their manuscripts seem to be rejected more on the basis of reviewers' subjective criteria (such as prestige of the author's affiliation) than on the basis of overall scientific merit (e.g., see Armstrong 1982b; Benwell 1979; Ceci & Peters 1984; Gordon 1977; Patterson 1969). Opposing arguments have been advanced (e.g., by Ingelfinger 1974). More recent criticisms of "blinding" manuscripts have been summarized by Ceci and Peters (1984, p. 1492): (1) an expensive publicity stunt used to placate authors but with little effect on quality, fairness, or interreferee reliability levels (Thomas 1982); (2) a process making it possible for authors to exaggerate their publication record, presumably by referring to their supposed research without having to cite author(s), journals, and publication dates, as proof of its existence (Howe 1982; Over 1982); (3) a mechanism enabling authors to leave out crucial information required for successful replication of their work (Lazarus 1982); and (4) a process that restricts the development of a constructive relationship between authors and editors (*Eight APA journals* 1972). Bradley (1981) also reports the results of a poll of psychologists revealing that more than 75% of them believed that the usual way authors' names and affiliations are removed from submitted manuscripts does not prevent reviewers from identifying the authors of such articles. (One consistent example of the failure of blinding occurs when names and affiliations are removed on the face sheet, but a footnote identifying the senior author and the institution at which the research was conducted is not.)

The Ceci & Peters (1984) review of the literature found no sound empirical evidence for the futility of blind review. Rather, the negative beliefs seemed to rest on the anecdotal experiences of selected authors (e.g., Machol 1981). Ceci and Peters accordingly tested hypotheses about the feasibility of blind review. They randomly selected 180 reviewers for 6 psychology journals (each covering a different area); 81% agreed to participate and 73% returned usable questionnaires. The journals were: *Journal of Personality and Social Psychology, Journal of Counseling Psychology, Human Learning, Developmental Psychology, Psychological Bulletin,* and *Psychometrika.*

Although the reviewers had predicted that they could correctly identify authors of manuscripts in 72% of the cases, their actual "hit rate" was only half of that (36%). Moreover, these results were not significantly affected by either the reviewers' age or the specific journal that was represented. The authors concluded:

> At a time when the integrity of the peer-review process is under siege, blind review would seem to be an obvious step toward regaining authors' confidence in the editorial process. If our findings from these six journals can be generalized to the 60 or so journals in the field (out of approximately 120) that routinely use blind review, including half of those published by the APA, then we have evidence that the personal identities and institutional affiliations of authors usually do not contaminate the evaluations of reviewers who are kept blind (Ceci & Peters 1984, p. 1494).

Although the results of Ceci and Peters are impressive, one must first ask whether the peer-review glass is to be perceived as 64% full or 36% empty. Moreover, further research is needed to determine: (a) whether more specialized fields of inquiry would produce different results, because of the smaller numbers of scientists working on similar problems; and (b) whether blinding raises or lowers the reliability or validity of the review.

Nonetheless, the importance of these findings should not be ignored. Perhaps a compromise would be optional blind reviewing, already the policy of some editors (e.g., see Adair 1981, p. 14). It would probably make sense to leave the responsibility of blinding a given manuscript to the author who makes the request, however. This strategy would be designed (a) to increase the probability of successful blinding (e.g., eliminating mechanical detection errors), since the author who made the request would presumably have a vested interest in maintaining anonymity; and (b) to free valuable time for editors and their staffs. *Optional* anonymity, however, might stigmatize some authors (e.g., does the author have something to hide?).[9]

With respect to NSF grant reviews, Cole and Cole (1981) note that initial attempts to blind such proposals compromised the integrity of the proposal in a number of instances, to the point that the "substantive content became very unclear. Moreover, since there was substantial disagreement about what was an identifying sentence, remark, or allusion, severe blinding depended heavily upon the blinder" (op. cit., p. 11).

Cole and Cole (1981, p. 12) adopted a compromise blinding procedure similar to the one used by many scientific journals. From each grant they removed the title pages, relevant author references, descriptions of research facilities, direct references to prior work of the principal investigators, and other obvious identifying information. (The data presented in Table 4 indicate no obvious differences between COSPUP "blind" and "open" reviews of the same proposals whether in chemical dynamics, solid state physics, or economics.)

7.4. Revealing reviewer identity. The call for journal editors to force referees to reveal their identity has sometimes been strident and cuts across fields of scientific inquiry (e.g., the behavioral sciences, Patterson 1969; Surwillo 1986; the medical sciences, DeBakey & DeBakey 1976; Ingelfinger 1975; Margulis 1977; Stumpf 1980; and Wright 1970; and the physical sciences, Robertson 1976). One author asks: "Why should the wish to publish a scientific paper expose one to an assassin more completely protected than members of the infamous society, the Mafia?" (Wright 1970, p. 404).

Despite the legitimate concern about possible unfairness in anonymous reviews, one needs to ask what effect open review might have on the younger reviewer who legitimately criticizes the work of an established titan in the field. What protection does such a reviewer have against possible retribution? In the reverse situation, the well-established critic would be less likely to suffer retribution from his less well known critic (e.g., see Cicchetti 1982; Scarr 1982).

Consistent with the view of Armstrong (1982b), a compromise solution appears appropriate. Referees ought to be encouraged to reveal their identities, but only if they so choose. Moreover, two levels of anonymity should be available as alternatives: anonymity from the

author, but not from fellow reviewers, and complete anonymity from everyone but the editor (this is *BBS*'s current policy). Ideas similar to these would also pertain to reviewers of grant proposals.

7.5. Author review of referees. To counteract possible unfairness or incompetence in refereeing, Hall (1979, p. 798) has suggested "author review." Each author of a submitted manuscript (or grant proposal) would be given the opportunity to evaluate referees on the basis of "fairness," "carefulness," "constructiveness," or whatever other factors the author (or grant applicant) deems relevant. Such information would then be recorded, filed, and periodically reviewed by the editor (or, for that matter, the program director, or executive secretary of a granting agency). Those referees receiving repeated low ratings might then be eliminated as future peer reviewers. The process of author review is practiced by *BBS*.

Another innovative strategy, also aimed at avoiding potential referee bias (here, Mahoney's confirmatory bias) has been used by a number of journals, including the *International Journal of Forecasting* (Armstrong 1982b). A referee is given a note containing information about an author's research design, methodology, and data analysis but *not* about the results or ensuing discussion. The referee is instructed to review the paper using the note, after which he can open a sealed envelope containing the completed manuscript. Armstrong (1982c) also recommends that authors be permitted to submit names of potential reviewers as well as those who, with reasons, should not review. All this must obviously be done in a nonbinding manner.

7.6. Rewarding referee contributions. Whereas much of the literature on manuscript peer review has tended to be highly critical of the performance record of referees, the opposite sentiment has also been expressed, especially for those referees who consistently write "commendable" reviews (e.g., Armstrong 1982b; Hunt 1971; Smith 1977). Accordingly, ways of acknowledging commendable referee performance have been suggested (e.g., letters of thanks, with copies to members of the editorial board of the journal; publication of lists of especially good referees; acknowledgments in footnotes for substantial referee suggestions; or even an offer to publish such reviews, see Armstrong 1982b; Hunt 1971). Several journals have adopted this or similar strategies. For example, *BBS* invites reviewers to serve as commentators if the article is accepted for publication.

7.7. Allowing authors multiple manuscript submissions. The injunction against authors submitting the same article to more than one journal is consistent across major journals in behavioral science as well as medicine (e.g., see the 1983 American Psychological Association (APA) Publication Manual for psychology journals; the American Sociological Association (ASA), policy for sociology journals, Peters 1976; and the policy for major medical journals, Relman 1978). Peters (1976) appears convinced that a policy of multiple submissions would generate healthy competition among journals, all vying for the same manuscript. He presents some cogent arguments against the rationale for the ASA policy statement, but there are a number of attendant problems to be considered.

Referees often review for a number of different journals in the same general area of inquiry. Thus, when an author submits a manuscript simultaneously to several journals, a given referee may receive requests to review the same article from several different journal editors. What is the conscientious referee to do under these circumstances: Send the same review to each journal? Select one journal only? If so, which one? As another example, what happens when more than one journal accepts an author's paper? How can an editor intelligently organize an upcoming issue when faced with a sudden and unexpected withdrawal of a previously scheduled manuscript by an author receiving a second acceptance from a preferred competing journal (e.g., see Hughes 1976)? Also, what justifies the time and expense of needless multiple refereeing?

More important, Lindsey (1978) predicts that if the model of multiple submissions "is adopted by journals, authors will not be guaranteed a careful and impartial review of their work. Rather they will be vying for the attention of editors to their work. In this competition, those with prestigious credentials will receive the closest attention. Rather than equalizing access, there is the danger that multiple submissions may have just the reverse consequence" (Lindsey 1978, pp. 110–11). Given these rather serious problems, none of which appears to have yet received satisfactory solutions, multiple submission does not seem to be a viable procedure.

7.8. Developing an author-to-editor appeal process. In earlier sections (6.1, 6.2), we discussed a number of variables that have been cited in the literature that: (a) could be objectively identified; (b) were irrelevant to the publishability of a given manuscript; but (c) have nevertheless been used at times by editors to justify rejecting a given journal submission. There are eight such variables:

1. documentable factual errors made by reviewers (e.g., suggesting an invalid procedure in lieu of a valid state-of-the-art one developed by the author)
2. faulty data analyses (which are readily reparable) in a work that has otherwise received a quite praiseworthy reviewer and/or editorial evaluation
3. prior publication in nonrefereed conference proceedings
4. replication of previously published work
5. null or negative findings (despite very favorable reviews)

or (6–8) without supporting arguments:

6. subject inappropriate
7. ideas insufficiently novel or original
8. insufficient space to accommodate the submission.

With respect to the first five variables, it would seem that a carefully documented, dispassionate letter could result in a journal editor's honoring a request that the rejected manuscript be resubmitted to a new set of referees who would review it independently and without prior knowledge of its rejected status. Should this strategy fail in the more general areas of behavioral science or medicine, then the author can choose, in all likelihood, equally prestigious alternate journals with (as we shall later report) a reasonably high probability of acceptance.

As we shall also see, the problem in the physical sciences is a quite different one.

If the author is told that the subject is inappropriate for the journal to which it has been submitted, a search should be made for a more appropriate alternate journal. The last two reasons for manuscript rejection (i.e, lack of originality or lack of journal space) seem somewhat subjective. A letter to the editor would accordingly seem pointless, and again, alternate journals should be considered.

7.9. Developing a peer-review appeals systems for grant submissions. Both the National Institutes of Health (NIH) and the Alcohol, Drug Abuse, and Mental Health Administration (ADAMHA) granting agencies have recently established formal peer-review appeal systems. According to Holt (1985), classes of problems that might justify an appeal include: the granting agency's refusal to accept an application; an applicant's disagreement over the assignment of his grant (whether to a specific study section or Institute); an author's doubts about the level of knowledge of specific study section members; or evidence of bias associated with the peer review. In a survey reported by Mitroff and Chubin (1979 p. 222), more than 70% of NSF applicants favor what is referred to as a "formal" appeal system as a remedy for mistakes. A similar process is in operation at the NSF, whereby a principal investigator (P.I.) whose grant has been disapproved can call the designated program officer, who will provide a detailed account of the specific reasons for disapproval. If the P.I. is still dissatisfied, other options are available, including a request that the Assistant Director of NSF reconsider the grant disapproval on the basis of arguments set forth by the P.I. These developments portend a healthier climate for handling legitimate claims about perceived unfairness or incompetence; they become especially important as less and less funding is made available to support worthy research endeavors.[10]

8. Concluding comments

Given that rejections of manuscripts and disapprovals of grant proposals can seriously affect one's research career (e.g., jeopardize or delay a potential promotion), what conclusions can be drawn from this report?

As Harnad (1986, p. 24) notes, this depends to a large extent on what one views as "closer to the truth": (a) that most published research is either "significant and essential to the progress of science"; or (b) that most of it is "neither significant nor essential." Harnad adds that there is some evidence to support both propositions. If one believes (a), then one becomes very concerned that at least some meritorious research will be rejected, disapproved, or delayed unnecessarily, thereby hampering scientific progress. On the other hand, if one believes proposition (b), then the potential problem of false negatives (rejecting meritorious scientific documents) becomes less significant. For if most research is unimportant and leads nowhere, then it matters less if some of it is rejected or delayed; indeed, this would even come as a welcome relief to many who, like Lock (1985), are deeply concerned about the literature glut.

Our own view tends toward proposition (a), especially with respect to the peer review of grants. It has recently been noted that although more and better scientists are being produced each year, the funds available to support research are not keeping pace (e.g., Ison 1985; Koshland 1985). In a debate reported in a *Science* editorial by Culliton (1984), NIH director James B. Wyngaarden addressed the problem he refers to as distinguishing "'shades of excellence' among competing grants that are all at the top." He went on to state that:

> in many institutes, there is money enough to fund those grants with top priority scores of 160 to 170, while those rated only slightly lower at 171 to 180 end up in the reject pile. Nearly everyone agrees that there is no objective way the peer-review system can make such fine-tuned distinctions about quality. (p. 1401)

Consistent with this statement, the current editor of *Science* notes that "with quality as high as it is today, and funding low, a committee of Solomons would have difficulty distinguishing between grants that should and should not be awarded" (Koshland 1985, p. 1387).[11] Given the decreasing availability of research funds, the current situation has become markedly more competitive, with NIH grants usually requiring priority scores of 125 or considerably less, in order to assure funding.

The inability of referees to make "fine-tuned" distinctions also affects the peer review of manuscripts submitted to prestigious scientific journals (e.g., the *New England Journal of Medicine (NEJM)*, the *Journal of Clinical Investigation (JCI)* and *Science*). Wilson (1978) reported that of all manuscripts rejected by *JCI* in 1970, 85% of them were later published elsewhere: "The journals in which these papers were published constitute a distinguished list of publications, with 14 journals accounting for one-half the papers." Similarly, the editor of the *NEJM* noted that 85% of his journal's rejections in 1975 were subsequently published or accepted for publication elsewhere (Relman 1978). Approximately 70% of these initially rejected manuscripts appeared in very distinguished medical specialty journals, general medical research journals (i.e., *Journal of Clinical Investigation (JCI)*, *Journal of Laboratory and Clinical Medicine*, *Journal of Applied Physiology)*, or general medical journals (i.e., *Journal of the American Medical Association (JAMA)*, *Lancet*, *British Medical Journal (BMJ)*, *Canadian Medical Association Journal*, *Medical Journal of Australia*). The remaining 24 articles were published in local, state, and other types of journals. It should also be noted further that of the 85% initially rejected articles, both for *NEJM* and *JCI*, the majority were either not changed or changed in only minor ways in their ultimately accepted versions. Similar results were published by Lock (1985, p. 67): 79% (or 1223) of the manuscripts submitted to the *BMJ* in 1979 were rejected; 68% of these "were published elsewhere, 15% in high-impact-factor journals, and 10% in high-impact-factor general journals" (based on the average number of citations of the published articles).

The results of these three studies are consistent with those of a MEDLINE (computer) search reported by the former editor of *Science*. Abelson (1980) noted that although *Science* rejects about 80% of submissions to the journal, "almost all of our rejected material has appeared in other journals" (p. 62). Finally, Garvey et al. (1979) have reported that the bulk of manuscripts rejected by social science journals are also subsequently published,

often without further revisions, and usually in journals as prestigious as the rejecting journals to which they were originally submitted.

The situation seems to be very different in the major astrophysics and astronomy journals. Only about one-third of the rejected manuscripts are subsequently published in other journals (Abt 1988). As noted recently by Hargens (1990), this phenomenon is consistent with the notion that unlike social and medical scientists, astrophysicists and astronomers are more likely to conclude that their rejected work does not merit being published elsewhere. This may in turn reflect more agreement on evaluation standards in well-defined areas of the physical sciences than in either less well defined areas of the same discipline (e.g., general physics, cross disciplinary physics) or in the more general areas of medicine or behavioral science, that have been investigated to date.

In considering the implications of such findings for peer review in general, one should probably be less concerned with the high rejection rates of general journals in social and behavioral science and medicine than with the overall increased rejection rates for grants submitted in all three areas. First, despite previous arguments to the contrary (e.g., Cole 1978; 1983), rejection rates for manuscripts both between journals and within the same journal have remained remarkably constant over time. Hargens (1988); and Hargens (1990) reports that the rejection rates for the late 1960s and the early 1980s for 30 leading U.S. journals in a wide range of disciplines were very highly correlated (Pearson R = .94). Moreover, these rejection rates were not significantly associated with either changes in journal submission rates between the two time periods or with whether journals levied page charges. Our analysis of the rejection rates in the 10 subfields covered by the *Physical Review Letters* shows that between 1981 and 1986, all possible rank-orderings of rejection rates (i.e., comparing each year's rankings with those of each remaining year) vary between .91 and .99. Considering both the relative ease with which authors in many areas of social and medical science succeed in publishing their previously rejected articles in other prestigious journals and the extent to which authors in the physical sciences choose not to do so (tending to regard their rejections as decisive), the focus of concern should be on the problems associated with the rather arbitrary rejection of grant submissions, in which the phenomenon cuts across various disciplines (e.g., physics, chemistry, economics) and may prevent or seriously delay the implementation of worthy research endeavors.[12]

ACKNOWLEDGMENTS

The authors gratefully acknowledge the extensive computer programming and data-analytic contributions of Robert Heavens, James Owen, and Lorraine Gambino. This research was supported by a Veterans Administration Merit Review Grant, MRIS 1416 (Dr. Cicchetti).

NOTES

*Senior Research Psychologist, Biostatistician, and Senior Research Scientist, West Haven VAMC and Yale University, 350 Campbell Ave., West Haven, CT 06516.

1. For a more detailed description of normative attributes and specific criteria for guiding referees and editors in the review of scientific manuscripts, see Bowen et al. (1972); Chase (1970); Cicchetti & Conn (1976); Cicchetti & Eron (1979); Gottfredson (1978); Greenwald (1976); Maher (1978); Scott (1974); Whitehurst (1983); and Wolff (1973). For corresponding information pertaining to grant reviews, see Allen (1960); Cole & Cole (1981; 1985); Cole et al. (1978); Mitroff & Chubin (1979); Noble (1974); and Wiener et al. (1977).

2. The general formula for the kappa or weighted kappa statistic is:

$$\text{Kappa}_{(\kappa)} = (PO - PC)/(1 - PC), \text{ in which:}$$

PO refers to the proportion of observed (or actual) rater (reviewer) agreement; PC refers to the proportion of agreement expected on the basis of chance alone; 1 − PC refers to the maximum possible difference between observed and chance agreement.

The level of statistical significance of kappa is determined by dividing κ by its standard error (s.e.) and referring the resulting Z value to a table of areas under the normal curve to determine the p value of kappa (e.g., a Z of kappa of 1.96 is statistically significant at the .05 level). The validity of this procedure was empirically demonstrated by Cicchetti (1981) and Cicchetti & Fleiss (1977). For weighting systems to be used with the kappa statistic, see Cicchetti (1976), Cicchetti et al. (1977); Cicchetti & Heavens (1979); Cicchetti (1978); Cicchetti & Sparrow (1981); and Heavens & Cicchetti (1978).

3. The formula for $R_{i(\text{Model III})}$, when the same set of raters(reviewers) evaluate each subject, also deriving from a one-way repeated measures, ANOVA can be defined as:

$$R_{i(\text{Model III})} = \frac{MSS - MSE^*}{MSS + (MSE^*)(R-1) + \frac{R(MSR-MSE^*)}{N}}, \text{ in which:}$$

MSS = mean square between subjects; MSE = mean square error (or residual); MSR = mean square between raters (or reviewers); R = number of raters (reviewers); N = number of subjects (abstracts, manuscripts, grants).

4. The formula for the intraclass correlation coefficient (R_i), Model I when different sets of raters or reviewers evaluate each subject (e.g., abstract, manuscript, grant proposal) derives from a repeated-measures (e.g., across reviewers) analysis of variance (ANOVA) model, and can be defined as:

$$R_{i(\text{Model I})} = \frac{MSS - MSE^*}{MSS + [MSS + (R-1)(MSE^*)]}, \text{ in which:}$$

MSS = mean square between subjects; MSE = mean square error; and R = the number of ratings (e.g., reviews) per subject (e.g., abstract, manuscript, grant proposal).

The level of statistical significance of a given $R_{i(\text{Model I})}$ value is determined by referring the quantity MSS (with its number of degrees of freedom [df]) by MSE (with its df) to a standard ANOVA table.

*Note. For the $R_{i(\text{Model I})}$ case, MSE pools the variance associated with raters with the variance associated with residual.

5. The formulae for the R_i for determining the reliability of dimensionally scaled data when the numbers and specific sets of examiners may vary at each assessment (e.g., in the usual peer-review process for evaluating grants), also derive from a one-way repeated measures ANOVA model and can be expressed as:

$$R_{i(\text{Model III})} = (MSS - mMSE)/[MSS + m(R_o - 1)MSE], \text{ in which,}$$

MSS and MSE are defined as in $R_{i(\text{Model I})}$; R_o = the average number of raters per subject; $m = N(R_o - 1)/[NR_o - 1) - 2]$; N = the number of subjects (e.g., NSF grants).

The level of statistical significance of $R_{i(\text{Model III})}$ is determined by application of the formula F = MSS/MSE, which, with (N − 1) and (M − 1) degrees of freedom is referred to appropriate

ANOVA statistical tables. Here, M = the total number of ratings summed across subjects.

6. For a more comprehensive treatment of appropriate and inappropriate reliability statistics, the interested reader is referred to Cicchetti (1988), Cicchetti & Feinstein (1990), and Feinstein & Cicchetti (1990).

7. These data derive from the *Physical Review and Physical Review Letters: Annual Report 1986* (published in January 1987). We wish to express our deepest appreciation to Dr. Peter D. Adams, Deputy Editor-in-Chief, American Physical Society for making available this information as well as other related material on the peer-review process for the *Physical Review*.

8. It has been noted by several investigators in the field of peer-review research that the extent to which editors use referee recommendations is an important and often neglected variable (e.g., Bailar & Patterson 1985; Patterson & Bailar 1985; Chubin, in a 1982 peer-reviewer comment). We agree, and we have data bearing on this issue for the 1,313 different manuscripts that were evaluated by at least 2 reviewers during the period from 1973 to 1978 and were ultimately accepted or rejected by the same editor of the *Journal of Abnormal Psychology*. This journal uses a reviewer-summary recommendation format in which the submission can be rated as one of the following: accept (as is); accept subject to revision; revise and resubmit for further consideration; or reject outright. The joint reviewer recommendations compared to the final editorial decisions were as follows (with numbers and/or percentages of manuscripts following each referee or editorial judgment):

Accept-accept (17), all 17 (100%) accepted by the editor;

Accept/revise-accept (86), 77 (89.5%) accepted and 9 (10.5%) rejected;

Accept-revise/accept-revise (100), 81% accepted, 19% rejected;

Accept/resubmit (61), 38 (62.3%) accepted, 23 (37.7%) rejected;

Accept/revise-resubmit (134), 69 (51.5%) accepted, 65 (48.5%) rejected;

Resubmit-resubmit (49), 15 (30.6%) accepted, 34 (69.4%) rejected;

Accept-reject (96), 20 (20.8%) accepted, 76 (79.2%) rejected;

Accept/revise-reject (219), 40 (18.3%) accepted, 179 (81.7%) rejected;

Resubmit-reject (200), 11 (5.5%) accepted, 189 (94.5%) rejected;

Reject-reject (351), 2 (0.6%) accepted and 349 (99.4%) rejected.

The total number of accepted articles was 370 (28.2%). The number rejected was 943 or (71.8%). These data are consistent with data for both the *American Sociological Review* and the *Physical Review*, which indicate that "referees' recommendations are the major factor determining the editors' dispositions" (Hargens 1988, p. 146). Consistent with such findings, Bakanic et al. (1987) reported a correlation of .81 between referees' mean overall recommendations and final editorial decisions for manuscripts submitted to the *American Sociological Review*.

9. It should be noted that *BBS* is systematically gathering and analyzing data on the relationship between levels of reviewer anonymity and the favorability and usefulness of the referee report, as indicated both by the referee's recommendations and the author's subjective ratings.

10. The authors extend appreciation to staff personnel at NSF for making available this information, which can also be obtained, in more detail, by requesting appropriate NSF publications.

11. It should be noted that the various funding agencies use different priority rating systems. For example, the Veterans Administration uses a 10-to-50 rating scale, which is the reverse of the NSF scoring system. Thus, *lower* scores represent *higher* quality grant proposals, and vice versa. In contrast to both these rating systems, reviewers for grants submitted to the American Heart Association used a 10-category, ordinally scaled grading system in which high priority (for funding) was defined as 1–3, intermediate priority as 4–7, and low priority as 8–10 (i.e., see Wiener et al. 1977). Concerning these varying scales of measurement, the empirical work of Cicchetti et al. (1985) indicates that no appreciable increment in interrater reliability is achieved by increasing the size of a rating scale beyond seven ordinal or quasi-dimensional points or categories. These data serve to validate and generalize the implications of the results of an earlier investigation that demonstrated that analogue rating scales (0 to 100 points) were no more reliable than three-category ordinal scales (here the Present State Examination [PSE], in assessing extent of psychiatric symptomatology (Remington et al. 1979).

12. It is of interest that Professor Jared Diamond describes many difficulties that composers have encountered in their attempts to have their works published or supported by grants. These and other parallels between struggling musicians and scientists were presented in the March 21, 1985, issue of *Nature*, in commemoration of Bach's 300th birthday.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Peer review: An unflattering picture

Kenneth M. Adams[1]

Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109-0704

Electronic mail: *gdvr@umich.cc.umich.edu*

Cicchetti brings a masterly touch to an enduring problem in modern science. The problem of peer review calls to mind the old saying about the weather: "Everyone complains about it but no one does anything about it." One may find this more amusing later in one's career than earlier.

Many researchers will regard Cicchetti's results with initial discouragement, if not outright embarrassment. Critics will be quick to point out the very real limits that low reliability places upon validity. Given these data, one must further suspect the peer-review process in science of being flawed. Yet several points deserve careful consideration.

First, the entire process of reviewing manuscripts or grants is one of considering *new* information. Decisions concerning the disposition of these entities are simple in their result, but often complex in their structure. Cicchetti has done well in trying to capture disposition as the most solid data point. In the case of a manuscript, a research report often contains components reflecting various aspects of the research enterprise: (1) command of existing knowledge, (2) skill at formulating the research design, (3) expertise in analysis of the executed study, and (4) wisdom in guiding the reader in how to understand the place of the study in our knowledge. Predictably, the ability of investigators in each of these areas will not be uniform. Equally predictable will be the varying degree to which reviewers may have special skills and capabilities to judge the success of the project in these components. Disagreement by colleagues about papers

may represent a healthy state of affairs in which the author and reviewer are pursuing their own kind of academic freedom. A similar situation exists with respect to the novelty of information in grants and their multiple components. (As an aside, it seems that many reviewers have forgotten that grant proposals are just that – proposals.)

Second, it is interesting that all sciences are roughly in the same range of reliability when it comes to reviews. This gives some indirect support to my first point about why novel information is judged imperfectly by experts in a number of realms. One problem with manuscript publication decisions is that reviewers expect to be asked to vote on the publication of a manuscript. I would add to Cicchetti's suggestions a plea that journals stop routinely asking reviewers whether a paper should be accepted. This recommendation on a reviewer's part may not be at all helpful for either the editor or the author of the manuscript. Many journals do appeal to reviewers to *avoid* making a recommendation on acceptance or rejection in their review comments intended for the author. It is a plea that often goes unheeded. Journal editors are probably in a far better position to weigh the various factors affecting potential publication than are most reviewers. Editors tend to send manuscripts to reviewers who have special expertise for consultation on certain points. In supplying this information, reviewers may have their own ideas of what should or should not be published, but editors often are not aware of the fabric of reviewers' editorial philosophy.

Third, it is probably inimical to true academic freedom to train reviewers. It is apparent, however, that the development of constructive reviewers is too often haphazard and uncertain. Good reviewing should be recognized by journals, institutions, and societies. We must find ways to groom reviewers without creating undue bias or suppressing scientific freedom.

Fourth, while the reliability of decisions in evaluating manuscripts and adjudicating grants is poor in both cases, the results are clearly more deleterious with grants. If one wants to publish a new manuscript, there will certainly be some way to do it eventually – even if it is not in one's preferred journal. A similar situation does not prevail with respect to the awarding of grants: no grant, no money. Reality in many areas of research dictates that if the federal government does not fund certain kinds of research, it just won't be done. This has brought the peer-review process under even more scrutiny, and the sense of frustration and arbitrariness that some "pink sheet" recipients feel is not going to be assuaged by the findings.

The remedies suggested for the manuscript review process need more dramatic implementation with respect to grants. The adjudication of grants is a far more social and political process than manuscript review. Steps must be taken to humanize the grant review process by putting reviewers' names on their opinions. The National Science Foundation (NSF) experience in trying to create blind applications did not work initially, but remains worth trying.

Fifth, Cicchetti's finding that reviewers can agree more easily on *less* desirable research than on more desirable research parallels the situation now usually extant in the individual review. As a general rule, reviewers spend far too little time being constructive, collegial, and consultative in their reviews. One sure-fire way to limit the impact of our character defects and make the review process more constructive would be the aforementioned requirement that reviewers reveal their identities in all cases. The temptation to be petty or take cheap shots when reviewing others' work would diminish thereby.

In closing, I would like to emphasize that the methodological basis of Cicchetti's investigation seems sound. Agreement between reviewers of manuscripts and grants is discouragingly low. The reasons for this are many, but the net result is that this study holds a mirror up to peer review that provides a distinctly unflattering picture. The remedial steps proposed by Cicchetti require urgent attention. I would underscore his suggestion that reviewers be identified, encouraged, and rewarded by whatever means available for providing quality consultation on manuscripts. The situation with grants is far more complex, but even the most stalwart defenders of current federal funding review methods cannot ignore this evidence suggesting that some scientific decisions resulting in funding or nonfunding are probably being made in a nonsystematic way. I would doubt that we are distinguishing between "shades of excellence" any more; a certain degree of caprice has entered the picture. In a complex world that can be helped by our research at a variety of levels, this should open our minds to constructive alternatives.

NOTE

1. The author is affiliated with The Veterans' Affairs Medical Center, Ann Arbor, MI.

Does the need for agreement among reviewers inhibit the publication of controversial findings?

J. Scott Armstrong[a] and Raymond Hubbard[b]

[a]*The Wharton School, University of Pennsylvania, Philadelphia, PA 19104*
[b]*College of Business and Public Administration Drake University, Des Moines, IA 50311*

As Cicchetti indicates, agreement among reviewers is not high. This conclusion is empirically supported by Fiske and Fogg (1990), who reported that two independent reviews of the same papers typically had no critical point in common. Does this imply that journal editors should strive for a high level of reviewer consensus as a criterion for publication? Prior research suggests that such a requirement would inhibit the publication of papers with controversial findings. We summarize this research and report on a survey of editors.

Prior research. Horrobin (1990) suggests that the primary function of peer review should be to identify new and useful findings, that is, to promote the publication of important innovations. This function is typically subordinated to the quality control aspects of peer review, however. The quality control approach looks for agreement among the reviewers. The result, Horrobin claims, is that competent research yielding relatively unimportant findings is more readily accepted for publication.[1] He provides numerous examples of harsh peer review given to important research that presents controversial results.

The popular press often reports difficulties associated with the publication of important research findings. The scanning tunneling microscope (STM) is a case in point. The STM is capable of distinguishing individual atoms and has been hailed as one of the most important inventions of this century. It earned a Nobel Prize in physics for its inventors. Nevertheless, the first attempt to publish the results produced by the STM in 1981 failed because a journal referee found the paper "not interesting enough" (Fisher 1989).

Armstrong (1982c) provides additional examples of lapses in the peer review system, along with summaries of empirical evidence that disconfirming findings about important topics are difficult to publish. Among these, the experimental studies by Goodstein and Brazis (1970) and Mahoney (1977) are of particular interest. They found that reviewers were biased against negative findings. They rejected these papers on the basis of poor methodology while accepting papers with confirmatory outcomes that used the identical methodology.

Given the above results, one might expect that if editors rely on consensus among reviewers for their publication decisions, few controversial findings will be published. This problem could be especially serious in social science journals. These journals generally have low acceptance rates and their editors may decide to publish only manuscripts with high agreement among reviewers.

A survey of journal editors. To assess how journals treat

empirical papers that present controversial findings, we conducted a survey of 20 current or recent editors of American Psychological Association (APA) journals. The two-page questionnaire, together with a stamped, self-addressed return envelope, was mailed out in March 1990. We followed up with phone calls 10 days after the mailing.

Replies were received from 16 of the 20 editors. One question asked: "To the best of your memory, during the last two years of your tenure as editor of an APA journal, did your journal publish one or more papers that were considered to be both controversial and empirical? (That is, papers that presented empirical evidence contradicting the prevailing wisdom.)" Seven editors could recall none.[2] Four said "yes" and indicated that there was one paper. Three editors replied that there was at least one. Two said that they published several such papers. It seems that controversial empirical papers do get published, but infrequently. Almost half the editors could not recall publishing such papers in the past two years.

We then asked about the peer review for the one published controversial empirical paper that they remembered most clearly. The question was worded: "How did the reviewers respond to this paper?" A five-point scale from "unanimously accepted" to "unanimously rejected" was provided, as well as a "don't recall" option. One of the nine respondents to this question reported unanimous acceptance, three reported "majority in favor," four reported "even split," and one answered "don't recall." In response to a question on this published paper's contribution to the discipline, one editor said "not important" four said "somewhat important," and four selected the highest rating, "important."

The editors were also asked if they had *rejected* any papers that were controversial and empirical. Six of the editors stated that they did not receive such papers, and four said they could not recall any. The six editors who rejected papers with controversial findings did so, they said, because of poor methodology and poor supporting arguments. Of the rejected papers that the editors "remembered most clearly," only one was "unanimously rejected;" a "majority not in favor" was reported for two, an "even split" for two, and a "majority in favor" for one. Three papers were rated as "not important," and three as "somewhat important."

These results suggest that one can get reviewer agreement on controversial empirical papers. Moreover, most of these papers are published without high levels of reviewer agreement. Apparently, editors do not rely solely on reviewer agreement.

It is interesting that our survey found only two instances of unanimous reviewer agreement for empirical papers with controversial findings. In one case, the recommendation was to reject. In the other, it was to accept. In the case of the accepted manuscript, it should be noted that the editor had invited this submission and had selected reviewers who, he said, were sympathetic to its content.

Our survey indicates that some controversial empirical papers do get published, even when there is disagreement among the reviewers. The willingness of editors to publish such papers is encouraging. On the other hand, 7 of 16 editors could recall no instances of publishing controversial empirical findings. Consequently, we consider some strategies to increase the odds of publishing this type of paper in the next section.

Possible solutions. Some methods that are currently used by journals should help.

1. Some journals' editorial policies allow the author to submit a list of possible referees, one of whom would be selected.
2. Items can be included on structured rating sheets so that reviewers rate the extent to which the findings are controversial. Editors can then give such ratings more weight.
3. Additional reviews can be sought when papers are judged to contain controversial findings. (This strategy was used for only one of the nine published papers and for only one of the six rejected papers in our survey.)
4. Special appeal procedures may help for controversial papers. This might involve other members of the editorial board.
5. Controversial papers can be reviewed initially *without revealing the findings*. This procedure is currently used by the *International Journal of Forecasting*. It has not been used frequently but, when used, it has been beneficial.
6. Provide a section of the journal for "Controversial Findings." The selection of an editor for such a section would indicate the journal's willingness to provide space for such studies. Unfortunately, the one application of this approach that we know (Armstrong 1982b) has produced only one submission, and the findings reported in that submission were not controversial, only the methods were.

Rather than looking for agreement, it might be useful to seek reviewers to act as advocates. This advocacy system would be used for papers that are designated as containing controversial results. A paper could be so designated by the author, the editor, or a reviewer, after which special advocacy procedures would be used. This might include some of the above mentioned suggestions. In addition, one could use more reviewers in an effort to find an advocate. An advocate could *insist on* publication; a note could be included with the published paper so that reviewers are, in a sense, willing to stake their reputations on the paper.[3] Through this note, the readers would receive information about the nature of the acceptance. All referees could be given the opportunity to write peer commentary on the paper. This procedure would greatly increase the likelihood that important papers would be published. The increased effort given to reviewing might also improve quality control.

Conclusions. Controversial empirical papers are expected to receive harsh treatment in peer review, but our survey indicates that such works occasionally get published, sometimes without much peer agreement. More can be done to encourage publication, however. We suggest ways to accomplish this, in particular, the use of an advocacy procedure that explicitly recognizes the need to promote this type of research.

NOTES

1. It is not clear that the quality control function is performed well. About one-third of the papers in biomedical journals were found to contain citation errors, and one-third also incorrectly quoted findings from the literature (Evans et al. 1990). In addition, Hubbard and Armstrong (1990) found that 60% of published replications with extensions in three leading marketing science journals failed to support the original findings.
2. "Unfortunately," according to one respondent. Also, in followup phone calls, several editors expressed the desire that more such work be submitted.
3. McNutt et al. (1990) found no differences in the quality of reviews based on whether or not they were signed by the reviewer. Also, those who signed the reviews were more likely to recommend acceptance.

Reliability, fairness, objectivity and other inappropriate goals in peer review

John C. Bailar

Department of Epidemiology and Biostatistics, McGill University School of Medicine, Montreal H3A 1A2, Canada

The following remarks are cast largely in terms of peer review of manuscripts for possible journal publication, but they also apply generally to peer review of grant proposals.

Cicchetti consistently misses the mark. The purpose of peer review is not reliability, but to improve decisions concerning publication and funding, and these authors simply do not discuss this critical matter.

Cicchetti fairly states the value of both redundancy and "creative" disagreement in peer review, but fails to acknowledge adequately that editors and grants managers choose (and should choose) reviewers for their different, complementary

expertise. For example, a report on a randomized trial of a new drug for the control of hypertension might be sent to a cardiologist, a pharmacologist, and a statistician. They would, and should, be alert to quite different kinds of strengths and problems, and there is no reason to expect either their detailed reports or their summary judgments to agree. Too much agreement is in fact a sign that the review process is *not* working well, that reviewers are not properly selected for diversity, and that some are redundant. Without this negative point, measures of inter referee agreement are of no value in assessing peer review mechanisms.

Cicchetti refers to the role of the reviewer in informing the judgment of the editor or grants manager, but does not adequately stress the point that reviewers are no more than sources of relevant information. I know of no leading journal where decisions about publication are made by a "vote" of the reviewers. As a former editor (of JNCI [Journal of the National Cancer Institute], 1974–1980) I had a section on the reviewers' form asking for a judgment about publication (publish as submitted, publish with minor revisions, etc.) and regularly found that it was of little value in sorting out the merits of a paper. There is no substitute for careful study of specific comments, integrated with the wisdom of editorial board members and, sometimes, special consultants. As a result, it was not unusual for us to publish papers that three reviewers had recommended for disapproval, and vice versa.

A further point is that editors can adjust for (or sometimes deliberately use) reviewer bias. There has been few studies of the comments of peer reviewers to date, and all have focused on what reviewers write, not on the critical issue of how they have affected the information base on which a decision was made. I knew and regularly used reviewers who could never bring themselves to criticize a colleague directly, though their detailed comments were full of insight. And I used others who could never find a paper good enough to publish; with appropriate interpretation, their comments, too, were helpful. On rare occasion, when it appeared that an editorial decision might be challenged on the basis of the position or prestige of an author rather than scientific merit, I deliberately chose reviewers from one or the other camp to ensure that a strong and balanced review would be on the record. Some other editors do the same, and our journals have been the stronger for it.

The paper by Peters and Ceci (1982) is a weak reed. Shortly after it was published, I wrote to Peters with some specific questions about their work. I made at least two telephone calls to verify his address at the time, but received no reply to my letter. Folklore to the contrary, few first-class letters are really lost by the Postal Service. I must assume that I received no reply because their answers would have undercut the strength of the conclusions in their paper. I cannot find my copy of the letter at this late date, but I recall that two points of special interest were *how* they "randomly" chose the papers they resubmitted (in more detail than was given in their paper), and how (also in detail) they revised titles and content to reduce the likelihood of detection of their own fraud. Most long-time editors have had the experience of publishing papers and almost immediately regretting the decision to publish, so that biased selection of winners or losers is simply not informative about practice in general.

I am concerned that Cicchetti accepts without comment the appropriateness of studies carried out without the consent of the subjects, whether journals (and editors) or reviewers. Substantial investments of time, and direct financial investments as well, have been requested under false pretenses in the name of "science." I know and understand the arguments that some research cannot be carried out if the subject is properly informed, but reject any notion that such research thereby becomes ethical.

Editors do, and should, base their editorial decisions in part on results. Many negative studies are never properly completed; others are presented in slap-dash fashion. Some are trivial because few knowledgeable investigators would have expected anything other than negative results; still others have samples too small to have much chance of showing a real effect even if one should be present. Many other negative studies are indeed published in the sense of "made public," but not as full-length original contributions. Instead, their results may be disseminated as abstracts, in short sections of later papers that extend the work, or even by word of mouth. Arching over all of this is the proper concern of editors about their readers' interests. I know of no evidence that readers are harmed by editorial decisions that depend in part on results. Many fewer people, and different people, may need to know that something did not work than would need to know what did work. A good editor must be even more concerned about readers' legitimate interests than about authors' complaints, and the "need to know" is chief among these. Thus, some kinds of bias against publication of negative results in the usual full form is entirely appropriate and should be encouraged.

Cicchetti's section 7, on improving the reliability of peer review, tacitly takes improved reliability as an important goal. But the fundamental objective of peer review, and of the manuscript selection process in general, is not "fairness" to authors (though that may be a welcome byproduct). It is to improve decisions. Will larger numbers of reviewers, better training, or instructions for reviewing improve decisions? The matter has not been studied, perhaps because no one has yet devised a good measure of the quality of decisions to publish or disapprove. I know of no good statistical evidence that blinding reviewers to authors, or authors to reviewers, affects editorial decisions in generally good or bad ways. There is substantial anecdotal evidence, however, that both the strengths and the weaknesses of a paper are appraised more accurately when reviewers know who the authors are, but not vice versa.

I find no recognition here that editorial decisions can, do, and should make use of criteria other than abstract scientific/technical merit. Such criteria include originality, the suitability of the topic for a given journal, readability and the appropriateness of length and style, the need for a balance of topics in journals with broad coverage, the importance of findings to readers, and even whether there is reason to suspect unconscious bias or deliberate error in the data or the analysis.

Overall, I believe that Cicchetti's paper shows a misunderstanding of the role of peer review as an aid to editorial decisions and grants management.

The predictive validity of peer review: A neglected issue

Robert F. Bornstein
Department of Psychology, Gettysburg College, Gettysburg, PA 17325

Cicchetti's analysis of inter-reviewer reliability in manuscript and grant proposal assessments is both timely and valuable, and will help to resolve a number of unsettled issues in this area. Cicchetti – like most researchers investigating aspects of the peer review process – focuses mainly on reliability issues in peer review. His analysis confirms that inter-reviewer reliability in manuscript and grant proposal assessments is generally quite low. An important question remains unanswered, however: What do we know about the *validity* of peer review? Peer review is, at least in part, an assessment tool designed to identify the best research efforts in a given sample of manuscripts (see Bornstein 1990; Eichorn & VandenBos 1985). Thus, we should be able to demonstrate empirically that peer reviews have predictive validity and that reviews can discriminate high-quality from low-quality research.

Unfortunately, designing studies to investigate the predictive

validity of peer review is considerably trickier than designing studies to assess inter-reviewer reliabilities. In particular, difficulties in selecting an appropriate criterion measure with which to assess research quality have hindered efforts to conduct empirical research on this topic. Researchers typically use journal citation frequency as a criterion measure in these studies, testing the hypothesis that, if manuscript reviews have predictive validity, then papers that receive highly positive reviews should be those that report the most important, well-designed studies. These papers should therefore be cited more frequently than papers that receive less positive reviews (Gottfredson 1978). Although citation frequencies have been used to assess journal quality and the eminence of individual researchers (Garfield 1972; Lindzey 1977), the use of citation indices as a measure of the quality of a particular piece of research is questionable for several reasons.

First, we make a number of assumptions regarding the quality of research based on the journal in which it appears. If a paper is published in a prestigious journal, we infer that it must be good and valuable research. Were the same paper to appear in a less prestigious journal, it would most likely be seen as less rigorous and important, and we would be less likely to cite it. Clearly, the well-known "halo effect" (Nisbett & Wilson 1977) influences our perceptions of psychological research.

Second, variables unrelated to research quality will influence the number of citations a paper receives. Mediocre research in an area that is tangentially related to a variety of topics will probably receive a greater number of citations than excellent research in a more obscure and narrowly defined area. Research on experimental design and methodology tends to be the most widely cited in all branches of science (see Lindsey 1978). This is not surprising, given that such papers have implications for a wide variety of topics.

Third, if a relationship between citation frequency and research quality does exist, this relationship is not likely to be linear. The relationship between research quality and citation frequency probably takes the form of a J-shaped curve, with exceedingly bad research cited more frequently than mediocre research (e.g., as an example of an idea or a line of research that turned out to be a blind alley, or as an example of what *not* to do in a particular area).

Finally, this outcome criterion does not allow the predictive validity of negative manuscript reviews to be assessed. Because studies receiving negative reviews may never be published (or may be published in obscure journals having very limited readerships), it is not possible to use criteria such as citation indices to assess the validity of these reviews.

In any case, there have been very few studies of the predictive validity of peer review, and the results of these have not been reassuring. Gottfredson (1978) compared reviewers' ratings of psychological research papers to the number of citations received by these papers in the first nine years following publication. He found only low to moderate correlations between reviewers' estimates of manuscript quality and impact and the number of citations received by a paper. Reviewers' ratings of research impact were most strongly predictive of subsequent citation frequencies ($R = .37$). Ratings of research quality did not fare as well ($R = .24$).

Thus, we know that: (1) inter-reviewer reliability in peer review is generally low (as Cicchetti et al. and others have demonstrated); and (2) we have no hard evidence that reviews have predictive (or discriminant) validity. To the extent that "confirmatory bias" (Mahoney 1985) and other variables unrelated to research quality demonstrably affect the outcome of peer reviews, the internal validity of the review process is also compromised. To anyone interested in the process of scientific inquiry and the dissemination of scientific knowledge, such findings are – to say the least – a bit unnerving. Because we regard peer review as a "test" or measure of the scientific worth of manuscripts and grant proposals (Bornstein 1990; Eichorn & VandenBos 1985), we should be able to demonstrate that this "test" is psychometrically sound. Yet, even a cursory reading of the American Psychological Association's *Standards for educational and psychological testing* (APA 1985) reveals that peer review fails miserably with respect to *every* technical criterion for establishing the reliability and validity of an assessment instrument (see APA 1985, pp. 9–44). If one attempted to publish research involving an assessment tool whose reliability and validity data were as weak as that of the peer review process, there is no question that studies involving this psychometrically flawed instrument would be deemed unacceptable for publication.

It is not too late to make changes in the peer review process that will help improve its reliability and validity. Cicchetti makes some useful suggestions in this area, and other researchers (e.g., Bornstein 1990; in press; Mahoney 1985; 1987) have also proposed procedures for improving the review process. At any rate, in addition to investigating reliability in manuscript and grant proposal assessments, we must now rigorously assess the predictive and discriminant validity of peer review. Altering the peer review process in order to maximize its reliability and validity may be difficult (in practical/procedural terms), costly (in monetary terms), and somewhat risky (the changes could create new problems instead of solving the old ones). I believe, however, that the costs and risks associated with changing – even experimenting with – the review process are far less than the costs and risks of continuing to support uncritically a process that, in its current form, has many significant flaws.

Does group discussion contribute to the reliability of complex judgments?

Patricia Cohen

Columbia University School of Public Health and New York State Psychiatric Institute, New York, NY 10032

Cicchetti is to be congratulated on a useful summary of our knowledge in this field. It seems reasonable to conclude that these complex human judgments cannot be made very reliably, a state of affairs that has also been demonstrated in other arenas, including student and personnel selection and the identification of diagnostic levels of psychopathology. As a long line of research in these areas has shown, when more objective indices are available, they will typically have higher validities than decisions based on human judgment alone. Such objective measures in manuscript evaluation might include the status of the institution and the publication record of the authors. Alas, in such a case the "objective" criteria lead directly to the kind of bias that peer review is designed to minimize.

When objective criteria are biased, the only sound alternative is to increase the number of evaluators, assuming that they will be less subject to such bias. Here practical constraints intrude, as it is hardly possible for all journals to obtain a sufficient number of reviewers for all articles to ensure a reliable composite review. The situation is somewhat different with regard to peer review of grant proposals, however. Here two reviewers typically examine the entire set of materials provided and report both a summary and their critiques to a larger panel. The larger group then discusses the material presented to them and may often review some portion of the proposal as well, should a specific issue require it. This larger panel may be thought of as a means of increasing the size of the review panel and is certainly intended to improve the reliability and validity of the resulting judgment.

To my knowledge, no hard evidence is available regarding the effectiveness of this subsequent segment of the review process. Because the judgments are very far from independent, and

because there are social and other disincentives to strong disagreement, panel ratings cannot be used as evidence of real consensus. Nevertheless, in a period in which the ratio of available funds to trained scientists is shrinking, it seems worthwhile to consider full scientific investigation of this review process. For example, it is at least possible that averaged ratings from four independent reviewers of a proposal would agree with an average from four other independent reviewers at a higher level than would the ratings between reviews from two independent groups that had been carried out in the current fashion. A larger number of noninteracting "primary" reviewers of grant proposals might be no more demanding of either scientists' time or of agency funds than the current procedure.

Other investigations of potential biases or sources of unreliability in the present grant review process are also easily imagined. For example, it is widely thought that inexperienced reviewers may be "tougher" on proposals than more experienced reviewers. A peer review committee officer has also told me that grants reviewed early in a session tend to be discussed more thoroughly and thus evaluated more critically than those reviewed later. Information on both of these issues could be compiled readily. If they were found to be nontrivial sources of bias, reviewers could be made aware of these problems in an attempt to minimize them.

Finally, it is worth emphasizing that measures of agreement may overestimate the reliability of the judgments concerning grants or manuscripts in the critical region of decision where the cut-off occurs. In the present framework of extremely tight funding, a cool reception by a single reviewer may be sufficient to preclude a fundable priority. Therefore, differences between rating scale use habits of different committee members can make some members much more influential with regard to the funding outcome than others. If ratings were standardized for each committee member before they were combined, or if members were each to place grants into the same preset distribution, potential abuses relating to this likely source of error would be eliminated.

Consensus and the reliability of peer-review evaluations

Stephen Cole

Department of Sociology, State University of New York at Stony Brook, Stony Brook, NY 11794

Research I have conducted suggests that the low levels of reliability in peer review evaluations described by Cicchetti are not an artifact of the peer-review system or of reviewer bias, but reflect the low levels of cognitive consensus that exists at the research frontier of all scientific disciplines (Cole 1983; Cole et al. 1981). I have argued that the level of cognitive consensus in the social sciences is not significantly lower than that in the natural sciences. Cicchetti presents some evidence supporting this view. Discussing the peer-review system used by *Physical Review Letters*, he quotes from the editors' policy statement on how difficult it is to make decisions; in only 10 to 15% of submissions do the 2 referees agree on whether the article should be accepted or rejected. Cicchetti concludes that if a more systematic study were undertaken, "we would predict that levels of referee consensus for *Physical Review Letters* would be of the same relatively low order of magnitude . . . characterizing general journals in many other disciplines" (sect. 4.5).

The assumption that the natural sciences have higher levels of consensus than the social sciences has been used to explain and justify the higher rejection rates of social science journals (Hargens 1988). I see the difference in rejection rates between natural and social science journals as resulting from differences in the amount of space available, the diffuseness of a field's journal system, and, most important, norms concerning the desirability of making Type I or Type II errors (Cole et al 1978; 1988). Natural scientists prefer to make Type I errors; social scientists, Type II errors.

My analysis leads me to be more critical than Cicchetti of current journal practices. He believes that since most articles in high-rejection-rate fields are eventually published and since authors of many rejected articles in low-rejection-rate fields do not resubmit, the system is working well. I agree with him for the low-rejection-rate fields, but disagree for the high-rejection-rate fields.

If there are *approximately equal* and *low* levels of consensus in fields like physics and sociology, respectively, this means that physics journals are publishing papers that many physicists believe are of little significance and sociology journals are rejecting papers that many sociologists would find useful. The policy followed in physics allows the diverse scientific community to decide what is useful and neglect the *published* articles that are not useful. The policy followed by the sociology journals allows a sample of two or three referees influenced by norms calling for high rejection rates to make this decision. This has many negative consequences for the development of the field. We must realize that as a result of lack of consensus and norms supporting high rejection rates, many of these rejections are "unjustified," thus giving the field a pervasive sense of inequity, bias against some work styles, and irrationality. This serves to reduce motivation and seriously interferes with the communication of ideas.

In physics, two journals, the *Physical Review* and the *Physical Review Letters*, publish a large portion of all the literature. By monitoring what is published in these journals physicists can be sure of being up-to-date on their research interests. In sociology the two leading journals publish a very small portion of the literature in the field. Much research that would be of use to some segments of the community is rejected from high-visibility journals and must be published in obscure sources. This makes it more difficult to keep up with the latest developments in areas of interest. Communication is further hampered by long delays, sometimes amounting to years, resulting from the inefficient publication system.

The only disadvantage for a field like sociology in switching its publication system to one more similar to that used by the physicists would be the increased cost of journal publication. But given the importance of publication for advancing one's career, it would seem that most authors would be willing to reduce the length of their papers and pay modest page charges, even if they had to pay these out of their own pockets.

Another possible argument against increasing the acceptance rate in high-rejection-rate fields would be the potential decrease in the quality of published articles. Among those who argue that journal rejection rates result from the level of disciplinary consensus there is the implicit assumption that because of a lack of agreed-on criteria in the social sciences most articles submitted to the journals are of "poor" quality and not "really" publishable. There are two problems with this assumption. First, it assumes that, because most of the articles submitted to natural science journals are accepted, they "really" deserve to be published. Many studies of citation patterns, however, have shown that the bulk of articles published in physics journals, for example, are rarely if ever cited (Meyer 1979). There is also qualitative evidence that natural scientists are just as likely as social scientists to disparage the quality of articles in their journals. For example, Mulkay and Williams (1971), in their study of physicists in England, report that "*all* our respondents thought that the vast majority of papers in the journals which they read were of poor quality or of little significance." (p. 74)

The second assumption is that the articles rejected by the social science journals "deserve" to be rejected. Stinchcombe and Ofshe (1969) conducted an analysis in which they assumed

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 ranks 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.

What should be done to improve reviewing?

Rick Crandall

Journal of Social Behavior and Personality, Box 9838, San Rafael, CA 94912

The title of this commentary uses the word *should* rather than *can* because there is a professional ethical issue here that is not being faced by journals. Cicchetti's target article makes it obvious that the review process is unreliable. This conclusion should be dealt with, yet most journals blithely continue to use the process without trying to improve it. We must question the validity of the review process, and in fact the whole editorial process, for at least four reasons. First, there is no evidence to support them. The fact that very few journals are making any effort to improve the editorial process is unethical, in my opinion. (Crandall 1990). Second, low reliability can limit validity. Third, Cicchetti and numerous others have documented the existence of systematic biases in the editorial process. It is frightening that, with the lack of other data on systematic "true" variance in the review or editorial process, one must wonder whether these biases account for the little reliability we do achieve. Fourth, in most fields, the majority of papers that are turned down are ultimately published in good journals.

It is clear to me that if we really wanted to, we could achieve high reviewer reliability. Although many remedies could help, the obvious one is to train reviewers. If we can train small dogs to jump up and down off moving horses at the circus, surely we can train scientists to act as reviewers, despite the complexity of the task. Thus, I believe that underlying the question of what should be done is why we haven't bothered. Although I may tend toward a moralistic interpretation of the reason, it is likely that "people" don't think better review reliability is really needed, or the payback would not justify the investment required.

Our journal has a major focus on testing and improving the editorial/review process. We have some unpublished data documenting the lack of existing training for reviewers by journals. Of 76 social and behavioral sciences editors who answered the question, "Do your reviewers receive any training, besides a general instruction sheet?" only two said yes, and their training was not major. It is clear that training is totally neglected. Reviewers must be expected to learn on the job at authors' expense. I had to laugh when I saw the recent American Psychological Association announcements recruiting members of underrepresented groups to be reviewers for journals. The only qualification mentioned was that they must have published articles in peer reviewed journals, because "the experience of publishing provides a reviewer with the basis for preparing a thorough, objective evaluative review" (*American Psychologist* 1989). This is logically analogous to requiring people to be executed before they can become hangmen. I guess it is better than nothing!

What training can be done? Some journals do some minimal training or screening of reviewers. I'm sure that editors try to weed out "bad" reviewers, so they should be gradually improving their pool. Cicchetti suggests that reviewers should be rewarded in some way. Several journals, including ours; have some explicit procedure for adding and subtracting reviewers from the published editorial panels depending on their work. The most common training approach is probably to exchange reviews after the decision so that each reviewer can learn from the other. This could be improved considerably if the editor's decision letter were sent to the reviewers or if they were told explicitly how good their review was and what the other reviewer did right or wrong. In other words, review the reviews! It would be a logical next step to have training manuscripts with prototype "ideal" reviews. When you train people (or animals) to perform a task, you also have explicit training goals and criteria. These should be written down and transmitted to reviewers.

It has been demonstrated that reviewers are, in fact, trainable. To use a simple, confirmed example, we have been successful in requiring reviewers to return reviews in two weeks. This was discussed and documented 20 years ago by a sociology journal editor (Rodman 1970).

Another group in need of training is authors. Many manuscripts that come in to our journal would probably be better received if they were better presented. Although I have argued that efforts to get authors to "improve" (Boor 1986) amount to blaming the victim and would be better spent testing the process (Crandall 1987b), author training could improve things a bit. What may be more important than presentation variables, however, is educating all authors to the "folklore" of how to improve their chances of getting published. We have attempted to do this (Anon. 1987; Wyer et al. 1987).

A number of other issues could be elaborated on in this area. I have mentioned only a few here. Elsewhere, I have suggested a number of simple standards that should be required of journals to make improvements (Crandall 1986). Among them are enlarging the pool of reviewers and editors so the excuse about being "overworked volunteers" could be eliminated, and requiring journals to make timely decisions and to commit to doing research to improve things.

I have a few brief reactions to other points raised by Cicchetti. He suggests using more reviewers to increase reliability. We do that, and it also increases speed, since you don't have to wait for a slow third reviewer. Their Table 3 data on the effects of adding reviewers on acceptance rates may also be relevant here, however. Causality goes in both directions. In addition to weaker papers being sent out for more reviews, they note elsewhere that the more reviewers you have, the greater the chance that an important flaw will be caught. Unfortunately, there is also a greater chance that a negative "picky" review will encourage an editor to reject the paper. Common sense, as well as experimental evidence (Amabile 1983), tells us that when a journal (or grant) review process rejects 75% of the submissions, a reviewer will look a lot smarter erring on the side of harsh criticism than on the side of leniency. I believe that reviewers have a negative bias unless trained otherwise.

A very important issue tends to be overlooked in discussions of reviewing. We have editors who are supposed to be capable of making independent decisions. With their exposure to all papers submitted, they should be "super reviewers." Yet, too many times editors may abdicate responsibility for editorial decisions. They should not be conducting a vote and averaging reviewers opinions. They can override reviewers! They should be using the reviewers to review and making the editorial decisions themselves. Too many times, I've seen cases in which editors do not seem to behave this way. I believe that many editors do not even read the papers for which they are supposed to have editorial responsibility. If they don't read them closely, how can they be the editors? Or how can they give reviewers feedback, as discussed earlier?

We need more highlighting of issues in the review process as Cicchetti et al. have done. We also need to take a broader look at the problems, and have a commitment by journals, associations, agencies, and reviewers to do a better job. I have directly challenged groups who publish journals to take responsibility and do something about these problems even more pointedly than I have here (e.g., Crandall 1987a). There has been little or no response. And I have seen little progress or commitment in the last 10 years. Why do you think this is?

Peer review: Explicit criteria and training can help

Fred Delcomyn
Department of Entomology and Neuroscience Program, University of Illinois, Urbana, IL 61801
Electronic mail: *bugoutfd@uxl.cso.uiuc.edu*

I doubt there is an active scientist in the United States who could not write an essay on peer review at least as long as Cicchetti's excellent target article. It is inevitable that our passions will be aroused by a process that can determine the success or failure of applications for funding or the acceptance or rejection of journal articles. After all, careers are at stake. Does peer review work? Yes, up to a point, as the target article shows. Peer review can be used to classify documents like grant applications into broad categories such as excellent, fair, and poor. Expecting it to allow one to distinguish between applications that are all in the excellent category, however, is like expecting to be able to measure the diameter of a nerve cell with a meter stick.

I think the prospects for refining peer review for the selection or rejection of manuscripts for publication is greater. Cicchetti makes several good suggestions, but other things can be done as well. I have two specific proposals: Make the criteria to be used in a review more uniform and explicit, and "train" reviewers.

Explicit criteria. What is the point of making review criteria more explicit? We all know what constitutes a good paper, right? Wrong. As Cicchetti points out, two reviewers can make similar comments about a paper and yet have opposite recommendations as to its acceptability. Let's cut down on this kind of conflicting advice by agreeing on the ground rules. These rules may be different in different disciplines, even in different fields, but there is no reason journals cannot develop an explicit set of guidelines for acceptable manuscripts, guidelines that can be published in the journals themselves.

What should these criteria or guidelines be? Some journals already provide a partial list in their forms to reviewers. The *Journal of Experimental Biology*, for example, asks reviewers for an assessment of experimental techniques, presentation of data, and quality of reasoning, among others. There are clearly many other aspects of a paper that can be evaluated. Based on my experience as a physiologist, I will make three specific suggestions.

(1) Do the experiments whose results are reported answer the questions set out in the introduction? I see no point in cluttering an already crowded literature with a publication that confuses issues by seeming to address one question when it actually addresses another (or none at all).

(2) Are the experiments carefully executed and controlled? The issue here is whether conclusions can confidently be drawn from the results, or are the procedures so flawed that no firm conclusions can be made.

(3) Are the conclusions that the author(s) draw supported by the data actually presented? There is a place for speculation and formulation of new hypotheses, but authors must obviously take care to separate their conclusions from their speculations.

What if the data seem to contradict someone's favorite hypothesis? Too often one hears of the struggles of a researcher to publish work that disputes someone else's data or interpretations. Here is where explicit criteria would be so helpful. If the research is well done, and the answer to each of the three questions above is affirmative, then there is no reason not to publish the work. It should not be the job of the referees or editor to settle scientific disputes. As long as there is no error that can be identified in the work, let it be published and let those whose work is called into question do the necessary experiments to settle the matter. That's the best way to make progress.

Using explicit criteria will not eliminate the ability of an editor to select what to publish. Such other more subjective criteria as importance or timeliness can still be used. Explicit criteria will just cut down on rejections because of the controversial nature of someone's work.

Reviewer training. Do we really need to "train" reviewers? Of course we do. I doubt that anyone who has reviewed more than a few years would say that their early reviews were as good as their later ones. Even experienced reviewers find their approaches to papers changing with time. One not only learns what to look for in a paper, one also learns how to phrase a criticism so it does not seem like a personal attack on the author.

What can be done to train reviewers? First, journals can draw up more explicit and detailed instructions to reviewers than are presently sent out. (In neurobiology, my experience is that usually no instructions at all are sent.) A copy of the criteria or guidelines for what constitutes a good paper would be a start. Some explicit statement that reviewers should stick to objective descriptions of the paper, and not make derogatory comments about the author(s) might also help.

Another approach I have found useful as a reviewer is to receive copies of the remarks of other reviewer(s) after I have sent mine in. Cicchetti mentions this possibility. From the standpoint of curiosity, it would be interesting to know who the other reviewer was, but this is not really necessary. What is important is seeing a colleague's opinion of the paper, to see if you missed an important point, or for younger reviewers, just to see how someone else handles the entire review process.

It is not likely that the system of peer review will change any time soon. If we have to live with it, the least we can do is to organize it in such a way that we make sure we all play by the same agreed-on rules. Periodic evaluations of the peer review system, such as this target article, are important steps to this goal.

Different rates of agreement on acceptance and rejection: A statistical artifact?

Marilyn E. Demorest
Department of Psychology, University of Maryland Baltimore County, Catonsville, MD 21228
Electronic mail: *demorest@umbc.bitnet*

An important substantive finding that emerges from Cicchetti's target article is that reviewers of manuscripts and grant proposals appear to have higher rates of agreement on rejection/disapproval than on acceptance/approval. This conclusion is based on category-specific rates of agreement as shown in Tables 5 and 6: Given that one reviewer makes a particular recommendation, what percentage of the time does the second reviewer agree? The data clearly indicate higher percentage agreement for negative recommendations (70%–83%) than for positive ones (41%–60%).

A statistical interpretation of these findings is that the higher agreement rates on negative recommendations reflect their higher prevalence. The omnibus agreement statistics reported throughout the review (intraclass correlation and weighted or unweighted kappa) corrected for chance levels of agreement. (Indeed it is the adoption of chance agreement as the null model for evaluating observed agreement that makes the reliability of peer reviews appear so dismally low!) The same standard has not been applied in evaluating agreement on a category-by-category basis, however. When category-specific agreement rates are corrected for chance, they are shown to be identical for acceptance and rejection.

To illustrate, consider the data presented for the *Journal of Abnormal Psychology*. The reconstructed agreement matrix is

shown in Table 1, assuming equal marginal rates for the two reviewers. As shown in the table of expected frequencies, by chance there would be 35% agreement for acceptances and 65% agreement for rejections. If the observed agreement rates are corrected for chance using kappa, the result is identical values of .14, which coincide with the values of the omnibus kappa and the intraclass R reported by Cicchetti. A moment's reflection (and a little algebra) reveals that for the dichotomous case it could not be otherwise. Any disagreement is simultaneously a disagreement about acceptance and about rejection. Reviewers cannot in principle disagree at different rates for the two categories, once the chance level of agreement is taken into account.

Because there are additional degrees of freedom when three or more categories are analyzed, it is possible that differential agreement could have been identified had the data not been dichotomized. Data on submissions to the *American Psychologist*, presented by Whitehurst (1984) and analyzed by Cicchetti (1985, Table 4, p. 567), provide an example. Table 2 shows the five rating categories, their prevalence, percentages of observed agreement, chance agreement, and chance-corrected agreement. Of the three observed percentages, the highest is clearly for outright rejection. This category was used 50% of the time by the reviewers, however, and therefore its expected agreement is also high. The chance-corrected agreement rates are quite comparable for unconditional acceptance and unconditional rejection.

Although Cicchetti is careful to interpret his results cautiously and note limitations on generalization, he speculates that journals with higher acceptance rates might demonstrate the "reverse phenomenon," higher agreement rates for acceptance than rejection. Indeed, just such a prediction would be made from a consideration of chance levels of agreement if the base rate for acceptance exceeds 50%. If correction for chance is deemed appropriate when evaluating overall agreement among reviewers, it must surely be relevant when considering category-specific rates of agreement. Although reviewers *may* use different criteria and judgment processes for negative and positive evaluations, substantive interpretation of differences in agreement rates for acceptance and rejection should be suspended pending evidence that the differences exceed what would be expected by chance.

Table 1 (Demorest). *Reviewer agreement for the "Journal of Abnormal Psychology" (from Cicchetti et al. Table 5: Review Number 2)*

Observed Frequencies

		Reject	Accept	Total	Category % Agreement
Review No. 1	Accept	258	204	462	44
	Reject	599	258	857	70
	Total	857	462	1,319	

Expected Frequencies

		Reject	Accept	Total	Category % Agreement
Review No. 1	Accept	300	162	462	35
	Reject	557	300	857	65
	Total	857	462	1,319	

Note: Intraclass R= .14; kappa for agreement on acceptance = (.44 − .35)/(1 − .35) = .14; kappa for agreement on rejection = (.70 − .65)/(1 − .65) = .14.

Table 2 (Demorest). *Category-specific agreement (%) for American Psychologist submissions (from Cicchetti 1985)*

		Type of Agreement		
Category	Prevalence	Observed	Chance	Corrected for Chance
1	10.3	55.6	10.2	50.5
2	8.6	26.7	6.7	21.4
3	18.4	68.8	18.3	61.7
4	12.6	54.5	12.5	48.0
5	50.0	75.9	50.0	51.7

Note: Categories are defined as 1 = accept as is, 2 = accept with minor revisions, 3 = reject and recommend resubmission after revision, 4 = reject and recommend resubmission to another journal, 5 = reject.

APPENDIX

The kappa coefficient for category-specific agreement may be formed in the same manner as the conventional kappa statistic: $(p_o - p_c)/(1 - p_c)$. Letting n_{11}, n_{12}, n_{21}, and n_{22} represent the frequencies in a 2 × 2 agreement matrix, the observed agreement rate for category 1 is the number of agreements in category 1, divided by the average number of category 1 ratings made by the two observers (Cicchetti 1985):

$$p_{o(1)} = \frac{n_{11}}{[(n_{11} + n_{12}) + (n_{11} + n_{21})]/2} = 2n_{11}/(2n_{11} + n_{12} + n_{21}).$$

The number of agreements expected by chance for category 1 is $(n_{11} + n_{12})(n_{11} + n_{21})/(n_{11} + n_{12} + n_{21} + n_{22})$. Thus, the chance agreement rate for category 1 is:

$$p_{c(1)} = \frac{2(n_{11} + n_{12})(n_{11} + n_{21})}{(n_{11} + n_{12} + n_{21} + n_{22})(2n_{11} + n_{12} + n_{21})}$$

Substituting these values in the formula for kappa and simplifying yields:

$$K_1 = \frac{2(n_{11}n_{22} - n_{21}n_{12})}{2n_{11}n_{22} + n_{21}^2 + n_{12}^2 + (n_{11} + n_{22})(n_{21} + n_{12})}$$

The same result emerges if the subscripts are interchanged and K is calculated for category 2 rather than category 1. Thus for the dichotomous case, the category-specific agreement rates, K_1 and K_2, are identical.

When nonreliability of reviews indicates solid science

Douglas Lee Eckberg
Department of Sociology, Winthrop College, Rock Hill, SC 29733

Cicchetti begins by arguing that low inter rater reliability in reviews is a scientific "problem" that challenges science's claim of special knowledge. Should we accept this justification for their research? I believe not. I reject the idea that inter rater agreement has a relationship to validity of scientific knowledge at any but extremes of reliability. There are several interlocking reasons for this.

Decisions to accept or reject submissions seldom rest on whether or not findings are "true." One seldom sees patently silly submissions; on this "truth" basis almost all submissions could be accepted. But as Cicchetti points out, decisions are

based on numerous criteria. Is the work sufficiently new, complete, and logically consistent? Is there adequate awareness of the literature and few enough errors of calculation or citation? Does the author address conventions and disagreements on methods? Is the methodology strong enough to support research claims? Does it meet all of these criteria well enough to be accepted into an underfunded journal with limited space?

Especially where journals are relatively general, there are many grounds for disagreement, and for good reasons. Research does not follow a precise model. Individual reviewers cannot be well versed in all the techniques and literature on which they must pass judgment. Given these impossibilities, one relies on one's general disciplinary training and looks for evidence of the above criteria where one *can* pass judgment without looking foolish later (or one may simply withdraw from the review, which does not solve the *editor's* problem). Different reviewers will therefore see different points, and may draw different conclusions about "worth." As with judgments of athletic performences at high levels of complexity (e.g., Olympic ice dancing or diving), there may be little difference between the performances of winners and losers, but "everyone" may still distinguish careful work from that of hack writers. What is acceptable to a reviewer or editor is not a problem of agreement *per se*, but of journal economics.

As Cicchetti tells us (sect. 8), a great many "losers" ultimately *are* published. The cost of delay? Time (perhaps occasionally a career, but this is a different issue). The gains? Truly "bad" work is screened out. For other work, the increased feedback by peers can lead to the benefits of substantial reworking. (The author of the target article says this is infrequent, but he gives no hard figures.) It is hard to make a case that this harms science. Even most "winners" pass quickly from sight, with hardly the ripple of a citation or recognition beyond the tenure committee (e.g., Crane 1972; Merton 1973, p. 448). What difference can it make to the general progression of science *which* of two or more highly complex and generally-decent-but-not-earth-shattering works is published *first?* Science progresses because researchers are motivated to continue trying to publish, to continue to take part in the great discourse on nature, and very occasionally, to be brilliant.

I shall use Cicchetti's article as an example of reviewer problems. I would not have accepted the article for publication, but would have asked for revision and resubmission, along with a scathing note. Why? First, my weaknesses. I am not terribly familiar with the kappa statistic or the various models of intraclass correlation. In any case, the work that goes into constructing them is invisible, and I accept them at face value. I am also not up on the peer review literature, though I did take part in the discussion of Peters & Ceci's work (Eckberg 1982). I therefore cannot judge most of Cicchetti's citations. I am impressed by the *number* of citations, however.

I *can* judge some things. I can tell that the article begins by drawing on the literature in philosophy of science concerning the special validity of scientific knowledge, but that this is not thorough and is basically dropped later. I take umbrage at the fact that in Table 2 the author imposes on the reader evaluative criteria for strength of agreement (from poor to excellent), and that this seems to spring from the author's brow. I am bothered that in section 4.5 the author "would predict" certain findings in an area where no research has been done. This is pure speculation. I am *especially* bothered that the numbers in some tables simply do not appear to add up.

The numbers of reviews/manuscripts of *Journal of Abnormal Psychology* and *Journal of Personality and Social Psychology* differ in Tables 1 and 2. Why? R_i's in Tables 2 and 5 are different. Is this because of the dichotomization of data? Tables 5 and 6 purport to show differences in proportions of agreements between reviews, on whether to accept or reject manuscripts or give high or low scores to grant applications, using χ^2 as the test of significance. In the lower two rows of Table 5 and all rows of Table 6, one can reconstruct category frequencies. In the upper rows of Table 5 one can quickly produce all the possible sets of frequencies fitting the presented data.

There are some *real* problems here for this reviewer. [Note: This issue has been clarified by Cicchetti in sect. 1.4 of his response, Ed.] In ascending order of importance: (1) Why does the author (in sect. 4.7) believe the evidence shows a greater propensity to "agree" on rejections, when the results are interpreted more parsimoniously as showing that these reviewers simply reject more often than they accept? Chance overlap would yield the same patterns; (2) How does the author decide who the "two" reviewers will be, given that (at least in the case of the grant reviews; Table 4) about four reviewers is the norm? We are not told; (3) Why is it that in all cases where frequencies can be determined, n of disagreements is precisely the same in both the Acceptance and Rejection (or High Ratings and Low Ratings) columns? This should not be the case, if reviewers are selected randomly. It appears that the author merely divides splits equally by hand. Even so, why are there always even numbers of splits? (4) I find that I can reproduce *none* of the χ^2 scores on Tables 5 and 6, though I have tried several different techniques. Some of my scores are close to those the author provides, but none are precise and a few are off *substantially*. Either the author has been very sloppy with his calculations, in which case all their original data are suspect, or he is using conventions with which a given social scientist might be unfamiliar, in which case he should explain his usage.

Some of my problems with this article are mere quibbles; others concern differences in interpretation. The methodological issues may be cleared up by the author in his replies; certainly, either no one else noticed the χ^2 "problem," or they followed the same conventions as the author and so *had* no "problem." From this experienced reviewer's standpoint, there would have been sufficient questions on enough issues to have warranted sending the piece back to the author. The article is long and complex enough, however, for me to realize that another reviewer and I might disagree on the importance or significance of various points. We might even agree to disagree. This is the nature of professional judgment; it does not mean that science is in trouble.

Journal availability and the quality of published research

Jack M. Fletcher

Department of Pediatrics, University of Texas Medical School, Houston, TX 77030

In concluding his review of the reliability of various peer review systems, Cicchetti recommends a focus on the relationship of peer review and grant funding, fearing that the unreliability of peer review leads to a failure to fund worthwhile grant applications. This focus is certainly justified in the current climate of research funding, particularly since publication and funding mechanisms use peer review procedures that Cicchetti justifiably identifies as poorly understood and potentially unreliable. One recommendation is always that more funds should be made available to reduce the probability that important research was not funded. If the relationship between space availability in journals and the quality of published research were examined, it would become apparent that the availability of journal space does not ensure that most quality research is published. Unfortunately, more journal space also ensures that considerably more research of poor quality will be published (Lock 1985).

If the quality of funded research were also evaluated according to changes in the availability of funds, similar conclusions would be forthcoming. More emphasis should be placed on the goals and internal mechanisms of the journals and their pub-

lishers, as well as those of the agencies and individuals responsible for funding research. Mahoney (1985) identified a number of factors responsible for the surge in published research and number of journals, including employment practices, requirements for funding, and fraud. It is possible that such factors also lead to a tendency to downplay editorial practices and commercial needs to ensure the promotion and funding of research within the community of scientists.

The number of new journals in many areas of science is currently increasing rapidly (Broad 1988). This increase is accompanied by concerns about the expansion of the scientific literature and possible decline in overall quality. There is presently so much journal space that the only recourse left to many editors (and publishers) may be to include research that would not pass muster either from peer review or common sense. In this sense, the unreliability of peer review can be used as a basis for accepting manuscripts. Journals survive either through subscriptions (particularly to institutions) or through affiliation with an organized group that provides a ready pool of subscribers and other more direct sources of support, financially and through submissions. Journals with the latter sources of support can generally afford higher rejection rates and can use peer review in a manner likely to support many submissions and rejections. The research cited by Cicchetti, however, is based primarily on general journals with higher rejection rates. These journals may use a single negative review as a basis for rejection; other journals that need manuscripts (and subscribers) may use a single positive review as a basis for acceptance. While these phenomena have apparently not been adequately studied, it would seem important to begin to establish mechanisms for periodically reviewing the quality of various journals. It would be interesting to know rejection rates, number of solicited reviews, results of these reviews, and other principally empirical results of the peer review process across journals. Evaluations of quality by scientific polling of professional groups and subscribers would represent another source of information.

Without more pressure on journals, editors, and publishers, Cicchetti's implicit warning that the unreliability of peer review leads to a failure to publish good research on a timely basis is somewhat diluted. If virtually all research can be matched with a publication outlet, why should the scientific community worry about the unreliability of peer review? There is apparently some type of implicit evaluative system among authors that leads them to choose and rank journals for their submissions. It would seem important to begin to try to make this evaluative system more explicit. The failure to do so may represent a tendency among researchers to operate in ways that will keep journal outlets open for publication and promote the system criticized by Mahoney (1985).

The extent to which a similar problem influences funding mechanisms is not well understood and certainly not parallel to journal practices. It is simplistic, however, to lament the government's inability to fund high quality research on the basis of ever lower priority scores for unfunded research. Project officers are encouraged to solicit applications to demonstrate the need for more funds in their area to program directors and funding sources (i.e., Congress). Peer review study groups are generally informed as to the priority scores (or percentiles) necessary to ensure funding of the committee's highest rated proposals. There is a tendency to approve weakly lower quality research so that percentiles for better research will be improved.

More telling, however, is the question of what happens to research quality when more funds become available. In the past decade, the federal government has placed substantially more emphasis on the war on drugs and on the AIDS problem. I am not in any sense disagreeing with these priorities; both problems clearly represent national emergencies. The question I am raising is simply whether increases in funding lead to greater availability of quality research addressing these problems. As with increasing the number of journal outlets, we may be producing a glut of information, much of which will be of questionable significance.

One unintentional implication of Cicchetti's target article is that when peer review is unreliable, the only safe solution to scientific problems of national importance is to spend freely. The fact that mechanisms for evaluating the quality of funding agencies and of journals are generally ineffective, nonexistent, or self-serving should be of great concern to the scientific community, particularly if the absence of mechanisms represents in part the need for the community to publish, promote, and fund its membership. Current proposals for reform tend to focus on individual levels of responsibility. The unreliability of peer review extends beyond the peers. Additional focus on the responsible institutions (e.g., journals) would also seem warranted.

Peer review is not enough: Editors must work with librarians to ensure access to research

Steve Fuller

Science Studies Center, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061

Electronic mail: *fuller@vtvm2.bitnet*

I would like to propose that peer review systems in academia function like markets in society at large, and that the "rationality" (if you will) of such systems be evaluated in much the same way as markets are. It is clear from Cicchetti's target article that peer review is such a vexed issue because the variety of views on how and why the system ought to work does not match up neatly with the equally wide range of ways in which the system in fact does work. Do we therefore conclude that peer review does not promote the growth of knowledge, or that we have yet to fathom the "invisible hand" principle by which it does promote such growth? The author himself seems to be struck most by the variety of peer review practices across the disciplines, but ultimately they are noncommittal as to whether any one practice ought to be used as the model for all the disciplines. The closest I could find to an explicit, normative commitment in the target article was a concern (in sect. 8, para. 3) that good scholarship not be lost to the world because of selection standards that are more stringent than reliable. I would guess that, given a chance to reform the system, Cicchetti would try to get other disciplines to approximate the peer review practices of cross-disciplinary research fields, in which most submitted articles eventually get published somewhere. I would like to subject this easy liberalism to the cold scrutiny of market analysis, however.

Is peer review supposed to promote the growth of knowledge or the careers of scientists? It is not obvious that the two goals can be jointly maximized, though Cicchetti seems to presume that the fairer the system is to the individual scientist, the higher the quality of science that is likely to result. But why presume this? Here is why *not*. Peer review is only one of several selection mechanisms, or markets, that operate in the production and consumption of knowledge. For example, the differences that Cicchetti found between the rejection rates of specialities (low) and interdisciplinary fields (high) suggest that the specialists withheld submissions until they could anticipate acceptance, while interdisciplinarians failed to do this. This prior difference in self-selection is probably because of the specialists having been trained to write for certain target journals, whereas the interdisciplinarians were not. Moreover, the specialists learned to associate quality work with acceptance in those journals, whereas interdisciplinarians learned to be more flexible (or cynical?) in their journal aspirations. Thus, the

rejected interdisciplinarians resubmit and are accepted elsewhere, while the specialists tend not to. On the surface, it looks as though the interdisciplinary way of doing things ensures that innovative scholarship is not lost, as opposed to the way of specialties, which are more likely to stifle any innovative impulse at the very start.

That is not the end of the story, however. There is one more market to consider, namely, knowledge consumers (i.e., reading scientists) who must choose from among the variety of knowledge products the ones that are best suited to their cognitive needs. The interdisciplinary markets are flooded with more articles in more obscure journals than the specialty markets are. Since the cognitive limitations of the consumer remain fixed as the number of markets and products grow, *physical access* is becoming an increasingly important determinant of which research turns out to be influential. Is it published in a journal that I routinely peruse? Is the journal copy readily available in the library? Does the article appear indexed in many databanks? The answers to these questions depend on issues quite incidental to the intellectual merits of a given article: Can I afford the journal, and does it publish other articles I normally find interesting? Is the current periodicals section properly policed and updated? Does the title of the article contain words that make the right associations with other words in the databank? All the best laid plans to reform peer review will have been for naught, if the high quality journal that publishes the high quality scholarship turns out to be low on physical access.

My point, then, is that interdisciplinary research may give an illusory sense of preserving good scholarship simply because of its more liberal publication policies. This illusion is fostered by focusing on the editorial office as the only clearinghouse for knowledge products. It is not that interdisciplinarians do bad work, but rather that their work is so diffusely placed that access to such work, and hence its ultimate impact, is limited. Given the inaccessibility of some journals, the work might as well have never made it into print. This suggests some policy implications:

1. Editors should forge closer links to the library and information systems that will determine the access that potential consumers have to journals and books.

2. The goals of peer review should be oriented more to the interests of a given journal's readership. At the moment, when there is a conflict of aims, peer review aims more at publishing papers that exemplify the methodological standards of the journal's discipline than papers that are likely to be taken up by the readership in their own research.

3. Regardless of whether one thinks that more scientists make for better science, the growing number of paper submissions may, at some point in the future, have to be checked by requiring that authors restrict the number of papers they publish over a given period. (In fact, Donald Campbell has suggested on occasion that the use authors put to such self-restraint could be weighed in tenure and promotion decisions.)

4. Tighter control of the knowledge markets, including increased self-selection of paper submissions, could encourage specialization and rigidify disciplinary boundaries. Much depends here on whether journal editorial policies are dictated more by the character of the field of study, or whether the field of study comes to have its character in virtue of the journal editorial policies. This is one of many intriguing issues that Cicchetti leaves hanging.

On forecasting validity and finessing reliability

J. Barnard Gilmore
Department of Psychology, University of Toronto, Toronto, Ontario, Canada, M5S 1A1
Electronic mail: *gilmore@psych.utoronto.ca*

Reliability is of such great concern in judgments of scientific worth because validity is. Where there is no reliability, there can be no validity. Even where reliability is found, there could be, and often there appears to be, distressingly low validity. The issue facing both peer reviewers and those who engage them is whether or not validity had been achieved with a given set of ratings.

These truths are familiar. They have been brought home to us many times in the past, as in the thoughtful work by Gottfredson (1978) and Mahoney (1985). Moreover, these are brutal truths, brutal, because achieving validity appears to be beyond all reasonable hope insofar as predicting the eventual importance of scientific work would require one to predict an unpredictable future. Let us have no illusions. The cherished arguments for generously supporting pure research, the arguments concerning the unpredictable sources of new scientific understanding, are the same arguments for doubting that we can forecast in advance which work needs financing and which work needs publication. Still, yes, choose we must. But nothing requires us to assert that we will have chosen well. And, nothing forces the conclusion that it would be the least bit foolish to make many of our choices by drawing lots.

Consequently, the concern with improving the reliability of potentially invalid ratings made by multiple reviewers, and with improving the measures of whatever reliability we do have, must not be overemphasized. Cicchetti asserts, for example, that some statistics are appropriate for measuring reliability and some are not. Instead, one might assert that the appropriateness or the lack thereof is more often to be found in the *meaning* ascribed to the statistics rather than to the choice of the statistics themselves. A careful reading of Garner & McGill (1956) makes it clear that some important differences in meaning and interpretation are appropriate to measures that are variance-based, such as kappa, versus measures derived from information theory, which reflect the proportional shared uncertainty (measured in bits) among raters. The "reliability" reflected in an uncertainty statistic reflects the proportion of our total uncertainty about the judgment of another rater, which will be reduced by the information contained in knowledge of an earlier rating. I would submit that this shared uncertainty index is closer to what we always intended to mean by "agreement" than is that percentage-of-total-variance index implicit in most kappas.

There are sound metric reasons, too, for sometimes preferring the uncertainty statistic to kappa. Garner and McGill remind us that variance-based statistics require an interval scale substrate to justify many of their interpretations, whereas uncertainty measures are always metric free, generalizable, and mutually comparable. The significant and sad fact is that most reliabilities measured with the shared uncertainty statistic turn out to be "lower" in relative size than those reported using kappas. (For a clear example of this, see Gilmore 1979.) Thus, the data presented in the target article may well deserve an interpretation that is even less optimistic than those marginally optimistic interpretations offered there.

In the conclusion of the target article the authors note that one of two assumptions (see Harnad 1986) prevails. One may assume that most published research contributes to the advancement of scientific work, in which case the rejection by journals of what would otherwise have proven to be helpful new data or new perspectives is indeed a serious matter. Conversely, one may assume that most published research does not contribute to the

advancement of science, in which case it seems far less serious that "good" articles would be rejected by unreliable reviews. I submit that the latter view only makes sense if one also endorses two unlikely additional assumptions: (1) that rejecting some of the few "good" articles along with the many "bad" ones will still reduce the absolute amount of false leads chased down, while not unduly delaying the eventual resubmission and publication (presumably by someone else) of the previously rejected, good idea; and (2), that researchers who cannot get funded or published will neither be missed nor unduly hurt when they then leave the society of scientists.

Pending the arrival of a better "science" of science, a better psychology of science, even a better political science of science, one can only observe that, in addition to all of its personal satisfactions, science is a complex social activity offering diverse social rewards of great value and diverse punishments of unsuspected force. The true sources of the most fruitful ideas and work in science remain mysterious. Scientists differ in temperament, in their values, and in their skills at communication. There is no one way in science. There are no 10 ways. Scientists have embraced electronic mail and have welcomed multiple new journals despite the undeniable difficulties and frustrations created by the information glut. One fact seems clear, then: Most of us will always want to know what others are thinking, doing, and finding. And whatever journal we are reading, whoever its reviewers or gatekeepers may be, still, we decide for ourselves what is flawed design, what is misleading interpretation, and what is "good" science.

As electronic means for sharing papers and creating reprints evolve further, I submit that it makes sense (every way except politically, which is, alas, perhaps the most significant way) to separate reviewing, editing, and publication. Let us have multiple professional reviewers, some of whom advise authors prior to publication and some of whom rate articles once they appear in print. But let us normally publish, electronically, all submitted papers when the author, as editor, thinks each is ready. Let us include with each paper careful abstracts, conscientious keyword lists, and brief professional reviewer ratings whenever available, to serve us as guides through the great wilderness of articles we might all then fear. Let there be a gatekeeper to electronic publication to keep out undue repetition from authors, and, when necessary, to enforce basic conventions of style and some reasonable quotas on how often one may publish in the system. Then, let authors vie not for space in prestigious journals, but for the attention of prestigious reviewers and readers. History and citations can later tell us what was useful and what was not. The problem of reliability would then be all but finessed.

Replication, reliability and peer review: A case study

Michael E. Gorman

Humanities Division, School of Engineering and Applied Science, University of Virginia, Charlottesville, VA 22903
Electronic mail: *meg3c@prime.acc.virginia.edu*

Cicchetti begins his paper with a brief discussion of Mertonian norms. Their concern is to see whether these norms are being followed by scientific journals. An issue he leaves for others is the question of how to evaluate empirically the efficacy of such norms. For example, Cicchetti discusses the way in which journals are biased against negative findings. Experimental simulations of scientific reasoning, in which science and engineering students work on abstract tasks, provide independent support for the value of seeking negative results, and also specify under what conditions a disconfirmatory strategy will be most useful (see Gorman & Gorman 1984; Klayman & Ha 1987).

Similarly, Cicchetti points out that there appears to be a very strong bias against replication studies. In a series of experiments I found that a strategy I called "replication-plus-extension" was superior to straight replication (Gorman 1989). Consider, for example, a student who wants to make sure that the triple "2,4,6" is really an instance of an abstract rule, given that the student knows as many as 1 triple in 5 will be subject to what Doherty and Tweney (1988) have called "system-failure error," that is, if it appears to be correct, it will actually be classified as incorrect and vice-versa. This student could propose "2, 4, 6" again. But what if the cost of an additional experiment is high? Then it makes more sense to propose a similar triple, "10, 12, 14" for example, which will not only replicate the previous one but extend the pattern to new instances. From a logical standpoint, this is a flawed strategy: "10, 12, 14" does not really replicate "2, 4, 6." But from a satisficing standpoint (see Giere 1988), the strategy makes good sense. In fact, students working on more complex tasks of this sort can employ it effectively (again, see Gorman 1989).

Cicchetti points out that one strategy for getting around journals' bias against replications is to embed the replication within another study or studies. Replication-plus-extension is an alternate strategy. Obviously, the two can be combined. I am aware of no empirical data regarding journals' preference for either strategy; future research should be directed at this question. For example, one could investigate replication-plus-extension through experimental or quasi-experimental designs by sending three versions of the same results to a wide range of journals, one of which was deliberately written as a replication, another as a replication-plus-extension and still another embedded with a novel finding.

Such reliability studies as Cicchetti's are important, but they should be complemented by two additional kinds of research: (1) Experimental studies directed at determining the normative value of philosophical or sociological prescriptions about science (see Fuller 1989, for a discussion). (2) Qualitative "biographies" of manuscripts, in which the same paper is followed through the revision and publication process, often spanning several journals. Cicchetti disparages these sorts of studies, yet they can reveal aspects of the peer-review process inaccessible to quantitative studies and suggest variables like replication-plus-extension that merit more rigorous exploration in quantitative designs.

Is there an alternative to peer review?

Richard Greene

Veterans Affairs Central Office, Washington, DC 20420

Cicchetti's target article is a major contribution to the peer-review literature. It is especially useful in its collection and analysis of the critical research studies in this field and raises a number of important criticisms of the peer-review process. I will concentrate my comments on the grant-review process, because this relates to my experience managing the national research program of the Department of Veterans Affairs (DVA).

Cicchetti refers to a number of studies showing that reliability among peer reviewers is highest when considering what is unworthy of support. The real problem comes in assigning a scientific priority to a set of studies that are all, or mostly all, supportable. The problem is well described by the author's citation of James B. Wyngaarden (Culliton 1984) referring to distinguishing "'shades of excellence' among competing grants that are all at the top." There is a consensus in the scientific community that the peer review process was not designed to measure the difference between two highly meritorious projects, one with a 155 priority score that will be funded and one with a 156 score that will not receive support. And yet, the

research administrator does run out of funds at 155. Should we abandon the current system, or shore it up for the present, hoping for a better time when increased research budgets will at least allow more approved projects to be funded? The reliability question would not go away, but fewer people with high quality projects would go unfunded and the problem would appear less pressing. Cicchetti does not advocate abandoning peer review, but he does offer several recommendations to improve the system. Let us consider them.

(1) "To improve the reliability of peer review, a minimum of three independent referees has been recommended." This is a very useful recommendation, and one that the research program in the Department of Veterans Affairs has used successfully for the last five years. DVA Medical Research Service uses four independent reviewers for each submitted research proposal in its Merit Review Program (equivalent to the NIH RO1 Program). Two written independent reviews are submitted by scientists who are not members of the Merit Review Board. These reviewers are selected because their own work is closely related to the applicant's problems. Two additional written reviews are prepared by the primary- and secondary-reviewer members of the Merit Review Board. Thus, when considering the application, members of the board can debate the relative importance of each critique and weigh the specific evaluations of four independent reviewers in light of this debate. The strength of this process goes beyond the number of reviews per proposal, to the discussion and analysis of the four independent critiques that take place during the review session, leading to a consensus evaluation by board members. This process still cannot accurately fine-tune shades of excellence, but it is a practical method of weighing the opinions of four independent reviewers.

(2) "Using author anonymity or 'blind' review'." This recommendation is impractical for the grant review process. Since past scientific productivity is a critical element of grant review, it would be extremely awkward to try to disguise the authorship of the proposal as well as the authorship of published papers from previous funding or training periods. Author anonymity would, in my opinion, significantly weaken the grant review process.

(3) "Revealing reviewer identity." Cicchetti comes down on the side of encouraging reviewers to reveal their identity voluntarily. Our experience in the grant review process is that anonymity is critical. I believe it is important to avoid personalization of the review process. It should be emphasized that the evaluation of a project is a consensus opinion of a committee, the membership of which is public knowledge. The strength of the process is that of evaluation by a *group* of experts.

(4) "Author review of referees." This recommendation is attractive and could in fact help weed out inappropriate reviewers. In addition to author reviews of referees, members of peer review committees could also identify reviewers who are problematic. Consideration will be given to implementing this idea in the DVA's scientific review process. The DVA's system has already adopted the practice of encouraging applicants to list potential reviewers and those they do not wish to review their proposals.

(5) "Developing a peer review appeals system for grant submission." This should be a critical aspect of any grant review process. The peer review system is a human enterprise and thus, not perfect. There must be a mechanism for applicants to appeal the results. The DVA system has had an effective appeals procedure for over a decade. Appeal has proved to be a complex and sensitive area and we have found it necessary to revisit the ground rules for appeals periodically.

Cicchetti has raised some important concerns about the peer-review system and has made some useful recommendations for improving it. While much of the stress in the current grant review process is a function of the small percentage of high quality research projects that can be funded, efforts to analyze and strengthen peer review are to be applauded.

Referee agreement in context

Lowell L. Hargens
Department of Sociology, University of Illinois, Urbana, IL 61801
Electronic mail: *hargens@uiucvmd.bitnet*

Cicchetti provides a valuable summary of the procedures and results of studies of interreferee agreement in peer review. Many will be surprised by the generally modest associations between referee recommendations, with most studies yielding intraclass correlations in a range between .20 and .35. Cicchetti characterizes these levels of agreement as "poor," and others have claimed that they indicate that chance plays an important, if not dominant, role in the assessment of scholarship (Cole et al. 1981; Lindsey 1988; Mahoney 1976). These modest associations should be viewed in the context of the entire peer review process, however; failure to do so gives a misleadingly pessimistic impression of the value of referees' assessments. Below I focus on referees' evaluations of manuscripts submitted to scholarly journals, but my arguments hold in general for peer review of grant proposals, too.

Those who argue that the modest associations between referees' evaluations imply that referee evaluations have low reliability usually base their interpretation on a psychometric perspective. This perspective, often called "classical test theory," views different referees' recommendations as parallel measures of a latent trait (see Lord & Novick, 1968); usually the trait is seen as the "scholarly quality" of a manuscript. Editors' discussions of the strategies they use in selecting referees cast doubt on the appropriateness of this perspective, however. They report frequently choosing referees they think will be sensitive to different aspects of a manuscript, perhaps one to judge its analytic procedures and another, its substantive contribution (Campbell 1982; Roediger 1987). If these different aspects are only moderately correlated across manuscripts, referees' assessments should show low agreement. In addition, in some fields scholars belong to competing "schools," and editors sometimes intentionally solicit evaluations from members of both sides of a controversy (Hull 1988). If an editor always followed this strategy for controversial submissions and a large proportion of submissions were controversial, referees evaluations might be *negatively* correlated. Thus, referees' evaluations are often not parallel measures of a latent unidimensional trait, and the low observed associations do not necessarily imply that peer-review evaluations are unreliable.

Editors' summary rejection of submissions may also produce low associations between referees' recommendations. Papers that are not sent out for review are necessarily omitted from referee agreement studies. If editors are right in claiming that summarily rejected papers are of very poor quality or are obviously inappropriate for the journals to which they have been submitted, then screening out such submissions reduces the range of papers evaluated by referees, and hence the association between referees' evaluations. Thus, reported associations between referee evaluations cannot be assessed without also considering journals' summary rejection rates. Fragmentary data on this question indicate that summary-rejection rates for prestigious social science and medical journals can be as high as 50% (Gordon 1978; Zuckerman & Merton 1971).

Even if referees' recommendations were parallel measures of manuscript "quality" and editors never rejected papers summarily, the modest levels of referee agreement summarized by Cicchetti should not be taken as an indication of the reliability of the entire review process. Under the assumptions of classical test theory, referee-recommendation intraclass-correlation coefficients estimate the reliability of the average *individual* referee's evaluation (Tinsley & Weiss 1975); the reliability of an assessment based upon two or three referee evaluations should be considerably higher. (See also Cronbach, 1981, who noted

that Cole et al. [1981] exaggerated the role of chance in the NSF grant-review process by basing their analysis on the unrealistic assumption that the reliability of the entire NSF peer-review process equalled that of a single referee's evaluation). For example, under classical-test theory assumptions, if the average referee's reliability equals only .25, a composite formed from 2 evaluations should have a reliability of .40, and a composite based on 3 should have a reliability of .50 (Hargens & Herting 1990b).

For most biomedical and behavioral science journals, editors reject papers that receive two negative evaluations, solicit a third evaluation for papers that receive split reviews, and personally read papers that receive two favorable reviews in order to advise authors about what, if any, revisions should be made before the paper will be accepted. Thus, most papers that eventually appear in these journals receive at least three evaluations; those receiving split reviews from the initial referees are sometimes reviewed by four or five people before they are accepted. Only papers that receive negative evaluations from both initial referees are evaluated by only two people, but, as Cicchetti shows, negative evaluations are substantially more reliable than other evaluations at these journals. Thus, the overall reliability of the peer review process as a whole should be significantly higher than the levels of individual referee reliability implied by the reported associations between referee recommendations.

I think it likely that Cicchetti will prove right in predicting that studies of referee agreement in general physical science journals, such as *Physical Review Letters*, will yield associations similar to those in the medical and behavioral sciences. I also suspect, however, that his speculation that specialized physical science journals will show greater agreement on positive compared to negative recommendation categories is incorrect. The data in Cicchetti's Table 6 and data on referee agreement for *Physiological Zoology* (Hargens & Herting 1990a) are inconsistent with that claim. If my suspicion is correct, even specialized physical science journals are likely to show modest levels of referee agreement. In part, this should happen because referee reports on submissions to such journals contain a relatively low proportion of negative evaluations (Hargens 1990), which will in turn tend to limit the overall reliability of individual referees' evaluations, because there are relatively few of the most reliable recommendations. Once again, however, the peer-review process at such journals tends to mitigate the damage that low associations between referee recommendations might cause. Specifically, these journals usually use a "when in doubt, accept" decision rule (Zuckerman & Merton 1971), and require at least two negative recommendations to reject a paper. As a consequence, these journals typically publish a substantial majority of submissions. If it is true that negative recommendations are more reliable than positive recommendations at these journals, then final editorial decisions will be largely determined by the most reliable (negative) recommendations rather than by less reliable (but more positive) ones. Regardless of the fate of these speculations, however, it is clear that the various elements of a peer-review system are interrelated, and that an assessment of any one element must place it in the context of the entire system.

Confusion between reviewer reliability and wise editorial and funding decisions

Charles A. Kiesler
Psychology Department, Vanderbilt University, Nashville, TN 37240
Electronic mail: *kiesler1@vuctrvax.bitnet*

Reviewing manuscripts for publication and grant proposals for funding is merely a means to an end, not an end in itself. The desired end product should be wise decisions about what is published and funded. Defining the reliability of such reviews as the correlation between reviewer ratings confuses process with outcome, and this is Cicchetti's main problem. In addition, there are important differences between the reliability of grant reviews and the reliability of article reviews; differences sufficiently important that I shall tackle them separately. Let me take up the issue of journals first.

A high correlation between reviewer ratings of submitted manuscripts should neither be expected nor desired. The expectation that these ratings should be highly correlated is naive; it almost assumes that reviewers are randomly drawn by the editor. As an editor, I intentionally act in ways that lower the correlation between ratings. For example, I give the manuscript to reviewers who have very different strengths or skills to bring to the manuscript: One might be a very sophisticated statistician, another a Freudian theorist. One would not expect a high correlation between them because they are evaluating different aspects of the manuscript; a valuable service to the editor. Sometimes, I also give manuscripts to two reviewers who I know will represent quite different points of view. I might select one reviewer who I know agrees with the general theoretical orientation and another who argues strongly against it. In this manner, I can see at one time both the very best and the very worst things one could say about the manuscript, and therefore make some judgment about how different and innovative it is. Furthermore, since I think a scientific field should develop an expanding pool of educated reviewers, I often give a manuscript to two sophisticated and experienced reviewers, and to another person (usually young) whom I have not consulted before. In this way I can discover good new reviewers, as well as show young people what is expected in a review. (I return all the reviews to their authors, so they can see each other's comments.)

All of these actions, which I submit contribute to making wise decisions about whether or not one accepts a manuscript for publication, are certainly counterproductive if one is seeking a high correlation between reviewers' judgments. But they are typical behaviors of a good editor who intends to play an active and decisive role in the final evaluation of a manuscript.

This notion of the editor having a very active role in the judgment of a manuscript seems lost on Cicchetti. Not only does he not discuss the kinds of processes described above, but he even ignores the role of the editor in making the judgment. For example, he recommends that there be three reviewers rather than two, to avoid a one-to-one vote. In my own case, when sending a manuscript out for review, I try to read just enough to make the judgment about whom to select as a reviewer. Then, when the reviews come back to me, I set them aside and review the manuscript myself with reference to the reviews. That's why my reviewers are always N + 1, and I can easily compare and contrast what the reviewers contribute.

The proportion of manuscripts one can accept is also a critical part of any investigation of reliability. I had just finished a stint as associate editor when Scott (1974) came out with his original article criticizing the low reliability of reviewers for *the Journal of Personality and Social Psychology* (*JPSP*) and I noted an interesting phenomenon. In only about 15% of the cases did reviewers of manuscripts for *JPSP* agree that the manuscript should definitely be published. Cicchetti would regard that as an unreliable review process. Only about 15% of the manuscripts *could* be accepted for publication in *JPSP*, however. If the review process is supposed to lead to a wise decision rather than producing a high correlation between ratings, the reliability of the *JPSP* process was very good. The outcomes were right in line with the needs of the editor to publish only a small subset of the manuscripts submitted. Whether that is the 15% to be published depends on whether or not editors see themselves as playing a very active role in the process. I think the

editor must play a very active role; Cicchetti apparently does not.

In general, the proportion of manuscripts (or grant submissions) that one can accept has an important influence on issues that concerned Cicchetti. When evaluating a submission to a journal (or a grant proposal), most reviewers are quite aware of the percentage of manuscripts (or grants) that can be approved. This awareness has a significant influence on the percentage of submissions that they rate as excellent, very good, and so on. Currently at the National Institutes of Mental Health (NIMH) grant submissions need to be close to a rating score of 125 (meaning an average rating of 1.25 on a 1–5 scale) to be funded. A rating of two on the NIMH scale reads "very good," but the experienced reviewer knows very well that a vote of two on a five-point scale for NIMH is a vote not to fund. Hence, reviewers even interpret descriptive labels on a scale differently depending on the percentage of potentially successful applicants.

The proportion of successful applicants is an important influence on the differences Cicchetti observed between the sciences. In general, the natural sciences and the behavioral sciences differ on this important statistic; the natural sciences typically have a higher acceptance rate on both grants and manuscript submissions. I argue that the differences that Cicchetti observed between these groups – what they specifically referred to as differences in emphasis on acceptance versus rejection – are a function of this difference in the probability of success. I would argue that if one were to equate the behavioral and the natural sciences for probability of success, whether regarding article or grant submissions, one would no longer find the differences observed by Cicchetti.

Cicchetti also seems confused about the role of biases in the judgmental process. The reputation of the investigator, the quality of the institution the investigator works for, and prior work by the investigator all influence the judgment that a reviewer might make. What Cicchetti seems to miss is that all of these biases artificially inflate the kind of "rating reliability" they emphasize. Cicchetti seems to imply that biases decrease reliability. In the sense that I mean the term they probably do, but such biases would increase the simple numerical correlation between reviewers' ratings that Cicchetti is concerned with.

Most of the issues described above apply to grants as well as manuscripts. There are some significant differences, however, that are worth noting. The potential impact of a delay is different for a grant than for a manuscript. As Cicchetti notes, 80 to 90% of manuscripts rejected by the journals to which they are submitted ultimately get published elsewhere. Having an article rejected by one journal may only mean a delay in publication of two to four months. A delay of four to six months necessitated by a grant resubmission (which would not be unusual and may be minimal) may force an investigator to shut down a research team that had been carefully built up over a period of years. In that sense, it is especially important that we focus on making wise judgments on grant reviews and give them an importance greater than manuscript review. For example, as rating scores inflate for grants, a "blackball" becomes a critical problem. If an agency is required to average all ratings of a proposal in the decision to fund, and if a score very near 100 is necessary to be funded, then a single reviewer can blackball a grant proposal by giving it a four or five (recognizing simultaneously that a score of two is a recommendation not to fund). This is a problem particularly for controversial, new, or innovative research.

The potential for blackballing grant proposals is a critical variable for protecting advances in science. We need greater flexibility for granting agencies, the sequestering of funds for especially innovative or new ideas, and a loosening of the requirement that all ratings of grant proposals be counted. Regarding the last, one might either report the median rather than the mean, or only average the N − 1 best ratings. That is, one might throw out the worst rating of any grant proposal and average the remaining ones.

Making wise decisions about publishing articles and funding grants is critical for normal progress in the sciences. Viewing the reliability of reviewers' judgments as simply correlations between ratings is to miss the most important part of that judgmental process. What is most important is that the outcome of the editorial decision or the agency's funding decision be a wise one, one that facilitates the development of our sciences. In no way does a wise decision depend upon a high correlation between the ratings of reviewers.

Do we really want more "reliable" reviewers?

Helena Chmura Kraemer
Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305
Electronic mail: *mn.kra@forsythe.stanford.edu*

First of all, congratulations to Cicchetti for his excellent target article. This paper represents a comprehensive, stimulating, and provocative discussion of issues that not only profoundly affect our individual professional lives, but the quality, consistency, and rapidity of progress within our respective fields. It is particularly interesting to read this paper from the perspective of the various roles each of us is asked to play in our professional lives: as author and reviewer of papers and proposals, as well as "consumer" of the results of published papers.

It is difficult, perhaps impossible, to be objective about one's own work, as a researcher, a reviewer, a "consumer," or an editor. The standards one might apply to a review of the review process are fundamentally different from these perspectives. Accordingly, the major contribution of Cicchetti (and others whom he cites) is the objective, unemotional, and quantitative approach to these issues. Only with such an approach is there hope of identifying or correcting faults in the review process. I doubt that I was invited to comment because of any such perspectives on the problem, more probably it was because of my research on the design and analysis of reliability studies and on kappa and intraclass correlation coefficients. I will briskly discharge my duties with regard to purely statistical issues and move on to more interesting themes.

I have a few points of disagreement on what was done: the choice of forms of coefficients, the use of null tests, the use of point rather than confidence interval estimates, and the use of asymptotic approximations to distributions rather than Jackknife or Bootstrap methods (Block & Kraemer 1989). If the authors and I were required to resolve such issues, I would predict we would happily reach solutions agreeable to us all, and that any resulting changes would scarcely affect the messages delivered in this paper. A kappa of .3 might become .2, or vice versa. A wide confidence interval might lessen interest in one reported study, whereas a short one might highlight another.

Instead, what requires reconsideration is not the magnitude of the reported reliabilities, but what to make of them. Difficulties arise because the word "reliability" is misleading when used with a general audience likely to interpret it in the sense of "to be trusted." A "valid" measure is one "to be trusted." A "reliable" measure may only be a highly reproducible wrong answer. Two facts about reliability are well known: (1) One may have a prefectly reliable (precise) measure that totally lacks validity (accuracy), and (2) one may improve reliability (precision) at the cost of validity (accuracy). Whether we err in judging the review process by assessment of interreviewer reliability is not therefore a trivial question.

My impression is that editors frequently seek reviewers with *different* expertise related to the various areas pertinent to the

submission. Reviewers are not selected to *reproduce* each other's results, but to *supplement and complement* each other. This is particularly true for submissions that are interdisciplinary. For example, a study reporting the results of a randomized clinical trial investigating the efficacy of an educational intervention for low birth weight, premature infants for enhancing cognitive development in the first three years of life, might well require review by a biostatistician, a neonatalogist, a pediatrician, an educator, and a psychologist, to fully and fairly assess its quality. By soliciting reviews from such varied professional fields, editors are acting in a way that might minimize reproducibility of the results. Are they wrong?

I think not. A simple-minded illustration: Suppose x_i represents the true scientific quality of submission i (sampled from those sent to a particular journal or agency), and x_{ij}, the assessment of reviewer j (sampled from competent reviewers) of submission i where:

$$X_{ij} = x_i + e_{ij}, \tag{1}$$

with e_{ij} representing the error of reviewer j's evaluation of submission i, that portion of the reviewer's assessment that is independent of the quality of the submission (which includes bias and other such errors, not all of which are random).

The *validity* of a reviewers' assessments for the scientific quality may be assessed by:

$$\text{Correlation } (X_{ij}, x_i) = (r)^{1/2},$$

where

$$r = \text{Variance}(x_i)/[\text{Variance}(x_i) + \text{Variance}(e_{ij})], \tag{3}$$

and the reliability between reviewers by:

$$\text{Correlation}(X_{ij}, X_{ik}) = r+(1-r)t,\ j \neq k, \tag{4}$$

where t is the correlation coefficient between reviewers' errors made for a submission. If the errors are completely independent, then the reliability and the validity are closely related (r and $r^{1/2}$). At the other extreme, if the errors are perfectly correlated ($t=1$), the reliability may be perfect, but the validity may well be near zero.

Now suppose we were to select a panel of R qualified reviewers and to use their mean as the assessment of quality. Then the validity of this mean as a measure of true quality would be:

$$\text{Correlation } (\bar{X}_i, x_i) = [Rr/\{Rr+(1-r)(1+(R-1)t)\}]^{1/2}. \tag{5}$$

One can see (Equation 5) that if $t = 1$, the reliability may be perfect (Equation 4), but there is no improvement in validity gained by soliciting more than one reviewer, and that validity may be near zero. The lower the correlation of errors (t), the lower the reliability may be (Equation 4), but if it is nonzero ($r > 0$), the greater the increase in validity may be with each additional reviewer (Equation 5). Maximal validity is obtained when the errors are independent, *and* one has as many reasonably reliable reviewers as possible (cf sects. 7.1,7.2).

Consequently, if editors do indeed deliberately select multiple reviewers to cover the various professional fields relevant to a submission, they are thereby probably *minimizing* the correlation of errors, thus *maximizing* the validity of the overall assessment, but thereby possibly *decreasing* the interreviewer reliability as well. With that in mind, do we really want more reliable reviewers? Perhaps those reliabilities reported in the .2–.4 range are of no concern. Not so, for the goal is to improve the *validity* of the review *process*. To the extent that this goal can be achieved by improving the *reliability* of *individual* reviewers, yes, we do want more reliable reviewers. The strategies discussed in section 7 should, however, be assessed and amplified with specific strategic goals in mind:

(1) To increase reliability by increasing the sensitivity of reviewers to the differential quality of submissions (increase Variance x_i)). For example, both reviewers who commend everything (sect. 7.6) and those who condemn everything should be removed from the review process. Reviewers with "blind spots" should excuse themselves from reviews in that area. My own "blind spots" include applications of Lisrel models. I have yet to see one whose scientific quality I have not questioned. Respected colleagues may have no problems in this area but might have troubles with meta analyses or quasi-experimental or observational studies (sect. 3.1), areas in which I believe I am able to distinguish "good" from "bad."

(2) To increase reliability by decreasing reliance on factors irrelevant to the quality of the submissions (decrease Variance (e_{ij})). Double-blinding, despite its weaknesses, remains the prime strategy here. Thus I agree with section 7.3 and disagree with section 7.4. Any reviewers can choose to reveal their identities to the author at any time. It need not be made a formal part of the review process. Use of multiple reviewers (sect. 7.2) also serves this purpose. Finally, in journal review, one might add strategies already common in grant review. No editor or reviewer from the same institution as the submittors should participate in the review process. Reviewers or editors should excuse themselves from the review of submissions of close personal friends or frequent professional collaborators, or from any other situation in which there might be an appearance of a conflict of interest.

(3) To increase validity but to decrease apparent reliability, by decreasing the correlation of errors (decrease t). No two reviewers from the same institution or who are close collaborators should review the same submission. Effort should be made to select reviewers across the broadest possible spectrum of specialties pertinent to the submission.

Ultimately, however, strategies to improve the review process focused on individual reviewers are not, I think, likely to optimize it. Again, let me propose a simple-minded illustration: Classify submissions as either Flawed or Nonflawed, and characterize the review process as in Table 1. There are two possible errors, that of accepting a flawed submission (impaired sensitivity to flaws) and rejecting a nonflawed submission (impaired specificity to flaws). I would argue that the Type I error, that is, the more serious error, is that of accepting a flawed paper, for such papers can mislead an entire field and may delay or derail progress. If the flaw is later detected and revealed, such a paper is an embarrassment to the authors, as well as to those who recommended acceptance. For a flawed grant proposal, time and money are wasted that might have been better invested elsewhere. On the other hand, rejecting a nonflawed submission (Type II error, I propose) frequently means only a delay, an annoyance to the submittor. In the long run, many such papers are published elsewhere (sect. 8); many such proposals are resubmitted and funded later.

The kappa coefficient used here (the so-called "unweighted" form) places *equal* weight on the Type I and II errors, whereas, on the basis of the argument above, I would prefer a form that

Table 1 (Kraemer). *A model for the evaluation of the probabilities describing the review process.*

	Decision of Review Process		
	Accepted	Rejected	Submissions
Flawed	P(SE)	P(1-SE)	P
Nonflawed	(1-P)(1-SP)	(1-P)SP	(1-P)
	Q	(1-Q)	1

P is determined by submissions to the journal or agency. Q is determined by resources of the journal (space) or agency (funding). SE represents the sensitivity of the review process to flaws in submissions, SP, the specificity.

places maximal weight on avoiding the Type I error. How to estimate such a kappa from reliability data may not be known, but the greater agreement reported among reviewers for rejection than for acceptance gives some hope that the review process may be operating better than indicated by the unweighted kappa in avoiding Type I errors.

Be that as it may, some of the strategies proposed (sect. 7.7–.9) are directed not at improving validity or reliability per se, but at reducing what is here labelled Type II error. I agree that Type II error should be reduced, but *not* at the cost of increasing Type I error. Well-done reviews leading to rejection may be beneficial in the long run. If a fatal flaw is detected, it prevents embarrassment, as one is allowed to withdraw quietly. If a flaw is remediable, the authors have the opportunity to revise the paper to one of substantially higher quality than the original. In my view, it is the editor's responsibility (not the authors') to detect and ignore poorly done reviews, or to ignore the occasional weak points in otherwise well-done reviews.

In place of the appeals processes Cicchetti suggests (sects. 7.8, 7.9), let me propose an *external* quality control panel. For each published paper, the names of the authors and those of the reviewers and editors and their recommendations would be filed with this group. This group would then receive and compile all challenges to the scientific validity of the results reported in the Abstract of the paper (i.e., ignoring typos or minor errors). A few such challenges now appear as letters to the editors or as papers submitted to the same or other journals, but are subject to the review of the same editors, reviewers, and sometimes the submittors, who may have erred in the first place. If enough evidence accumulates in these challenges to indicate a major flaw, one sufficient to raise questions about the validity of the overall conclusions, the journal should publish a summary of the challenges compiled by the quality control panel, along with the names of the authors, and those of the editors and reviewers who recommended publication. No attempt at adjudication should be made. It should be required that any paper on which such a question is raised should continue to be listed in the authors' CV, followed by such a note as, "Results questioned (reference)."

I share what I perceive as Cicchetti's view that, with respect to the review process, the cup is more full than empty, but that there is merit in seeking to fill the cup further. I would differ in proposing that the review process be judged more by the results it produces (valid findings) than by the procedures it uses to produce those results, such as the "reliability" of reviewers. The approach used by Cicchetti does an excellent job, however, of discussing what should be done, regardless of which criteria are emphasized. Finally, the value of the discussion is as much in its potential to cause readers to reevaluate their roles in the review process as in the specific proposals presented.

Why is the reliability of peer review so low?

Donald Laming

Department of Experimental Psychology, University of Cambridge, Cambridge, England CB2 3EB
Electronic mail: *drjl@phx.cam.ac.uk*

I compliment Cicchetti on a careful and detailed survey of studies of peer review in many different disciplines. Of the 58 tabulated correlations between independent referees, only four fall short of 0.18 and only four exceed 0.40 (of which the highest reduced to 0.38 on replication). What is to be made of these low levels of interreferee agreement? Cicchetti is dispassionate in his presentation and I am not sure how he feels about these results. But most scientists would, I believe, say these levels are not good enough and need to be improved. I am going to argue, on the contrary, that significant improvement may not be possible.

1. Summary of argument. Laboratory studies of absolute judgment of simple stimuli (frequencies of pure tones, for example, or sound pressure levels) show that such judgments are nevertheless relative – relative, usually, to the preceding stimulus in the experiment. This means that successive stimuli are compared with a constantly shifting frame of reference that limits the accuracy of judgment much more than any specifically sensory confusion. There are three quite different statistics from studies of the judgment of sound intensity that indicate that variation in the frame of reference accounts for about two-thirds of the variability of the judgments. Now transpose that result into the field of peer review. Two different referees use two different frames of reference for the evaluation of a submitted article or grant proposal. If those different reference frames contributed two-thirds of the variability of each evaluation, the correlation between independent peer reviews would be limited to about 0.33. I now fill out the details of my argument.

1.1. Absolute identification of simple stimuli. The most compelling example of the limited accuracy of absolute judgment comes from Pollack (1952). Pollack presented a series of tones to his subjects with frequencies selected at random from some number (m) of chosen values in the range 100 to 8,000 Hz, the number of different values ranging from 2 to 14 in different parts of the experiment. Each tone was presented for 2.5 sec. at about an 85 dB loudness level. The subject identified the tone by assigning it a number in the range 1 to m, and was then told the correct identification. As the number of different auditory frequencies and response categories increased above four (up to which point identification was nearly error-free), errors increased at such a rate that the accuracy of identification never exceeded a level equivalent to the use of just five categories without error. This result – specifically the limit of five categories without error – is not peculiar to frequencies of tone, but is obtained for many other sensory attributes as well, with only a few exceptions. (See Garner 1962, Chapter 3; Laming 1984, p. 155, Table 1).

This surprisingly low limit does not depend on sensory confusability. Jesteadt, Wier and Green (1977) found that there were about 2,000 just noticeable differences between 100 and 8,000 Hz. Moreover, in a series of supplementary experiments, Pollack (1952) manipulated several variables that affect discriminability without materially increasing the accuracy of identification of single tones. The only manipulation that increased accuracy was the presentation of a fixed reference tone (of a frequency known to the subject) prior to each stimulus to be judged (Pollack 1953). The limit to the accuracy of absolute judgment has to do with lack of a stable frame of reference.

The same conclusion may be drawn in a quite different manner from a study of magnitude estimation by Baird et al. (1980). When the intensities of two successive noise bursts differed by not more than 5 dB, the respective judgments (log magnitude estimates) correlated about +0.8. So some 0.64 of the variability in the judgment of the second stimulus was inherited from error in judging its predecessor (see Laming, in press). This result has been replicated several times. It is found in the work of Jesteadt, Luce & Green (1977), Green et al. (1977), Luce & Green (1978), and Green et al. (1980). All these experiments used the intensity of a pure tone as the stimulus attribute to be judged, but Baird et al. (1980) have demonstrated the result with the area of an arbitrary geometric figure as well.

1.2. Transmission of error in absolute judgments. If each stimulus, and the judgment assigned to it, is used as a reference point for the judgment of its successor, any error in the first assignment will be transmitted to the second. Herein lies a substantial source of inaccuracy. The experiment by Baird et al. indicates that about two-thirds of the error judgment may be accounted for in this way. There are two other experiments (more particularly, different statistics from two other experiments, not a mere replication of this present result) that also point to a proportion of about two-thirds.

J. C. Stevens and Tulving (1957) reported on a class of 70 undergraduate students making their first-ever magnitude estimates of loudness. Subsequently, S. S. Stevens (1971, Figure 8) plotted the cumulative distributions of those judgments (after "modulus equalisation" [Stevens 1971, p. 428] to remove differences in the absolute scale of different subjects' judgments) to show that those distributions were approximately log-normal. The inverse gradient of the cumulative distribution function (cumulative normal probability versus log estimate) estimates the standard deviation, and the variabilities of successive log-magnitude estimates, calculated in this manner, are tabulated by Laming (1984, p. 168, Table 2). For the very first judgment by each of the 70 subjects the variance was 0.010. For the second judgment, the variance was 0.020, and so on, increasing to an asymptotic level of about 0.030. So the variability contributed by a single stimulus presentation amounted to about one-third of the asymptotic value. The other *two-thirds* must have been inherited from the preceding judgment.

Of necessity, magnitude estimation requires that the subject receive no knowledge of results, lest it bias the judgments. In absolute identification, on the other hand, feedback after each trial is the rule. It is possible to compare the two procedures by conducting an absolute identification experiment without feedback, however, using the same set of stimuli and the same presentation schedule in both kinds of experiment. Braida & Durlach (1972, Experiment 1) is a case in point.

My third estimate comes from an as yet unpublished replication of Braida & Durlach's experiment. The stimuli were 10 1-kHz tones of 0.5 sec. duration, ranging in level from 50 to 86 dB SPL in 4 dB steps. For the first 3,000 trials the subject was asked simply to estimate the loudness (without being told that there were only 10 different stimuli). For the next 3,000 trials the subject was asked to identify the stimulus, but without feedback. The final 3,000 trials were again absolute identification, but with the correct response indicated immediately after each judgment. The data from all three tasks were analysed using Torgerson's (1958, Chapter 10) model of categorical judgment (see also Braida & Durlach 1972). This model estimates d' for the separation between adjacent pairs of stimuli (cf. Luce et al. 1982) and Figure 1 plots the cumulative d' for one subject performing each of the three tasks. There is not much difference between discrimination in the magnitude estimation and absolute identification tasks, both without feedback. When immediate knowledge of results is provided, however, d' improves. Comparing the aggregate from 50 to 82 dB, d' improves by the factor 1.79. This is equivalent to a decrease in the model variance to 0.31 ($= 1.79^{-2}$) of its former value. Evidently, in the point of reference used for the ensuing judgment, immediate knowledge of results substitutes the correct response for the response actually made, thereby preventing the transmission of error from one trial to the next. The proportion of error inherited from the preceding trial by this particular subject is 0.69.

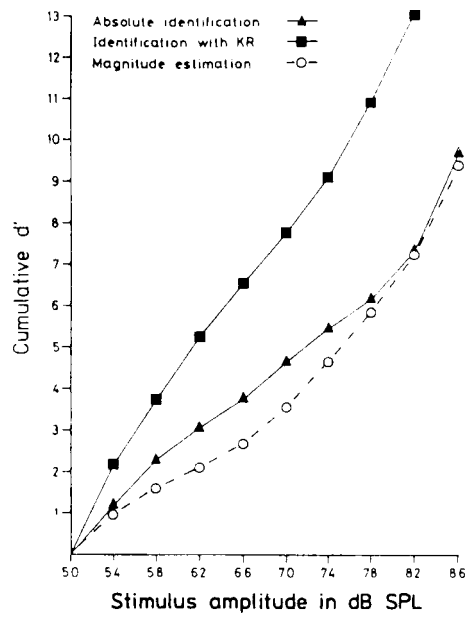

Figure 1 (Laming). Cumulative d', cumulated from the smallest stimulus (50 dB) upward, for magnitude estimation, for absolute identification without feedback, and for identification with immediate knowledge of results.

I have no theoretical foundation for this proportion of about two-thirds; it probably signifies no more than a fortuitous selection of experimental sources. But while, of necessity, the experiments are all somewhat similar, my three estimates are obtained from different kinds of experimental statistic. For this reason, two-thirds is a defensible value to transpose to the domain of peer review.

1.3. Application to peer review. In section 2 of his target article Cicchetti discusses a set of evaluative attributes and specific criteria for peer review of journal articles and grant proposals. He seems to envisage that these criteria are internalised by referees, possibly in the manner in which some musicians exercise "perfect pitch." An alternative scenario treats those criteria as no more than empty verbal formulae, which do not generate any particular behavioural correspondence between the bases of judgment by different referees. Instead, different referees formulate their judgments against different backgrounds of ideas which, according to the foregoing analysis, should account for about two-thirds of the variability of their assessments.

If those different frames of reference are truly independent, then the scope for agreement between two referees is limited to the remaining one-third of the variance, and that one-third corresponds as closely as one could reasonably expect to the spread of correlations reported by the authors. Test theory (Gulliksen 1950, p. 13) tells us that the reliability of any one assessment may be measured by the correlation between two independent judges, here 0.33. An editor has two or more such assessments on which to base a decision whether to accept an article for publication, however. The Spearman-Brown formula (Gulliksen 1950, p. 63) tells us that the combination of two independent referees raises the reliability of the editorial decision to 0.50. But I think most scientists would still regard this as unacceptably low. Some further exploration of the process of assessment is needed to discover what improvement may be possible.

2. Comparison with the marking of examinations. Public examinations in the United Kingdom (GCE "O" and "A" levels) typically achieve a mark-remark correlation of 0.9 or better (Murphy 1978; 1982). The re-marking in Murphy's studies was undertaken by an independent examiner and in that respect is comparable to peer review. I exclude from this comparison examinations in mathematics, physics, and kindred subjects, because for those subjects examiners are provided with, in effect, a list of the admissible answers and the marks to be assigned to each. Even though a professional article falls within one of those disciplines, it is nevertheless open-ended (the choice of topic is always at the author's disposal), so the correct comparison must be with examinations in subjects such as

English, history, and sociology. In these subjects a reliability of 0.9 or better is still commonly achieved.

The difference vis-à-vis peer review is the use of a marking scheme, however imprecise, which is practised by the examiners. This difference is immediately apparent in comparison with university examinations (Byrne 1980; Cox 1967; Eells 1930; Hartog et al. 1936; Laming 1990; again excluding mathematical and physical sciences), which usually have no such scheme. Take away any pretence at a marking scheme and the reliability of examination marks falls to near the levels reported for peer review. There is a substantial argument to be made in favour of my alternative scenario, that the "specific criteria" referees are assumed to use have little or no behavioural reality.

3. Scientific progress? The argument that follows next is, I suppose, a flagrant abuse of classical test theory. But it provides the vehicle for a particular pessimistic view of scientific progress that needs to be exposed to scrutiny.

Suppose that referees typically accord a weight w to an article or a grant proposal to be appraised, the residue $(1-w)$ being contributed by the variability in the background with which the article is implicitly compared. The quantity $(1-w)$ corresponds to the estimate two-thirds in section 1.3. Suppose, however, that the frames of reference implicitly used by different referees are not independent, but correlate r with each other, because the referees are chosen from within a common scientific tradition. The correlation to be expected between two independent assessments is then $([1-w]r+w]$). If a journal editor bases his decision on the reports of n different referees, application of the general Spearman-Brown formula (Gulliksen 1950, p. 78) suggests that the editorial decision will have a reliability of

$$r' = n[(1-w)r+w]/\{(n-1)[(1-w)r+w]+1\}. \quad (1)$$

In Equation 1, r' is the concordance between the articles ultimately published and the criteria referees are supposed to apply. It is also the correlation between the frames of reference with respect to which subsequent referees will formulate their assessments of the next generation of journal submissions. What happens to successive values of r'? Do they converge to a limit and, if so, what is the value of that limit?

For admissible values of n and w the process does, indeed, converge, and the only possible limit is 1. That is fine; the process of peer review converges on a common frame of reference that, in a scientific discipline, is presumably in concordance with the state of Nature. But the uncomfortable import of the correlations reported in the target article is that this does not seem to be happening. The only reconciliation of the theoretical argument and the empirical data that I can at present think of runs as follows: Once attention is confined to the rather narrow stratum of potentially plausible grant proposals and publishable papers, referees are, for the most part, unable to tell the meritorious from the rest, and scientific "progress" is principally a random progression.

It is clear from their espousal of proposition (a) in section 8 that Cicchetti does not share my pessimism. He takes an optimistic view of scientific progress, but on what evidence? The optimistic view envisages that most published research will have some detectable effect on the state of the field 50 or even 100 years hence. It is difficult to see what evidence could be brought to bear on such a proposition. But I have had occasions to consult journal articles in my subject (experimental psychology) from 50 and sometimes 100 years ago. On those occasions, I have often glanced at the table of contents of the journal volume being consulted, just to see what else was there. It is interesting to discover an article of historical significance that one has heard of in a different context. But, usually, nine articles out of ten, even 19 out of 20, have proved to be completely unknown. The present state of my subject would be no different if those articles had never been published. Is the situation any different today?

Should the blinded lead the blinded?

Stephen P. Lock
British Medical Journal, Tavistock Square, London WC1H 9JR, England

Given the apparent inherent variation in opinions, not only between referees themselves, but between referees and editors, what can be done to improve things? My personal hierarchy of proposals would start with the editors' subcategorizing the questions they expect reviewers to consider. For example, instead of answering the question, "Is the work original?" the referee could indicate, "New to me; known to me: (a) by rumour, (b) by personal communication, (c) from presentation at a conference (with or without abstract), (d) from published work, or (e) from retrieval from database."

Next, I would advocate two unconfirmed hunches. First, that the quality of a decision is enhanced by having an editorial "hanging committee" (named by analogy with the selection body of the London Royal Academy of Arts), which discusses most of the articles with "grey" reviewers' reports (that is, 2–4 on a 1–5 scale of reject/accept). Second, that for a very general journal better quality reviews are obtained if the choice of reviewer is delegated to an assistant editor in the subfield; expert knowledge by one competent reviewer is more helpful for making a decision, in my view, than having two or more opinions from referees with no specific expertise.

Paramount among my suggestions, however, is the need for blind review – or at least for editors to study it under their own circumstances. To earlier suggestions of the cogency of this view by Mahoney (1977) and Peters and Ceci (1982) must be added the results of the rigorous study by McNutt et al. (1990). Not only did the last show that blinding was feasible for the editorial office and successful for 76% of reviewers, but on a 3-point scale there was a 21% improvement in the quality of reviews, as well as a striking increase in the proportion of excellent reviews among the blinded reviewers. So, in addition to replicating these findings for other journals, another study that is now urgent would determine the effect of blinding on the editors themselves, particularly in some recent findings (Garfunkel et al. 1990). Some 25 manuscripts that had clearly been revised and accepted for publication in the *Journal of Pediatrics* were sent for re-review by two additional referees, and then reevaluation by three experienced, independent assistant editors. Most manuscripts were thought by the new reviewers to have defects that warranted further revision, but, though one of the participating assistant editors would have requested revision more often than the other, there was infrequent disagreement among them about the basic decision to accept or reject.

My second group of comments relates to publication bias, in particular, the preference for original over replicative work and for manuscripts reporting positive results. Possibly, now that editors have recognised the pitfalls of this attitude, which were well discussed at the First International Congress on Peer Review in Biomedical Publication (Chalmers 1990; Chalmers et al. 1990; Dickersin 1990; Sharp 1990), the problem will diminish, particularly if authors appeal on this account. Nevertheless, the editorial decision must depend on circumstances: For example, whatever the findings, I believe that the *British Medical Journal* would be interested in publishing other studies (however long and detailed) of the incidence of leukaemia in children of fathers who had been exposed to high doses of radioactivity in their work in various atomic power stations, thus confirming or refuting the recent work by Gardner et al. (1990) at Sellafield. The recent introduction of structured abstracts (Ad Hoc Working Group for Critical Appraisal of the Medical Literature, 1987) may also make it easier for editors to find the space for confirmatory reports or reports with negative results. With a limit of 400 words and a tightly defined vocabulary, these allow a detailed statement of the study's objectives, setting, methods,

analysis, endpoints, findings, and conclusions. Thus, in the future, after peer review of the full article, editors might like to suggest that some reports be printed in structured abstract form with the substantive report incorporated into an on-line database.

Finally, in the past editors in many disciplines have been able to sleep regardless of mistaken decisions because of the concept of Western plurality: What gets rejected by the *Lancet* will be published by the *British Medical Journal* or *Gut*. Cicchetti shows disquietingly that this is not so for some journals in some disciplines and for research grant applications in any of them. We urgently need retrospective and prospective studies on these findings: Did the rejected ideas stand up in the light of history? Was the rejection the result of lack of rigour by the researcher, even though the original ideas were sound? Were they ever studied by anybody else? And should society see that some journal space/money is reserved for "zany" ideas in all disciplines as in David Horrobin's journal *Medical hypotheses?* Some of these studies need not be too time consuming and might make fitting contributions to the 1992 Second World Conference on Peer Review. (Details are available from Drummond Rennie, *Journal of the American Medical Association*, 535 N. Dearborn St., Chicago, IL 60610.)

Justice, efficiency and epistemology in the peer review of scientific manuscripts

Michael J. Mahoney

Department of Psychology, University of North Texas, Denton, TX 76203

Cicchetti has written a valuable and comprehensive critique of the reliability of peer review for journal manuscripts and grant applications. Besides addressing some of the many subtleties, complexities, and practical issues involved in peer review, the author has identified an important and well-replicated phenomenon in this area, namely, "that reviewers are indeed substantially more in agreement on which scientific documents to reject than on which to accept." This may be a heartening conclusion for those evolutionary epistemologists who view selection processes as primarily negative, but the authors offer a provocative discussion of caveats in the interpretation of this pattern.

I agree that the high rate of rejection for grant proposals is of greater concern than the rejection of journal manuscripts in the social and behavioral sciences. In their recent report, the National Research Council (1988) stated that basic research in these sciences merits a 30% increase in funding over current levels. Unfortunately, federal funding for such research has been declining sharply since 1983. In fact, although federal support for research in the other sciences has increased by 36% in the last 15 years (1972 to 1987), federal funding for basic research in the social and behavioral sciences has been reduced by 25% during that same interval. Needless to say, those psychologists who recognize the need for a scientific basis for their activities will now have to work even harder to reverse the trend of declining support for their research.

I also agree that allowing authors to engage in multiple manuscript submissions is not a viable solution to the problem of high rejection rates from journals. Indeed, the project recently reported by Epstein (1990) illustrates some of the problems of this practice as a research strategy, let alone as standard practice for scientific authors. Epstein apparently plagiarized an article and submitted it to 146 professional journals in (or related to) the field of social work. His methodology and quantitative results were very weak, and yet he offered an interpretation of his study that was harshly critical of the professionals who had unwittingly invested perhaps a total of 1,000 hours in his project.

With other scientists, I believe that studies of peer review are a priority for future science studies. In this regard, it is reassuring that the American Medical Association sponsored a special conference on the topic (Rennie 1986). "I do not believe, however, that such a compelling priority justifies violations of human rights and the professional codes of ethical conduct developed to protect them. The questions of how, when, and why we 'draw the lines' demarcating ethical and unethical conduct will remain with us, of course, and we can only hope that their challenges will teach us some important lessons about ourselves and our methods in the process (Mahoney 1990, p. 54)."

Reflections on the peer review process

Herbert W. Marsh[a] and Samuel Ball[b]

[a]School of Education, and [b]Faculty of Education, University of Sydney, Sydney NSW 2006 Australia

The peer review process is one of the most highly regarded and frequently used procedures for evaluating the academic merit of academic manuscripts, grant proposals, tenure/promotion applications, and academic monographs and textbooks. Hence, peer review is of utmost importance to the academic community and we welcome the comprehensive review by Cicchetti. It brings together discussions of theoretical issues, methodological/statistical concerns, a diversity of empirical studies, and practical suggestions for the interpretation and application of the peer-review process. Given the scope of his review, we will limit ourselves to comments on a few specific aspects of the peer review process for academic journals.

Reliability of the editor's decision. Marsh and Ball (1981; 1989) noted that low single-reviewer reliabilities should not be confused with the reliability of the decision of the review process. First, the reliability of the mean response by multiple reviewers depends on the number of reviewers; a single-reviewer reliability of .36 results in a reliability of .53 for 2 reviewers, .65 for 3, and 0.69 for 4 (using the Spearman-Brown equation). Second, the editor serves as an implicit additional reviewer, further contributing to the reliability of the final decision. Third, the editor's decision is typically based on additional factors not considered in single-reviewer reliability estimates, such as the detailed written comments provided by reviewers, author responses to reviewer criticisms in revisions and/or separate correspondence, and sometimes further reviews of the revised manuscript. Fourth, the exclusion of manuscripts judged to be grossly inappropriate further attenuates reliability estimates, in that agreement among reviewers would probably have been best for these manuscripts. Thus, the editor's decision is likely to be substantially more reliable than that of the single reviewer.

Policy decisions and practices that operationally define the peer-review process. A general framework for the review of academic manuscripts is common to most journals. Editors, chosen for their broad expertise, generally screen manuscripts for appropriateness and then assign them to one or more reviewers with particular expertise in relevant areas. The reviewers are asked to provide written critiques, ratings, or recommendations on the advisability of publication. Editors, relying on these reviews and their own appraisals of the manuscripts, decide to accept (perhaps, subject to revision) or reject the manuscripts, or to seek further review. Many details, however, are left to the discretion of the editor including: (1) the criteria to be used by the reviewers; (2) the form of the review (ratings, written critiques, etc.); (3) the number of reviewers; (4) how reviewers are selected; (5) whether authors and reviewers are anonymous; (6) what is done with divergent or inconclusive reviews; (7) the extent to which the editor's decisions are

dictated by the reviews; (8) whether revised manuscripts are sent for further review to the same or different reviewers; and (9) whether authors have the right to challenge reviews or request re-reviews. These policy decisions are typically made in an ad hoc fashion, and editors have little guidance in establishing the policy practices that constitute the peer review process. Whereas research reviewed by Cicchetti addresses some policy practices, too little research has been done summarizing existing policy options or testing their effectiveness.

The design of review surveys: Multidimensional components and externally anchored scales. Several studies reviewed by Cicchetti considered ratings on specific multiple criteria in addition to overall ratings, and better agreement was found for these (e.g., attention to relevant literature and research design). Some disagreement among reviewers may arise from the way they weight the different components in determining their overall recommendation, or it may be limited to one particular aspect of the manuscript. There is surprisingly little effort to determine what the factorial structure of responses is, however, or whether more reliable composites could be obtained by averaging the different subscales.

Marsh and Ball (1989) developed a 21-item reviewer survey based on a content analysis of written critiques. Factor analysis of responses to this survey clearly identified four factors affecting the outcome of reviews: research methods, relevance to readers, writing style and presentation clarity, and significance or importance. Multitrait-multimethod analyses of agreement among multiple raters of the same manuscripts provided modest support for convergent validity and for the distinctiveness of the rating components, but it also indicated a substantial "halo effect" in the ratings by a given reviewer. It is interesting that halo effects associated with the overall recommendations were much smaller than with responses to the 21 rating items (even though one was also an overall rating item). The explanation seemed to be that the response categories in the overall recommendation were much better anchored to concrete behaviors (e.g., accept as is or with slight revisions, reject outright) than the 9-point rating scale for the 21 items. Consequently, single-reviewer reliabilities based on various combinations of the 21 rating items were no higher than for the overall recommendation. The results suggest the potential usefulness of multidimensional rating scales, but also point out the importance of having well-anchored response scales that minimize halo effects and response biases idiosyncratic to how each reviewer interprets the response scale.

In discussing attempts to improve reviewer reliability, Marsh and Ball (1989) also noted that such proposals must be evaluated in terms of their likely impact on validity. For example, there are relatively objective characteristics on which reviewers could agree that are unrelated to manuscript quality. In addition, specific strategies may affect differentially reliability and validity. For example, editors are likely to send the same manuscript to reviewers having different perspectives, and this strategy may lower reliability but increase validity.

The process of peer review: Unanswered questions

Linda D. Nelson

Department of Psychiatry, Medical Center, University of California, Irvine, Orange, CA 92668

Peer review is essentially a classification system that involves both *process* (i.e., the activity that led to a decision) and *outcome* (i.e., the decision itself).

Although Cicchetti states from the outset that a major objective of his study was to analyze the peer review *process*, *outcome* appears to be the focus of his target article. Dependent variables (e.g., accept, reject, resubmit) were carefully examined in the context of their own and others' work, with the results clearly displayed in tabular form (e.g., rates of interrater agreement). Their recommendations regarding the appropriate statistics for evaluating and determining standards of reliability added new and potentially useful information to the study of peer review. His conclusions regarding possible interactions between the nature of the discipline (e.g., general vs. diffuse) and acceptance rates were interesting and highlighted bases for differential outcomes in levels of agreement. Focusing part of his discussion on peer review as it relates to major funding sources (e.g., agreement on grant proposals by type of study and area of discipline) offered new interpretations regarding outcome to this important phenomenon. In short, the author is to be commended for his efforts in updating and expanding our understanding of peer review as it applies to manuscripts and grant proposals.

Rather than replacing outcome as a topic, per se, I would have liked some additional discussion on the process involved in peer review. This point is important to me as a psychologist and researcher because it challenges the interplay between what a person thinks and what a person does. The author cites Kuhn (1962) as someone who stresses the importance of peer evaluation on scientific activity. It is noteworthy that Kuhn's remarks (1962) were used to support the importance of considering the relationship between process and outcome in psychotherapy 16 years later (Orlinsky & Howard 1978). To ignore the influence of process on outcome in peer review misses an important link between what (or how) individuals think and what leads them to certain conclusions. Although Cicchetti touches on this in his discussion of reviewer bias, I am not certain from his presentation of supportive literature whether bias is actually an operative factor affecting outcome: One experiment used to support the role of author affiliation status and review outcome was soundly criticized (see commentaries on Peters & Ceci, 1982); another (Mahoney 1977) relied on a "qualitative analysis" of reviewers' remarks.

Cicchetti further implies that peer review can engage referee/reviewer variables that are so powerful that evaluative criteria (e.g., adequacy of methodology) become secondary factors in decision making. He states

> On the basis of the best controlled studies of the peer review process to date, we are forced to conclude that referees do at times apply subjective criteria (that) cannot be described as "fair," "careful," "tactful," or "constructive," despite the fact that such traits are widely accepted as desirable characteristics of referees.

The author then cites, as an example of this phenomenon, the increased likelihood of reviewers accepting manuscripts based on type of results (e.g., positive) instead of a manuscript's overall worth (e.g., adequacy of methodology).

The notion that unfair, subjective criteria may be imposed by journal "gatekeepers" is a provocative one. Even more unsettling is the contention by Mahoney (1977) that this phenomenon represents "confirmatory bias," wherein reviewers deem acceptable manuscripts that coincide with their beliefs and reject those that do not. Does this mean that papers tend to get published on the basis of statistical significance? Furthermore, are we to assume that reviewers tend to agree with any alternative hypothesis set forth in a paper in such a way that failure to reject null hypotheses is viewed as inconsistent with their beliefs? Considering direction of results as an operative variable in peer review is further clouded by the fact that the main experiment used to justify the notion involved a direct manipulation of the operative variable in question (Mahoney 1977). An investigator's choice of an independent variable, and Cicchetti's conclusions regarding its impact on the review process, represent subjective determinations, in this instance regarding which reviewer variables within the complex process of peer review characterize prepotent predictors of outcome. As a journal reviewer, I would venture to say that unfortunately, papers with

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.

Is 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:

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

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

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

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,

yields a correlation for the two highest levels of evaluation (acceptance) of −.017, whereas for the two lowest levels of evaluation (rejection) the analogous correlation is .107. The difference between R's of .107 and −.107 surely seems more modest than the difference between agreement rates of 70% and 44%, although in the same direction. Indeed, Cohen (1988) treats the former difference as "small" ($q \cong .12$) and the latter as "medium" ($h \cong .53$).

Conclusion. In this commentary on the indices of reliability evaluated or used in the target article, three are recommended as information-efficient: Pearson R, intraclass R, and Cohen's *kappa* for the 2 × 2 case. It is further recommended that 3 others *not* be used: rate of agreement, *kappa* for tables larger than a 2 × 2 (with $df > 1$), and χ^2 or any other tests of significance, because they depend on sample size as well as on reliability per se.

Toward openness and fairness in the review process

Byron P. Rourke

Department of Psychology, University of Windsor, Windsor, Ont. N9B 3P4, Canada

Cicchetti's target article is thorough and thoughtful. The issues addressed are important, and each is dealt with in a systematic manner. Suggestions for future research are clear and relevant. The overall approach to the topic is disinterested and scientific. There is little to criticize in this presentation. I would point to some issues that seem important to me, however, that either were not mentioned or did not receive the sort of emphasis I would have given them. These are as follows:

1. Signed reviews. From my perspective, it would seem desirable for all reviews to be signed. I say this in the full realization of realization of the problem that the author mentions regarding "younger" scientists and the recriminations that they may suffer as a result of criticizing the work of more established researchers. There is also the possibility that the younger scientist may mute criticisms if forced to acknowledge their source. I would argue, however, that insisting that all reviewers acknowledge their identities is fair and just. It would be easy to implement; in any case, it will eventually become the rule rather than the exception. (Hence, with regard to this issue, I would disagree with the position taken by Cicchetti.)

With respect to fairness, I would simply point out that hiding behind the cloak of anonymity opens the door to the worst sort of blackballing. Stating opinions that one must stand by – and defend, if necessary – is part and parcel of what the social dimension of science is all about. With respect to ease of implementation, I would suggest that mandating the acknowledgment of reviews would be difficult at first, but reviewers (i.e., active scientists) would soon become accustomed to the process. Finally, as is now the case for many governmental granting bodies, the freedom of information "movement" will eventually target journals to achieve the same degree of openness as now exists in the area of grants.

2. The role of the journal editor. Journal editors can do much to increase the probability of positive or negative reviews through their choice of the consulting editor(s) (CE). Experienced journal editors have sufficient knowledge about the likely reactions to particular pieces of research of many, if not most, of the members of their editorial boards. Even though CEs are retained on the editorial board because of their perspicacity as well as objectivity, some can be "depended on" to look askance at particular research designs, methods of analysis, subject populations, and any number of other important dimensions of scientific papers. Furthermore, it is my impression that negative biases will have a much greater effect than will positive biases: That is, it is likely that the negative biases of CEs will have a larger (negative) effect than will their positive biases about particular kinds of research. Indeed, it has often been my experience that CEs will bend over backward to find fault with research that is clearly similar to their own (i.e., of a sort toward which they would be expected to have some positive bias).

This being the case, what is the journal editor to do? Random assignment of CEs is one alternative. In all but the most narrowly focused journals, however, this would result in many nonexperts reviewing the work in question. The only other solution is to attempt to balance the review process by choosing CEs who represent conflicting stances vis-à-vis the project in question. If this seems to call for a rather well developed sense of where CEs stand with respect to important issues in the field, I would emphasize that this is exactly what I mean to convey. There is no substitute for a fair, judicious, and experienced journal editor if the process of evaluation is to proceed in a fair and judicious manner.

3. A good article will get published somewhere; bad articles however, also tend to get published somewhere. Cicchetti does not see the high rejection of rates of, say, the *New England Journal of Medicine*, as much of a problem. He points out, with good reason, that the vast majority of articles rejected from the handful of very prestigious journals eventually see the light of day elsewhere. This is all well and good. What is not so good is the practice followed by some authors of resubmitting articles to any number of journals until they get lucky. For example, over the past 14 months I have had the opportunity to review an article for four different journals. I commented on the article the first time around, and recommended a number of changes in it. When I received the unamended article from a second journal, I informed the editor that I had reviewed the piece previously, and forwarded my first review to him. The third and fourth times around for the article – still in its unamended form – were handled by simply informing the editor that I had seen the article before, that I had recommended changes in it, that these had not been made, and that I did not want to go to the trouble of commenting any further on the manuscript.

I should be quick to add that my experiences in this regard are not unique: I have shared similar stories with many of my neuropsychological colleagues. What usually transpires in such cases is that the author eventually finds space for the article in a low prestige journal or, worse, a high prestige generalist journal that does not have the editorial expertise available for the proper evaluation of the manuscript.

What should be done about this situation? I think the answer is quite simple: Demand that authors submit a statement to the effect that their article has been submitted to journals, X, Y, or Z and that the article has been rejected. In addition, the reviews of the article could be provided to the editor of the journal that must now decide on the acceptability of the work.

4. Journal editors should provide the verbatim reports of CEs to authors. This point was not raised in the target article. Nevertheless, it is an important one. Justice and fairness require that authors see for themselves the reviews of the work submitted. A précis of CE comments simply will not do. This procedure would enhance the fairness of the process even more if the suggestion cited above regarding the acknowledgment of the identity of the reviewers were adopted.

Finally, I would point out that the suggestions made by Cicchetti regarding future research in this field are quite important. These efforts will not only enhance our knowledge about decision-making in the publication of scientific articles and the allocation of grants but will also aid immeasurably in generating new (and, one hopes, fairer) modes of operation by and for editors and granting agencies. [The commentator is co-editor of *Journal of Clinical and Experimental Neuropsychology* and *The Clinical Neuropsychologist*. *Ed.*]

yields a correlation for the two highest levels of evaluation (acceptance) of −.017, whereas for the two lowest levels of evaluation (rejection) the analogous correlation is .107. The difference between R's of .107 and −.107 surely seems more modest than the difference between agreement rates of 70% and 44%, although in the same direction. Indeed, Cohen (1988) treats the former difference as "small" ($q \cong .12$) and the latter as "medium" ($h \cong .53$).

Conclusion. In this commentary on the indices of reliability evaluated or used in the target article, three are recommended as information-efficient: Pearson R, intraclass R, and Cohen's *kappa* for the 2 × 2 case. It is further recommended that 3 others *not* be used: rate of agreement, *kappa* for tables larger than a 2 × 2 (with $df > 1$), and χ^2 or any other tests of significance, because they depend on sample size as well as on reliability per se.

Toward openness and fairness in the review process

Byron P. Rourke

Department of Psychology, University of Windsor, Windsor, Ont. N9B 3P4, Canada

Cicchetti's target article is thorough and thoughtful. The issues addressed are important, and each is dealt with in a systematic manner. Suggestions for future research are clear and relevant. The overall approach to the topic is disinterested and scientific. There is little to criticize in this presentation. I would point to some issues that seem important to me, however, that either were not mentioned or did not receive the sort of emphasis I would have given them. These are as follows:

1. Signed reviews. From my perspective, it would seem desirable for all reviews to be signed. I say this in the full realization of realization of the problem that the author mentions regarding "younger" scientists and the recriminations that they may suffer as a result of criticizing the work of more established researchers. There is also the possibility that the younger scientist may mute criticisms if forced to acknowledge their source. I would argue, however, that insisting that all reviewers acknowledge their identities is fair and just. It would be easy to implement; in any case, it will eventually become the rule rather than the exception. (Hence, with regard to this issue, I would disagree with the position taken by Cicchetti.)

With respect to fairness, I would simply point out that hiding behind the cloak of anonymity opens the door to the worst sort of blackballing. Stating opinions that one must stand by – and defend, if necessary – is part and parcel of what the social dimension of science is all about. With respect to ease of implementation, I would suggest that mandating the acknowledgment of reviews would be difficult at first, but reviewers (i.e., active scientists) would soon become accustomed to the process. Finally, as is now the case for many governmental granting bodies, the freedom of information "movement" will eventually target journals to achieve the same degree of openness as now exists in the area of grants.

2. The role of the journal editor. Journal editors can do much to increase the probability of positive or negative reviews through their choice of the consulting editor(s) (CE). Experienced journal editors have sufficient knowledge about the likely reactions to particular pieces of research of many, if not most, of the members of their editorial boards. Even though CEs are retained on the editorial board because of their perspicacity as well as objectivity, some can be "depended on" to look askance at particular research designs, methods of analysis, subject populations, and any number of other important dimensions of scientific papers. Furthermore, it is my impression that negative biases will have a much greater effect than will positive biases: That is, it is likely that the negative biases of CEs will have a larger (negative) effect than will their positive biases about particular kinds of research. Indeed, it has often been my experience that CEs will bend over backward to find fault with research that is clearly similar to their own (i.e., of a sort toward which they would be expected to have some positive bias).

This being the case, what is the journal editor to do? Random assignment of CEs is one alternative. In all but the most narrowly focused journals, however, this would result in many nonexperts reviewing the work in question. The only other solution is to attempt to balance the review process by choosing CEs who represent conflicting stances vis-à-vis the project in question. If this seems to call for a rather well developed sense of where CEs stand with respect to important issues in the field, I would emphasize that this is exactly what I mean to convey. There is no substitute for a fair, judicious, and experienced journal editor if the process of evaluation is to proceed in a fair and judicious manner.

3. A good article will get published somewhere; bad articles however, also tend to get published somewhere. Cicchetti does not see the high rejection of rates of, say, the *New England Journal of Medicine*, as much of a problem. He points out, with good reason, that the vast majority of articles rejected from the handful of very prestigious journals eventually see the light of day elsewhere. This is all well and good. What is not so good is the practice followed by some authors of resubmitting articles to any number of journals until they get lucky. For example, over the past 14 months I have had the opportunity to review an article for four different journals. I commented on the article the first time around, and recommended a number of changes in it. When I received the unamended article from a second journal, I informed the editor that I had reviewed the piece previously, and forwarded my first review to him. The third and fourth times around for the article – still in its unamended form – were handled by simply informing the editor that I had seen the article before, that I had recommended changes in it, that these had not been made, and that I did not want to go to the trouble of commenting any further on the manuscript.

I should be quick to add that my experiences in this regard are not unique: I have shared similar stories with many of my neuropsychological colleagues. What usually transpires in such cases is that the author eventually finds space for the article in a low prestige journal or, worse, a high prestige generalist journal that does not have the editorial expertise available for the proper evaluation of the manuscript.

What should be done about this situation? I think the answer is quite simple: Demand that authors submit a statement to the effect that their article has been submitted to journals, X, Y, or Z and that the article has been rejected. In addition, the reviews of the article could be provided to the editor of the journal that must now decide on the acceptability of the work.

4. Journal editors should provide the verbatim reports of CEs to authors. This point was not raised in the target article. Nevertheless, it is an important one. Justice and fairness require that authors see for themselves the reviews of the work submitted. A précis of CE comments simply will not do. This procedure would enhance the fairness of the process even more if the suggestion cited above regarding the acknowledgment of the identity of the reviewers were adopted.

Finally, I would point out that the suggestions made by Cicchetti regarding future research in this field are quite important. These efforts will not only enhance our knowledge about decision-making in the publication of scientific articles and the allocation of grants but will also aid immeasurably in generating new (and, one hopes, fairer) modes of operation by and for editors and granting agencies. [The commentator is co-editor of *Journal of Clinical and Experimental Neuropsychology* and *The Clinical Neuropsychologist. Ed.*]

its fictions and errors, but never make any real progress. Referring specifically to "educational and psychological studies" (p. 310) as examples, Feynman (1985) has characterized this type of science as "cargo cult science: They follow all the apparent precepts and forms of scientific investigation, but they are missing something essential, because the planes don't land" (p. 311). What is missing is this: "It's a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty – a kind of leaning over backwards." (op. cit.).

Lately, signs have multiplied that Feynman's characterization of the behavioral sciences is less humorous than it sounds. Especially if one adopts a historical perspective, one finds two recurrent invariants: Both (a) absurd and sterile research trends, and (b) transparently erroneous claims persist far longer than would be expected on the basis of Rubin's roulette theory. Estes (1975) cites axiomatic measurement theory as an example of (a): "One reason for the relative paucity of connections between measurement theory and substantive theory in psychology may arise from the fact that models for measurement have largely been developed independently as a body of abstract formal theory with empirical interpretations being left to a later stage." (p. 273). That much is clear. What is unclear is why it took 20 years to notice that "the difficulty with this approach is that the later stage often fails to materialize" (op. cit.).

Similarly, in his recent autobiography, Luce (1989) cites mathematical learning theory as another example of false starts: "In learning, hundreds of papers studying and testing stochastic operator and Markov models have, in my opinion, come to very little" (p. 286). Purely random peer review might have produced this insight earlier. "At the risk of offending some colleagues," Luce can think of "only three areas where mathematical modelling can be shown to have had a profound impact." One of them is psychological testing, which, in his view, "is more mathematized than most people realize" (p. 285). This may have been a mixed blessing, however, since it provides numerous examples of (b). Perhaps the best known instance is the Burt scandal, which was less a scandal about Burt than about his peers: "What . . . are we to make of the fact that Burt's transparently fraudulent data were accepted for so long, and so unanimously, by the 'experts' in the field?" (Kamin 1981, p. 105).

My own experiences are linked to the rediscovery of the factor indeterminacy in the early '70s (Schönemann 1971; Schönemann & Wang 1972; Steiger & Schönemann 1976). The significance of the indeterminacy is that is vitiates all claims that "intelligence" can be operationally defined as "g." This flaw of Spearman's factor model went unnoticed for a quarter of a century, until Wilson (1928) finally pointed it out in a review of Spearman's (1927) *Abilities of man*.

More recently, I found that one of the most popular formulae for estimating "heritability," Holzinger's h^2, which is supposed to estimate the proportion of genetic variance in the total (genetic plus environmental) variance, is erroneous because Holzinger (in Newman et al. 1937, pp. 94–116) had made a mistake in his derivations. As a result, h^2 contains no environmental variance at all (Schönemann 1989). When this mistake was finally spotted after 60 years of uninterrupted use of h^2, several statistical editors refused to publish the correction for a variety of reasons that had nothing to do with the facts at issue: "As in the earlier review by another journal, the referees do not claim to have found mathematical errors in your development" (Solomon 1989). [See also Wahlsten "Insensitivity of the Analysis of Variance to Heredity-Environment Interaction" *BBS* 13 (1) 1990.]

A final example of (b) is Rosenthal & Rubin's (1978) failsafe solution of the "file-drawer problem" of meta-analysis. They proposed a formula intended to estimate the ("failsafe") number of suppressed studies from the number of published studies. Because they retrieved 345 studies on the experimenter expectancy effect, they estimated this number as 65,000 and dismissed the bias hypothesis as unreasonable for these data because, "it is unreasonable to suppose that there existed enough unretrieved nonsignificant studies to overwhelm the studies we were able to retrieve" (p. 385). Darlington (1980) soon noticed a problem with this reasoning, however: "Imagine that all the tested null hypotheses in a certain area are true, and that the only results published are the 5% of studies which achieve significance by chance. Suppose the 345 studies were published this way . . . then we are imagining that the total number of studies performed was $T = 20 \times 345 = 6900$. Thus, a correct analysis of the data from the 345 published studies should in fact lead $T = 6900$" – not 65,000, as Rosenthal and Rubin's failsafe formula had predicted.

Almost a decade went by after Rosental and Rubin published their *Behavioral and Brain Sciences* target article, before *Statistical Science*, patterned after *BBS*, finally published an obliquely worded criticism of Rosenthal and Rubin's failsafe logic (Iyengar & Greenhouse 1988). After first praising the failsafe method as a "clever formulation of the file-drawer problem," the authors point out "several drawbacks that limit its usefulness" (p. 115). One such drawback is "the assumption that the unpublished studies are in fact a random sample of all studies that were done" (p. 110), because it conflicts with the very file-drawer hypothesis the failsafe number is supposed to cure: "Now if there were publication bias in favor of studies with statistically significant findings, then the Z values for the unpublished studies would not be a sample from the standard normal distribution" (p. 115). In this case, too, the recorded evidence is at odds with the charitable null hypothesis that the long delay in correcting Rosenthal & Rubin's claims was solely due to chance. In fact, not just one, but two authors repeatedly tried to alert editors that something was amiss with the failsafe argument Rosenthal (1979) had described in more detail in the *Psychological Bulletin*.

Shortly after the article appeared, Darlington (1980) submitted a Note in which he challenged the failsafe argument with the simple counterargument cited earlier, concluding: "(Rosenthal's) formula appears to be incorrect, grossly overestimating X in some cases and grossly underestimating it in other cases" (Darlington 1981, Abstract). The editor encouraged him to revise his paper and then rejected the revision. A few years later, Thomas (1985) reached the same conclusion: "The solution proposed by Rosenthal for the 'file drawer problem' is a product of faulty reasoning and should be forgotten" (Abstract). Pinpointing the flaw in Rosenthal and Rubin's reasoning precisely: "The conclusion is inescapable. In general Z* [the failsafe number; and by similar argument Z'] are not standard normal, i.e. n(0,1) in distribution" (p. 9), Thomas anticipated Iyengar and Greenhouse by more than five years. In the end, neither Darlington nor Thomas received any credit.

To summarize: As long as the validity of peer review is negative, as these and numerous other examples suggest, the rational course of action is to diminish its reliability further, not to enhance it.

Disagreement among journal reviewers: No cause for undue alarm

Lawrence J. Stricker
Educational Testing Service, Princeton, NJ 08541

The modest agreement among reviewers of journal manuscripts amply documented by Cicchetti is not a cause for undue alarm.

1. Interreviewer reliability is greater than it seems. The average intraclass correlation or kappa of approximately .30 between reviewers' ratings describes the reliability of ratings by a single reviewer.[1] Reviewers are analogous to test items in this situation, and the value of .30 is akin to the reliability of one test item.

When items or ratings by different reviewers are combined, reliability systematically increases. Applying the Spearman-Brown prophecy formula (Gulliksen 1987) to a reliability of .30 for the ratings of 1 reviewer, the estimated reliability is .46 for the combined ratings of 2 reviewers, .56 for the ratings of 3, .77 for the ratings of 7, and so on. The use of three reviewers is increasingly common, and the reliability in this case, though less than ideal, is substantially better than that for a single referee.

2. Disagreement among reviewers is useful. Journal editors often select reviewers deliberately for their dissimilarity. Reviewers may be chosen because they differ in the nature of their expertise – one a substantive specialist, another a methodologist, for example – or in their theoretical viewpoints (e.g., Bakanic et al. 1987). It is not surprising then that reviewers disagree when assessing the same apparently carefully defined components, such as "importance," or "design and analysis." And, insofar as the reviewers attend to different aspects of a manuscript's acceptability, the validity of the combined evaluations is improved (e.g., Harnad 1985), just as using predictors that measure different portions of the criterion variance maximizes the multiple correlation of the predictors with the criterion.

3. Editorial decisions should not be based solely on reviewers' ratings. Concern with the agreement among reviewers seems linked in large part to a model of the journal editor as a kind of psychometric clerk who simply adds up the scores that a manuscript gets from each reviewer and then accepts the paper if it achieves a passing score. Numerous anecdotes (e.g., Goodstein 1982), as well as the close associations reported by Cicchetti and others between reviewers' recommendations and editors' decisions, suggest that this model may indeed describe the behavior of some editors. But good editors are not clerks. They read the manuscript, appraise the reasons reviewers give for their recommendations, and weigh all the information about it (e.g., Goodstein 1982). This kind of active decision making takes time and specialized knowledge, and may be too much of a burden for the sole editor of a journal that receives many submissions. The workload can always be divided up among a set of associate editors, however, each of whom has complete responsibility for processing papers in a particular area, a practice followed by the *Personality and Social Psychology Bulletin* and some other journals.

NOTE

1. The anomalous and unreplicable intraclass correlation of .54 for *American Psychologist* manuscripts (Cicchetti 1980, and unpublished; Scarr & Weber 1978) may arise because nine of the 87 papers were resubmissions. The nine were presumably revised in response to the initial reviews, making it likely that these papers would receive favorable ratings, especially if the original referees were used. When these manuscripts are excluded, the correlation drops to .45.

Chairman's action: The importance of executive decisions in peer review

Peter Tyrer

St. Charles Hospital, London W10 6DZ, England

Cicchetti has provided valuable data in support of a maxim I sometimes repeat to disconsolate writers of rejected manuscripts: "A determined author can get any rubbish published." The low levels of reviewer agreement found in this wide-ranging review may be regarded as unsatisfactory by some, but encouraging to potential authors. After noting that worthlessness appears easier to detect than excellence, our author-in-waiting must be reassured by levels of agreement between assessors that barely exceed those of chance when several referees are used. Even taking into account the omnibus quality of the R_1 and kappa statistics, which can obviously conceal islands of excellent agreement, the levels of agreement cannot be regarded as good by any scale of values. Bearing in mind that research careers and the funding of departments depend so much on peer review of scientific papers and grant applications, it is sad that apparently random factors play such a major part in success.

The target article also explodes the myth that papers concerned with the "hard" physical sciences are assessed with greater levels of agreement than papers of "soft" social and psychological subjects. The reasons for poor agreement, to paraphrase Shakespeare's Cassius, "lie not in the words but in ourselves, that we are underlings."

The editor of a scientific journal and the chairman of a grant-giving body are faced with much conflicting information before coming to a final judgment. Cicchetti outlines a number of ways of improving the reliability of peer review, but even if levels of agreement are improved, the position of the editor (or chairman) can be a very important one. Disagreement is often resolved by the taking of "chairman's action," whereby an executive decision is made to reject one or more of the views referees used in coming to a decision, or, alternatively, to send the manuscript (or application) to another referee independently. It needs to be appreciated that the editor usually has a completely free hand in choosing referees for any article. The bias of the editor can influence whether an article he would like to see published goes to a reviewer who is likely to provide a favourable report. Alternatively, a paper the editor does not want published can go to a tough and critical referee. The opinion of the editor is particularly important when contentious papers are being reviewed.

The vagaries of editorial and referee judgment are particularly important for a young worker on the threshold of a research career. In this vulnerable stage there is a danger that one or two rejections may mean the abandonment of a research endeavour, when a more hardened worker would be inclined to persist. To make allowances for the poor levels of agreement, and the importance of editorial interest and bias, it would be valuable for potential authors to be aware of the particular interests of the journals for which they are writing. For example, one important journal in the United Kingdom not only tries to write the first and last words about any relevant topic, but is prepared to take considerable risks in trying to achieve this. Another journal will bend over backwards to include topical material in its columns, and so the time of submission is all-important. Others, from the examination of their contents, consistently give *proportionately* much more space to one or two aspects of a subject even though there is no indication of this in the guidance to contributors. For example, for controversial issues such as the merits and disadvantages of community care in psychiatry, one well-known journal has a bias toward its merits and another toward its disadvantages. Only an informed author knows which to select first.

The fact that many papers rejected by one journal are subsequently published in other equally prestigious journals (Wilson 1978) suggests that bias of interest is much more potent than assessment of merit. In view of this, much more attention should be paid to giving guidance to potential authors, particularly young research workers, in preparing their manuscripts and choosing the appropriate journal for their submission (Freeman & Tyrer 1989).

One other implication of the findings, which is also a major criticism of peer review, is the low likelihood of approval for papers and grant submissions concerned with "ground-breaking" research. Cicchetti provides data suggesting that agreement about such submissions is likely to be poor and that the "safe" option of rejection is most likely. In such circumstances the whole policy of peer review appears stultifying, and one sometimes longs for the good old undemocratic days when the publication of a paper was dependent only on the editor's whim or on the cranky beliefs of millionaire philanthropists. Such

people relied more on personal intuition than the collective views of often faceless referees, and their success rates did not compare unfavourably. The citation rates (not necessarily an index of excellence) for the *Lancet* persistently show it to be amongst the top six medical journals, but it is only in recent years that outside referees have played any significant part in the acceptance of articles for publication. Although it would be unwise to turn the clock back to these early beginnings, a quarter turn anticlockwise would not be out of place.

Do peer reviewers really agree more on rejections than acceptances? A random-agreement benchmark says they do not

Gerald S. Wasserman

Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907
Electronic mail: *codelab@psych.purdue.edu*

Cicchetti ably summarizes the accumulating evidence indicating that peer review reliability is unimpressive. And he correctly concludes that this finding has implications that should influence the structure of peer review systems. He weakens this conclusion, however, by adding the notion that reviewers agree fairly strongly with each other when they reject, even if they do not agree strongly when they accept.

This notion is troubling for two reasons. First, it leads to a seductive rationalization for inaction: It is easy and quite comforting to say that there will never be enough money (or journal space or whatever) to allocate to all good research. Therefore, we should be content if we can make sure that scarce resources are not wasted by allocating any of them to bad research. The present systems would supposedly do this despite their low overall reliability, if they really did consistently exclude bad research.

The second reason is that the notion is counterintuitive: Acceptance and rejection are just opposite sides of the same dichotomy. What is true for the one should be true for the other. This intuition prompted me to examine the evidence that led Cicchetti to his notion. Specifically, I compared his actual results with benchmark results one would obtain if no real agreement existed and if reviewers were making purely random judgments. This examination shows that the intuition is correct and his notion is unfounded. I will illustrate the examination with a detailed analysis of the data Cicchetti presents in his Table 6. And I will present expressions to calculate the general case:

Figure 1 gives a graphical representation of Cicchetti's data. It represents the peer review process as a sequential flow chart, even though the reviews are actually done independently. The input to the process is 150 grant proposals, of which Table 6 indicates that 52 got high ratings and 98 got low ratings. I have interpreted these tabular entries to mean that the reviewers' average rating was high for 52 of the 150 proposals and low for the other 98. I have added the further assumption that the average individual reviewer's performance is given by the collective average of all the reviewers' performances.

The proposals are read by one peer reviewer (Rater 1) who accordingly judges (on average) that 52 proposals are high and should be accepted (YES), while 98 proposals are low and should be rejected (NO). Then these proposals are read by the second peer reviewer (Rater 2), who also says YES to 52 (i.e., 28+24) proposals and NO to 98 (i.e., 24+74) proposals (which shows that the flowchart representation does not depend on which reviewer actually judged first).

I backed into the relation between the two raters' judgments by using the agreement percentages in Table 6. They show, as one would expect from the weak reliabilities reported in Table

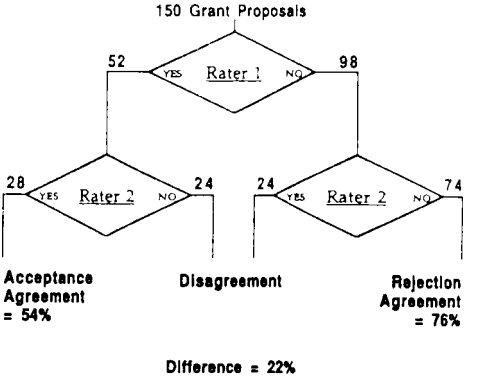

Figure 1 (Wasserman). Flow chart intended to represent the evaluation of grant proposals by two peer reviewers. Data were taken from Table 6 of the target article. See text for detailed explanation.

6, that the reviewers' judgments are weakly correlated: Rater 2 gives 28 YESs to the 52 proposals rated YES by Rater 1; this figure is computed from the tabular acceptance agreement of 54%. On the negative side, Rater 2 gives 74 NO's to the 98 proposals rated NO by Rater 1; this figure is computed from the tabular rejection agreement of 76%.

Figure 1 shows, as noted in the target article, that rejection agreement is 22% higher than acceptance agreement. It is on this kind of difference that Cicchetti bases his notion. But the first question should be: Against what benchmark should these numbers be evaluated? Figure 2 shows a benchmark created by a completely random process. In this case, the peer reviewers do not read the proposals. Instead, each reviewer has a bucket that contains 150 balls with 52 white balls and 98 black balls. Each proposal is "judged" by reaching into the bucket and

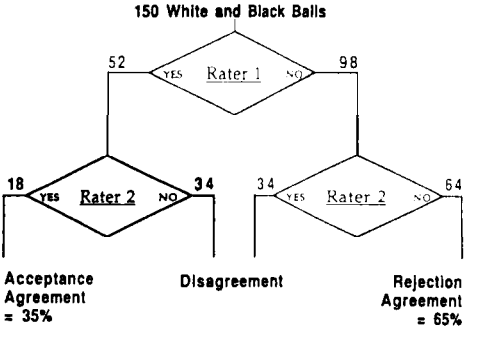

Figure 2 (Wasserman). Flow chart representing a benchmark based on purely random agreements between two separate reviewers evaluating 150 proposals. The first reviewer does not read the proposals, but instead evaluates the proposals by pulling balls from a bucket with 52 white balls (YES) and 98 black balls (NO). The second reviewer has another such bucket to use independently.

pulling out one ball. If the ball is white, the reviewer says *yes*; if black, the reviewers says *no*. This process gives a random-agreement benchmark that can be compared with the actual data. This comparison reveals the following effects:

The apparently greater rejection agreement over acceptance agreement is an illusion. Two outcomes support this conclusion when the data are examined in the metric of percentages: First, the difference between acceptance agreement and rejection agreement is *lower* in the real data (22%) than it is in the random benchmark (30%). Second, actually reading the proposals produces *less* of an improvement over benchmark in rejection agreement (11%) than in acceptance agreement (19%). Both effects are opposite to what one would expect from the target article narrative.

The benchmark difference between acceptance and rejection agreements is a simple function of the acceptance rate. This can be seen by examining the generalized random result, as illustrated in Figure 3: The result is obtained by expressing the number of proposals as n, the typical peer reviewers' *yes* ratings as p (in proportions), and the reviewers' *no* ratings as $q = 1 - p$. Then the acceptance agreement is $np^2/np = p$, the rejection agreement is q, and the difference is $q - p = 1 - 2p$. These expressions show that acceptance agreement, of course, goes to zero as the proportion accepted goes to zero. At the same time, rejection agreement goes to 1.0 and the difference goes to 1.0. Equality of acceptance agreement and rejection agreement only occurs in the random benchmark for the special case when $p = q = 0.5$.

A metric does exist in which one finds what intuition predicts, namely, a perfect equivalence between acceptance agreement and rejection agreement. One obtains this result if one avoids the slippery ground of percentages and proportions. Rather, one simply *counts* proposals. Doing this shows that, relative to the chance count in Figure 2, peer review in Figure 1 increases the count for both forms of agreement by exactly 10 proposals each.

The general expression for the above increase in agreement count is given by $2npq\phi$, where n, p, and q are as defined above, and ϕ is the (fourfold point; McNemar 1955, p. 202) correlation between reviewers. (For the data of Table 6 and Figure 1, $\phi = .29$.) This expression is easily interpreted by reference to Figure 3: There are *two* ways of reaching a disagreement and when ϕ is zero, as it is in both Figures 2 and 3, each way produces a count of npq. If the correlation were perfect ($\phi = 1.0$), however, then no disagreements would exist. In that case, the acceptance agreement count would be given by $np^2 + npq = np$, and the rejection agreement count would be given by $nq^2 + nqp = nq$. This confirms the intuition about symmetry: In general, peer review increases each form of agreement by $npq\phi$ counts. In the particular case of Figure 1, $npq\phi = 150{*}0.35{*}0.65{*}0.29 = 10$.

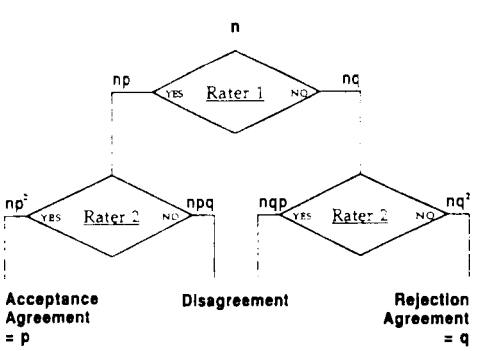

Figure 3 (Wasserman). Flow chart representing the generalized random-agreement benchmark. Note that agreement here is expressed in proportions, not in percentages. The number of items reviewed is n, the acceptance proportion of the average reviewer is p, and the rejection proportion is $q = 1 - p$. See text for detailed explanation.

The real peer reviewers described in Table 6 and Figure 1 do agree slightly more often than one would expect from the random benchmark in Figure 2. The effect is very slight, however: They only get together on 20 more proposals out of 150. By contrast, it will not escape notice that the random benchmark accounts for most of the variance in reviewer behavior. Hence, in this particular case at least, the benchmark is a fair model of a peer reviewer.

What to do about peer review: Is the cure worse than the disease?

Thomas R. Zentall
Department of Psychology, University of Kentucky, Lexington, KY 40506
Electronic mail: *zentall@ukcc.bitnet*

Peer review is among the most important professional services that scientists provide. It determines what research gets funded and what research gets published and in what journals. As Cicchetti so carefully documents, it is a system that is seriously flawed because of inherent subjectivity and reviewer bias.

The question is, what changes can be made in the system to eliminate the flaws? Any *major* change in the peer review process is likely to create its own problems, perhaps even more serious ones. Furthermore, given that the review process depends on the voluntary contribution of reviewer time, one needs to weigh the potential benefits that might accrue from change against the costs involved.

The two major issues raised by Cicchetti are, (a) the surprising, but well-documented, low reliability of grant and manuscript reviews, and (b) unfairness in the review process due to reviewer bias. Many of Cicchetti's suggestions address reviewer bias, but any variable that brings out a pervasive reviewer bias is likely to *increase* reliability, though perhaps at the expense of fairness. Thus, the issues of reviewer bias and fairness may be negatively correlated. Increasing fairness in the review process may be a valid goal, but can these biases be removed and, if so, at what cost?

Blind review. Voluntary blinding defeats its purpose because those most likely to benefit from their reputation would be least likely to blind, and, as Cicchetti notes, mandatory blinding may be impossible to enforce. On the other hand, is it really unfair to include knowledge of the author's reputation in one's judgment of suitability for publication? Just as statistical tests address the question of reliability of findings, so too, the reputation of the author may provide indirect, supplementary information about the reliability of the findings (though clearly, the latter should be given less weight).

Cicchetti also notes a related bias due to self-citations of "in press" research. Shouldn't the fact that related findings have gone through a (typically, stringent) review process argue for their increased reliability? Ideally, research findings should be able to stand on their own, but in reality experimental results are usually evaluated in the context of prior research, and in-press self citations are a part of that literature. Because of the inaccessibility of these papers, however, it is reasonable, albeit cumbersome, for editors to request that preprints of such citations be included with the manuscript to be reviewed.

Blind reviewer. Would reviews be fairer if the reviewers could not hide behind their anonymity? I suspect that the resulting fear of retribution would greatly reduce referees' participation in the review process. The cure may be worse than the disease.

Bias against negative findings. The bias against negative findings is more complex. Would a nonsignificant difference between groups be significant with greater power (e.g., more subjects)? Could the failure represent a Type II error (failure to observe a difference when a real difference exists)? Negative findings sometimes occur when research is not done carefully (resulting in increased within-group variance) or they may be due to inadvertent fluctuations in the experimental treatment (resulting in reduced between-group variance). There is a reluctance to publish negative results because there are many more ways to fail to observe an effect than ways to observe it.

In some cases, failure to replicate represents the useful establishment of boundary conditions of a phenomenon (i.e., when an effect is found under some conditions but not others). On the other hand, when the negative findings occur under conditions comparable to those in which the original findings were reported, and those negative findings are not just an example of Type II error (i.e., they can be replicated), they should be suitable for publication (see, e.g., Roberts 1976).

Improvements. How can the system be improved? One of Cicchetti's suggestions is an appeal process. An informal appeal process already exists, in which authors who feel that an incorrect decision has been made can appeal to the editor. To allow newer contributors better access to the appeal process perhaps it should be formalized (e.g., authors whose submissions have been rejected would be informed that they have the option of responding to the reviewers comments).

A second suggestion by Cicchetti is to increase the number of reviewers. The larger the sample of reviewers, the more reliable their combined judgment is likely to be. The increase in reliability would, I think, offset the added cost in reviewer time.

Third, I would like to see distributed to reviewers a set of guidelines that warn of potential biases and suggest that reviewers try to avoid them. This may appear too simplistic, but it is a cost-effective strategy that could result in the significant reduction of unfair biases.

As to the lack of reviewer agreement, it may be that such variability is an inherent characteristic of the field. There does not appear to be good agreement on what constitutes quality research, what are minor methodological flaws, or what are important findings. Much of this disagreement represents strong theoretical and methodological (confirmation) biases that, realistically, cannot be eliminated. It may be possible for editors to reduce the effect these biases have on the review of a manuscript through the careful selection of reviewers who do not have strong biases against the kind of research or direction of findings submitted, and by directing reviewers to avoid introducing their biases into the review process.

Author's Response

Reflections from the peer review mirror

Domenic V. Cicchetti
VA Medical Center, West haven, CT 06516[1]
Electronic mail: *dicchetti@yalevm.bitnet*

In an earlier *BBS* target article on peer review it was noted that "the area that seems to be most promising – that of cross-disciplinary comparisons – is still relatively unresearched" (Peters & Ceci 1982, p. 252). Within this suggested framework, a number of hypotheses received support in the current target article, namely, that across the various disciplines: (1) agreement is better on manuscript and grant submissions of perceived poor quality than on submissions of good quality; (2) better-defined (specific and specialized) areas of scientific inquiry have higher acceptance rates and use fewer reviewers than less well-defined (general and less focused) areas of scientific interest; and (3) levels of chance-corrected interreferee agreement are rather low (R_i usually $\leq .40$). The disciplines thus far investigated have included psychology, sociology, medicine, and physics. These issues were discussed in the context of research design considerations, statistical or data analytic approaches, and suggestions for improving the quality of peer review. Another important issue, discussed briefly, was how editors or granting officials use the information supplied by referees (reliable or not) to arrive at publication or funding decisions.

I am gratified by the generally positive evaluations of my work and its perceived heuristic value in generating follow-up research. My Response focuses on the areas of concern expressed by the various commentators. These fall into five categories: (1) methodological, statistical, and data analytic strategies; (2) interpretation of the results; (3) using peer reviews to improve editorial/funding decisions; (4) improving the peer review process; and (5) future research in peer review.

1. Methodolological, statistical, and data analytic strategies

1.1. Corrigenda. Let me begin by pointing out several minor errors of omission and commission that have been corrected in the revised target article (compared to the preprint that was circulated to the commentators). The first, my own discovery, pertains to the data reported in Table 3, section B, which depicts the parallel relationship between acceptance rates for manuscripts submitted to *Physical Review* and the use of one or more reviewers. The significant relationships now become even more apparent because the last two subfields appearing in the table (Particles & Fields, General Physics) interchange positions to reflect the same ordering as in Table 3, section A. This means (as previously) that as the subfields tend toward more general focus (Nuclear Physics, Condensed Matter, General Physics, Particles & Fields), *both* the percentage of accepted manuscripts and the percentage of manuscripts using a single reviewer decrease significantly ($p < .000001$ in the former case, $p = .0003$ in the latter).

The second error, Table 5, was caught by one of the commentators, Eckberg. In the second row of the table, the number of rejected manuscripts should read 577 rather than 578. Since the correct N of 577 was used in the calculations, the resulting chi square(d) value of 57.895 ($p < .00001$) is correct.

The third error is that the ordering of the second and third endnotes had to be reversed to be consistent with correct footnote citation in the text. Finally, through another typographical error that escaped my review, the denominator of the Formula for R_i (Model I), now the third footnote, had to be amended, by removing the previously

initial term MS +. I am indebted to another of the commentators, **Hargens**, who relayed this information to me several months ago by telephone. Readers will note that Rosenthal's commentary also questioned this R_i formula. Finally, in Table 6, the missing R_i for combined data (.32) has now been inserted, and the R_i for NSF and COSPUP reviews of proposals in *Chemical Dynamics* should be .16 rather than .12.

Next examined are the more formal and involved criticisms of the methodologic, statistical or data analytic techniques presented in the target article.

1.2. Interpreting levels of kappa and R_i. There is concern on the part of Eckberg about the presumed arbitrariness of the Cicchetti & Sparrow (1981) strength of agreement values for kappa and intraclass correlation coefficients, namely: POOR (below .40); FAIR (.40–.59); GOOD (.60–.74); EXCELLENT (.75 and above).

These values are similar to those provided by Fleiss (1981), although he uses a wider range to encompass values between .40 and .74 (designated as FAIR to MODERATE). Earlier, Landis and Koch (1977) proposed six evaluative categories: less than zero = POOR; 0–.20 = SLIGHT; .21–.40 = FAIR; .41–.60 = MODERATE; .61–.80 = SUBSTANTIAL; .81 and above = ALMOST PERFECT (see also Feinstein 1987, p. 185). These examples show the similarity of guidelines research biostatisticians recommend to differentiate mere statistical significance (kappa or R_i larger than 0) from significance that may be of practical or clinical usefulness, as well. The general concept is analogous to Cohen's (1988) suggested effect sizes (ES) for interpreting sample correlation values (i.e., an R_i of .15 representing a SMALL, .30 a MEDIUM, and .50 a LARGE effect, when compared to expected values of zero).

More important, these guidelines are consistent with the frequency with which high and low kappa values are reported for many clinical phenomena. Koran (1975a; 1975b) has shown that when kappa is used to assess interexaminer reliability levels of the presence or absence of a wide range of clinical signs and symptoms, values rarely exceed .70.

Concerning the application of these guidelines, **Eckberg** questions the plausibility of a specific hypothesis (sect. 4.5), namely, that if a formal study were conducted on the reliability of peer reviews for manuscripts submitted to *Physical Review Letters (PRL)*, it would be the same order of magnitude (e.g., R_i below .40) that characterizes general journals in many other disciplines. Given that an average of five or more *PRL* reviewers is required to arrive at consensus, coupled with a 45% rejection rate (Adair & Trigg 1979, sect. 4.5), I would consider the hypothesis reasonable rather than what Eckberg characterizes as "pure speculation."

1.3. Choice of statistical tests. It was suggested by **Gilmore** and **Rosenthal** that other statistical tests may have been at least as appropriate as the ones that were used.

Gilmore prefers his "shared uncertainty index" (Gilmore 1979) to the kappa and R_i approaches described in the target article. For the reasons cited in section 3.3, I do not share his preference. In the list of multiple and unique advantages of kappa over any and all of its competitors, I would add that:

a. Kappa has been widely generalized to fit (1) varying scales of measurement (Cicchetti 1976; Cohen 1968); different types of rater and subject reliability research designs, for example; (2) 3 or more raters (Fleiss 1971; Fleiss et al. 1979; Landis & Koch 1977); (3) differing numbers of raters per subject (Fleiss & Cuzick 1979); (4) multiple diagnoses per patient (Kraemer 1980; Mezzich et al. 1981); (5) multiple observations on small numbers of subjects (Gross 1986); (6) single subject reliability assessments (Kraemer 1979); (7) separate reliability assessments for each category on a given clinical scale (Cicchetti 1985; Cicchetti, Lee et al. 1978; Spitzer & Fleiss 1974).

b. Other generalizations include those in which (8) rater uncertainty of response is the focus (Gillett 1985); (9) the rating categories have not been defined in advance (Brennan & Light 1974; Brook & Stirling 1984); (10) multiple raters are analyzed pair by pair, when each pair rates the *same* set of subjects (Conger 1980) or *different* sets of subjects (Uebersax 1981; 1982); (11) the data are continuous with a focus on the *duration* rather than the *frequency* of joint events (Conger 1985); (12) jackknifing functions are used to reduce bias in estimating standard errors of kappa (Davies & Fleiss 1982; Kraemer 1980).

c. Kappa has also been (13) subjected to a number of empirical studies for testing and confirming or modifying the way it can be applied appropriately (e.g., Cicchetti 1981; Cicchetti & Fleiss 1977; Fleiss & Cicchetti 1978; Fleiss et al. 1969; 1979).

d. Kappa (nominal data) and weighted kappa (ordinal data) have been shown under certain specified conditions to be (14) equivalent to various models of the intraclass correlation coefficient (R_i) (e.g., Fleiss 1975; 1981; Fleiss & Cohen 1973; Krippendorff 1970; Shrout et al. 1987). Finally,

e. Kappa and kappa-type statistics have also been used in conjunction with a number of multivariate approaches to reliability analysis: (15) cluster analysis (Blashfield 1976); (16) signal detection models (e.g., Kraemer 1988); (17) latent structure agreement analysis (Uebersax & Grove 1989); and (18) latent structure modeling of ordered category rating agreement (Uebersax 1989); and (19) Kraemer (1982) has shown, in the 2 x 2 case, the relationship between kappa values and the sensitivity and specificity of a given diagnostic procedure.

Rosenthal writes, in the 2 x 2 case, of three "more-information-efficient" indices, kappa, R_i, and the standard Pearsonian product moment correlation (R), or the phi coefficient. He describes these indices as mathematically equivalent for that reliability research design in which the same two examiners independently evaluate all subjects (or objects). Rosenthal prefers their usage to three "less-information-efficient" statistics, namely, "rate of agreement" or what Rogot and Goldberg (1966) refer to as the "crude index of agreement" (uncorrected for chance); chi square(d); and unweighted kappa for 3 x 3 and larger tables. I agree with some of Rosenthal's conclusions.

First, kappa, R_i, and R (or phi) will be identical only when marginal frequencies or category assignments are identical for each of any two independent reviews. For peer review, if the acceptance (approval) and rejection (disapproval) rates are the *same* for both independent sets of reviews (e.g., 20% acceptances and 80% rejections),

then the data in the resulting 2 x 2 or four-fold table will produce *identical* results, whether one applies kappa, R_i, or phi (e.g., see Cicchetti 1988; Cohen 1960; Fleiss 1975; 1981). The example cited (the reviews for manuscripts submitted to the *Journal of Abnormal Psychology, JAP*, Footnote 6 of the target article) illustrates this equivalence, as Rosenthal correctly notes. This occurs because there is no intuitively obvious way to distinguish "first" reviews from "second" reviews. Therefore, the required Model I R_I that is applied to the data will produce equal rater marginals (category assignments to "accept" and "reject") for the two independent sets of reviews. In such a situation, the three mathematical formulae (for kappa, R_i, and phi) become equivalent. These identities also hold in the Model II case (same two raters throughout) providing, again, that the category assignments are identical. When these assignments are not identical (the much more usual case), Kappa, R_i, and phi (or R) will assume different values, the difference depending on specific distributions of the two category assignments.

As an example of the effect of unequal category assignments on the values of kappa, R_i, and phi, consider the data presented in Table 6 (target article). Here there was interest in distinguishing two identifiable sources of *average* ratings, namely those made by NSF and those made by COSPUP. The full data on which the condensed Table 6 entries are based, for the area "Economics," are shown in Table 1: Here, R_I (Model II) = Kappa = .44. If we had instead considered that the distinction between NSF and COSPUP ratings are not of concern and used R_I (Model I), which would take into account that different pairs of reviewers viewed different proposals, its value would be .38. In either case, R (or phi) would = .41. Thus, Kappa, R_i, and R are identical when category assignments are identical) but *not* under any other combination of category assignments (the more usual case).

Concerning Rosenthal's second point, I would agree that chi square(d) should not be used as a measure of examiner agreement, for the reasons he cites, as well as because chi square(d) measures associations of any type, whereas kappa and R_i measure agreement per se. I would partially agree with Rosenthal's caveat about applying unweighted kappa as an omnibus statistic to 3 or more categories of interest. Although the overall value of kappa might be of somewhat limited interest, the decomposition of kappa into levels of specific agreement (observed and chance-corrected) on a category by category basis, would, in fact, be quite informative (e.g., Fleiss 1981, p. 220). For peer review, there might be interest in the extent to which reviewers agree on such conceptually distinct evaluation attributes (nominal variables) as: importance of the problem under investigation; adequacy of research design; and interpretation of research results. Each evaluative attribute could be scored as "acceptable" or "unacceptable." If the reliability design were such that the same two reviewers evaluated all submissions independently, then the generalization of kappa developed by Davies and Fleiss (1982) would apply. If the reviewers varied from one submission to another, then the kappa statistic developed by Fleiss (1971) and extended by Fleiss et al. (1979) would be relevant. Again, while the overall (omnibus) kappa value averaged over the 3 categories of interest might be of limited value, the levels of observed and chance-corrected agreement on each evaluative attribute would be quite meaningful. On the other hand, if the overall kappa value were not even statistically significant, one would be less interested in the specific category reliability assessments. For these reasons, and the ones expressed in my reply to Gilmore, I would conclude that kappa is more "information-efficient" than its competitors.

Table 1. *Average NSF and COSPUP ratings of 50 proposals in the field of "Economics"*

	COSPUP:		
NSF:	Low Ratings (10–39)	High Ratings (40–50)	All Proposals
Low (10–39)	29	3	32
High (40–50)	9	9	18
All Proposals	38	12	50

Finally, with respect to Rosenthal's application of kappa to the acceptance and rejection figures for the *JAP* data given in Table 5, my two values are .14, as is true for overall kappa (again, the 2 x 2 equal marginals case). Although these values, as well as the 70% and 40% agreement levels, are describing the same data, each conveys valuable, though different, information, as explained more fully in my upcoming replies to Demorest and Wasserman.

Reanalyzing data from Tables 5 and 6 respectively, Demorest and Wasserman arrive at the same conclusion, namely, that chance-corrected agreement on rejection (disapproval) is no better than on acceptance (approval). They are both right. The phenomenon, as Demorest correctly notes, however, is specific to degrees-of-freedom limitations inherent in data deriving from a 2 x 2 contingency table. As noted in my discussion of Rosenthal's commentary, overall kappa values are always mathematically identical to specific kappa values for acceptance and rejection (e.g., see also Cicchetti 1980; Cicchetti & Feinstein 1990; Fleiss 1975).

A very important and relevant issue, however, discussed neither by Demorest and Wasserman, nor by the target article itself, still needs to be addressed. As noted recently (Cicchetti 1988, p. 621), the same kappa value can be reflected in a wide range of observed agreement levels. Some will be of substantive (practical or clinical) value and others will not. It thus becomes necessary to set some specific criterion for judging the usefulness of both observed and chance-corrected levels of agreement as they may occur together. My colleagues and I have suggested that one should require a minimum level of agreement of 70% *before* correcting for chance, and an accompanying level of at least .40 ("fair" agreement) *after* correcting for chance (see Volkmar et al. 1988, p. 92). If we apply these criteria to the data presented by Demorest, in Table 2, namely, category-specific agreement levels for reviews of manuscripts submitted to the *American Psychologist*, the only category that meets these standards is category 5 ("reject"), for which the observed level of reviewer agreement is 75.9% and the chance-corrected level (weighted kappa) is .52. Consistent with these results, reviewer agreement levels on 866 manuscripts submitted to a purposely unidentified Major Sub-

Table 2. *Category-specific agreement levels for 866 submissions to a Major Subspecialty Medical Journal*

Reviewer Recommendation	Average Frequency of Usage (%)	Type of Agreement		Corrected for Chance
		Observed (%)	Chance (%)	
Accept/Excellent	5	53	34	.29
Accept/As Is	7	65	53	.27
Accept/Revise	21	78	68	.33
Resubmit	24	78	75	.12
Specialty Journal	10	74	72	.08
Reject	33	81	66	.44
All Recommendations	100	77	67	.30

Note. Weighted kappa (Cohen 1968; Fleiss, Cohen & Everitt 1969) was used with a weighting system developed and recommended by Cicchetti (1976); Cicchetti & Fleiss (1977); and Cicchetti & Sparrow (1981), in which: complete reviewer agreement is assigned a weight of 1, followed by disagreement which is one ordinal category apart (.8), two categories apart (.6), three (.4), four (.2), and five categories apart (0, i.e., "Accept/Excellent" vs. "Reject Outright"). The corresponding R_i value for these data was .37, which was shown in Table 2 of the target article.
Source: from Cicchetti & Conn (1976).

specialty Medical Journal (Cicchetti & Conn 1976) are shown in Table 2. The only reviewer recommendation category that meets the Volkmar et al. (1988) criterion is "reject," with an observed rate of agreement of 81% and a chance-corrected level of .44.

In summary, the data indicate that the accompanying levels of observed agreement are substantially higher for negative than for positive evaluations and that the phenomenon holds, more generally, in the 3-or-more category case where there are varying levels of chance-corrected agreement possible.

1.4. Interpreting the data in Tables 1, 2, 5, and 6. The question is raised by Eckberg why the numbers of manuscript reviews for the *Journal of Abnormal Psychology (JAP)* and the *Journal of Personality and Social Psychology (JPSP)* vary from Table 1 to Table 2. For *JPSP* manuscripts, the two samples were different ones. The *JAP* data in Table 2 are based on a complete sample of 1,319 manuscripts submitted between 1973 and 1978. They focus on overall reviewer recommendations (scientific merit). The data in Table (target article) 1 are based on evaluation criteria (deriving from specific rating forms) that reviewers applied to *JAP* manuscripts submitted between 1976 and 1978. For the approximately 50% of the remaining manuscripts (1974–1975), these rating forms were unavailable for reviewers. To clarify this issue in Table 1, row A now reads: "For manuscripts submitted to the *Journal of Abnormal Psychology* (1976–1978)," rather than (1973–1978).

Referring to the data presented in Tables 5 and 6 (target article), Eckberg wonders why I conclude that reviewers agree more on rejection than acceptance, rather than that reviewers simply reject more often than they accept. He also wonders whether the chi square(d) values in Tables 5 and 6 are incorrect.

Concerning the first question, the data do, in fact, indicate substantially more agreement on rejection than on acceptance. This phenomenon is conceptually independent of the fact that reviewers recommend rejection much more often than acceptance. Take the data for *JAP* (first entry of Table 5). Of the 462 manuscripts that received positive reviewer recommendations, how many were in agreement? This is 44%, or 203. For those 857 manuscripts receiving negative recommendations, however, there was agreement on 70%, or 600. The question raised here is simply whether there is significantly more agreement on rejection than on acceptance. The chi square(d) value of 83.99 means that the difference is statistically significant at beyond the .00001 level.

The figures reported in both Tables 5 and 6 are all correct as they are reported in the target article. Two factors will cause chi square(d) values to vary, however. The most obvious (and least important) pertains to how many places beyond the decimal point are considered. This produces differences from simple rounding errors. The conceptually more serious source of variation arises from whether the chi square(d) test (here with 1 degree of freedom) is applied with or without the Yates (1934) correction factor. Fleiss (1981) argues correctly (p. 27) that "because the incorporation of the correction for continuity brings probabilities associated with χ^2 and Z into closer agreement with the exact probabilities than when it is not incorporated, the correction should always be used." Soper et al. (1988) demonstrated in a recent computer simulation that the random application of the chi square(d) test to neuropsychological data resulted, as expected, in values that were indistinguishable from nominal or chance levels (e.g., .05 or .01) when the continuity correction was used. When it was not, many more significant chi square(d) values were produced than were warranted by the data. These results support Fleiss's arguments and are also consistent with the earlier recommendations of Delucchi (1983, p. 169) and of Lewis and Burke (1949), much earlier.

Given the necessity of using the correction for continuity, what effect would its nonusage (albeit incorrect) have on the chi square(d) and p values shown in Tables 5 and 6? These range from trivial to substantial depending on the size of the continuity-corrected chi square(d) value and the number of cases on which the test is based. Thus the chi square(d) value for *JAP*, based on 1,319 cases, increases to 85.08, which, "p-wise," is indistinguishable from the reported continuity-corrected chi square(d) val-

ue of 83.99. In distinct contrast, the continuity-corrected chi square(d) value of 3.413 ($p = .06$), for the 72 manuscripts submitted to *Developmental Review* (entry 3 of Table 5) increases to 4.46 ($p = .02$), when the correction for continuity is not used. Similar effects can be noted for the data in Table 6.

Eckberg asks two additional questions: (1) In the case of comparing NSF and COSPUP open reviews (Table 6), how was it decided who would be the two reviewers? Each average COSPUP rating for a given grant proposal (first "reviewer") was compared to each average NSF rating (second "reviewer"). (2) Why is the number of disagreements exactly the same in both the "Acceptance" and "Rejection" columns (Table 5) and in the "High" and "Low Ratings" columns (Table 6)? This is because the disagreed-on cases for acceptance and rejection cannot differ in the 2 x 2 case, because of degrees of freedom restrictions (see also Cicchetti 1988, Tables 6–10, pp. 611–615, and p. 619).

1.5. Interpreting the data in Table 3. Based on experience with behavioral psychology journals, Cone notes that journals with lower submission rates will tend to have higher acceptance rates. Therefore this variable needs to be controlled in peer review research. He concludes that the data presented in Table 3 (target article) provide partial support for this notion in the case of manuscripts submitted to the *Physical Review (PR)*. For example, the Nuclear Physics section of *PR* has a higher acceptance rate and lower submission rate than those sections with two or three times as many submissions as Condensed Matter or General Physics.

A more comprehensive analysis of these data do *not* support Cone's contention. Thus, the two sections with the lowest submission rates, Nuclear Physics and Particles & Fields, with a combined submission rate of 31.5% (or 1658/5264), have a combined acceptance rate of 73.3% (or 1215/1658). There is a similar combined acceptance rate of 75.3% (or 2717/3606) for the two sections (General Physics and Condensed Matter) with more than twice the percentage of submissions (3606/5264 or 68.5% vs. 1658/5214 or 31.5%). Chi square(d), corrected, 1 df = 2.35 (p = n.s.). More important, the strength of association (Effect Size (ES), Cohen 1988) between manuscript submission rate and acceptance rate, as measured by phi (or χ^2_{1df} (uncorrected)/N) is only 0.02, a zero-order effect.

In a related issue, pertaining again to the type of data presented in Table 3, Cone contends that there is no evidence for my assertion that "manuscripts requiring more than one reviewer tend to be those that are problematic." This is based on a misunderstanding about how the single initial referee system works. In the field of physics (e.g., *Physical Review, PR*), the editor sends a manuscript initially to a single reviewer. If the reviewer recommends acceptance, the editor typically supports that decision. Only when the initial referee detects a problem (i.e., recommends rejection) is the manuscript sent to a second referee. If the second referee also recommends rejection, then the editor typically rejects the article. If the second reviewer recommends acceptance, however, then the paper is viewed as "problematic." Such a manuscript is usually sent to a third referee who will decide the fate of the submission (see also Hargens 1988).

Kiesler's comments about "explaining" differences between natural and behavioral scientists in terms of their "success" with manuscript or grant applications seem confused, so I am unable to respond. They presumably have something to do with the data presented in Table 3, but I simply can not follow his arguments. Clarification in *BBS* Continuing Commentary is suggested.

The next several sections of my Response focus on varying interpretations of the overall results presented in the target article, namely, that across disciplines and type of submission (manuscript, grant) levels of interreferee agreement (corrected for chance) tend to be rather low (R_i usually below .40).

2. Interpretation of the results

2.1. Reliability levels are correct as reported. A majority of commentators accepted the low levels of reliability as valid, and offered a number of suggestions for improving the reliability (and at times even the validity) of peer reviews (**Adams, Bornstein, Cohen, Cole, Colman, Cone, Crandall, Delcomyn, Fletcher, Gilmore, Gorman, Greene, Kraemer, Laming, Lock, Mahoney, Nelson, Roediger, Rourke, Salzinger, Tyrer, and Zentall**). These views are discussed in later sections of the report.

Cole feels that both editors and granting officials need to admit that since reliability is so poor, much high quality research is rejected or disapproved, whereas some poor quality research is accepted or funded. Therefore, editors should gradually increase the number of manuscripts they accept and granting officials should put funding aside for meritorious but disapproved proposals. The major problem with this otherwise good idea is that the time required to reverse a funding decision may equal or exceed the time required to revise the proposal and resubmit it to the same or a different funding agency. **Zentall, Roediger,** and **Laming** doubt that levels of reliability could ever be improved substantially. Zentall argues that much of the disagreement reflects deep theoretical and methodological (confirmational) biases. Similarly, Roediger argues that the corpus of psychological literature has demonstrated consistently that human judgments of such complex issues as hiring decisions or making clinical diagnoses, are of questionable reliability and validity. Hence, the similarity in results for peer reviews is to be expected. Laming, in a most imaginative commentary, argues by analogy with the results of a number of psychophysical studies across sensory modalities that the constantly shifting frames of reference with which successive stimuli are compared limit the accuracy of human judgments to the extent that about 2/3 of the variability in judgments can be attributed to the variability in frames of reference. Thus, it is the absence of a stable frame of reference that sets limits on the extent of judgmental accuracy. Applying this knowledge to the field of peer reviews of manuscript and grant submissions, Laming concludes that the shared variability between independent reviews would be restricted to an upper limit of about 0.33. He ends his commentary on a rather sombre and pessimistic note that he contrasts to my own more optimistic view of progress in science (in general) and peer review (in particular). Laming's pessi-

mism is based on his examination of journal articles in his field of interest (experimental psychology) that were published between 50 and 100 years ago. He concludes that if more than 90% of this research had never been published, the state of experimental psychology would be no different than it is today.

In contrast to the pessimism shared by **Laming, Roediger,** and **Zentall,** I must state emphatically that the progress in my own field of inquiry, assessing the reliability and validity of standard and state-of-the-art diagnostic instruments in both behavioral science and medicine, has been nothing short of dramatic. Thus, my colleagues and I have developed highly reliable and valid instruments over a wide range of disorders:

1. In behavioral science, for example, adaptive behavior (Sparrow et al. 1984a; 1984b; 1985), alexithymia (Krystal et al. 1986), personality disorders (Cicchetti & Tyrer 1988; Tyrer, Cicchetti et al. 1984; Tyrer, Strauss et al. 1984), anxiety (Tyrer, Owen et al. 1984), affective behaviors of demented patients (Nelson et al. 1989), and dissociative disorders (Steinberg et al. 1990); and

2. In medicine, the Yale Observation Scales for identifying seriously ill febrile children (McCarthy et al. 1982; McCarthy et al. 1990), new methods for classifying cataracts both in vitro (Cicchetti et al. 1982); and in vivo (Cotlier et al. 1982), and accuracy of the barium enema in diagnosing (a) Hirschsprung Disease (Rosenfield et al. 1984), and (b) acute appendicitis (Garcia et al. 1987).

For each of these diverse areas, we have consistently shown levels of reliability in the GOOD to EXCELLENT range (usually kappa or R_i values .90 and above), as well as good evidence for validity. When I became actively involved in research more than two decades ago, there was little optimism that the low levels of reliability and accuracy of judgment (especially in the behavioral sciences) would ever become "respectable." Yet, less that a decade ago, the field of psychiatric diagnosis had improved dramatically as encapsulated in the writings of Grove et al. (1981, p. 408):

> For years, achieving adequate diagnostic reliability in psychiatry was considered to be a hopeless undertaking. A number of landmark studies suggested that psychiatrists looking at the same patients frequently disagreed about the appropriate diagnoses. As a consequence, the importance of diagnosis was minimized in both research and clinical work. . . . The reversal of nihilistic attitudes about psychiatric diagnosis has led to a rigorous (and successful) attempt to rework the entire American diagnostic system used by clinicians, DSM-III, which demonstrated in field trials that good agreement could be achieved even in routine practice.

The specific details about how I believe that similar breakthroughs can be made in the field of peer review (namely, improving both its reliability and validity) are expressed in a later section of the report.

2.2. Reliability levels are worse than indicated. Examples are given by **Schönemann** from the published literature in which false claims about a number of phenomena have been made and perpetuated (e.g., indeterminacy, heritability, the results of mathematical modeling). Because manuscripts with high reliability (editors and reviewers agree they should have been published at the time) have "negative" validity, this can only mean that reliability is lower than one would think, perhaps at random or chance levels.

Although **Schönemann's** argument has a certain face-validity appeal, I am hard pressed to calculate the actual frequency with which the unfortunate phenomena he reports occur relative to the mammoth corpus of research that has been published. In mathematical terms, we are faced with trying to interpret a ratio with both unknown numerator (the number of invalid published research findings) and unknown denominator (the total number of nonredundant published findings). In short, it is not possible for me to draw a cause and effect conclusion on these matters given the data presented thus far. Perhaps, given the enormousness of published research in such diverse outlets, one could never arrive at a valid conclusion.

2.3. Reliability is better than indicated. Several commentators (**Hargens, Marsh & Ball**) mention that reliability levels may have been underestimated by taking into account only the recommendations of two independent reviewers. **Marsh & Ball,** for example, note that in addition to the initial two reviews, the editor often has his own review, author revisions, and further reviews of the revised manuscript on which to base a decision, thereby probably increasing the reliability of the process. The additional review, however, whether by the editor or a third reviewer, is often not an independent one and so may be heavily influenced by the results of the initial two reviews. Despite this problem, there is a factor mentioned by both **Hargens** and by **Marsh & Ball** that one can test empirically, namely, that the editor's process of weeding out very poor quality manuscripts (rejected without being sent out for review) might reduce the variance and subsequently increase the levels of inter-reviewer agreement, because these very submissions are the type we have shown to produce the highest levels of consensus. **Hargens** cites both Gordon (1977) and Zuckerman and Merton (1971) to suggest that the editor's sole "summary-rejection" rates for prestigious journals in both social science and medicine may reach levels as high as 50%. Fortunately, I have been able to analyze further some additional data deriving from reviews for the *Journal of Abnormal Psychology (JAP)* during 1973 and 1977. As given in Table 3, and based on 996 submissions, there was an overall R_i (or kappa) value of .24 with 73% agreement on rejection, 51% on acceptance, and 65% overall agreement. In addition to these 996 submissions, the editor received 384 additional manuscripts. He rejected 333 (86.7%) and accepted the remaining 51 (13.3%). If we make the assumption that the rejected manuscripts would also have been rejected by another independent reviewer because of their obvious poor quality or inappropriateness for *JAP,* the results show that: Overall agreement increases from 65% to 74%; agreement on rejection increases from 73% to 82%; agreement on acceptance remains at 51%; and R_i (or kappa) increases from .24 to .34. In conclusion, even if one assumes that the reliability of negative editorial reviews is perfect, it may not have a profound effect on increasing the reliability of the peer review process. Thus, whereas the agreement level on rejection improves, the lack of a corresponding increase in reliability for acceptance keeps the R_i value at relatively low levels.

Table 3. *Effect of editorial summary rejection of 333 manuscripts on the overall reliability of peer review of manuscripts submitted to Journal of Abnormal Psychology (1973–1978)*

A. Based on two independent reviews

First Review	Second Review: Accept	Reject	Total
Accept	181	172	353
Reject	173	470	643
Total	354	642	996

$PO_{(overall)}$ = 65.4% $PO_{(accept)}$ = 181/353.5 = 51.2%
$PC_{(overall)}$ = 54.1%
Kappa (or R_i) = .24 $PO_{(reject)}$ = 470/642.5 = 73.2%

B. Adding the 333 editor's rejections to the reject-reject cell

First Review	Second Review: Accept	Reject	Total
Accept	181	172	353
Reject	173	803	976
Total	354	975	1329

$PO_{(overall)}$ = 74.0% $PO_{(accept)}$ = 51.2%
$PC_{(overall)}$ = 60.9% $PO_{(reject)}$ = 82.3%
Kappa (or R_i) = .34

The great majority of commentators viewed the target article as a worthwhile endeavor, although they differed on their specific interpretation of what the results mean; two remaining commentators, however, **Kiesler** and **Bailar**, questioned the value of such research. These two commentators share the minority view that the only meaningful goal of peer review is to improve decisions about which submissions should be accepted (or approved) and which should be rejected (or disapproved). As such, the issue of reliability is essentially irrelevant to them. They also express the view that high levels of agreement signal that there is too much redundancy in the peer review process, that it is not working well, and that a balanced review has not been achieved.

Kiesler is convinced at a basic conceptual level that high levels of reliability are incompatible with what he terms "wise" editorial and funding decisions. He states specifically that to expect high levels of reviewer agreement is "naive" because it falsely assumes that reviewers are randomly drawn by editors. I would submit that herein lies the most serious error in Kiesler's reasoning. In fact, if he were to choose reviewers randomly in his own general area of focus (the broad field of psychology), this procedure would almost guarantee levels of reviewer agreement even *lower* than what has been reported. Given that Kiesler needed a Freudian theorist as well as a sophisticated statistician to obtain a balanced review (using his hypothetical example), the probability that such expertise could be obtained on the basis of purely random selection procedures would indeed approach zero. In fact, any set of reviewers selected at random in any general focus area (behavioral science, medicine, general subfields of physics) would almost perforce, be expected to disagree to a greater extent than those chosen specifically for their areas and levels of expertise. **Rourke** correctly intimates that the validity of the comments of randomly selected reviewers would also be comprised because of insufficient knowledge about the area they would have been asked to evaluate. (A similar view is expressed by **Lock**.) In short, the balanced selection of reviewers should, if anything, enhance both the reliability and the validity of the resulting reviews.

If we accept **Bailar**'s commentary at face value then to expect the peer review process to be "reliable," "fair," and "objective" would be considered an "inappropriate" goal. A careful reading of Bailar's comments suggests that as an editor he chose to work around the obvious unreliability, unfairness, and subjectivity of the peer review process for the *Journal of The National Cancer Institute (JNCI)*. As one example, his regular use of reviewers who were clearly biased (i.e., would never recommend publication or would never criticize their colleagues) would prompt other commentators to act quite differently (I agree). Thus **Kraemer** would remove reviewers who "condemn everything" or have an apparent conflict of interest with the author(s) of the paper under review. Similarly, other commentators would rather remove than live with or "work around" other obvious biases in the peer review system (I again agree). These biases include "confirmatory bias" against "negative" research findings; well-conceived replication studies (**Gorman, Lock, Salzinger, Schönemann, Zentall**); innovative research (**Armstrong & Hubbard, Lock**); the time of day that grants are evaluated, subjective "rating scale use habits" of grant reviewers, and the hypothesized harsher (more negative) evaluations provided by less experienced grant reviewers (**Cohen**).

In summary, for **Bailar** to allow individuals who are clearly biased or who may have a potential conflict of interest to remain as "regular" reviewers stretches to the breaking point my limits of permissible peer review practices. Consistent with the views of peers at large, I am totally opposed to the practice. It is also somewhat curious that Bailar voices concern that ethical issues were not discussed in the target article. His comments follow closely his voicing obvious frustration with not being able to discuss such issues directly in connection with the Peters & Ceci (1982) publication about eight years ago. The fact of the matter is that about 20% of the authors' reply was devoted to the ethical issue. **Mahoney** addresses the ethical issue more broadly and I endorse his sanguine remarks heartily.

Another issue that both **Kiesler** and **Bailar** seems to have overlooked is that high quality research (worthy of support) is integrally related to: (a) asking important questions; (b) designing and executing the research in an exemplary manner (utilizing proper controls); (c) using state-of-the-art instrumentation (and/or test materials); (d) writing clearly and succinctly; and (e) presenting a compelling discussion of the results and their implications (or heuristic value) for furthering scientific advancements in the field. Because of the interrelatedness of these five evaluation attributes, my many years of experi-

ence reviewing manuscripts and grants over a broad spectrum of disciplines (behavioral science, medicine, biostatistics), as well as my activities on editorial boards and grant review committees, have indicated to me that when the peer review process is working properly (i.e., reviewers are selected for their varying areas of competence and they take their reviews seriously) it is not unusual to find high levels of agreement on at least the final recommendation, if not on a number of manuscript or grant attributes as well.

To clarify the relevance of **Bailar**'s example, there is no a priori reason to believe that the cardiologist, pharmacologist, and statistician should not agree that a given clinical trial evaluating a new hypertensive drug is or is not worth supporting simply because each represents a different area of expertise. They would surely agree more than alternative reviewers selected randomly. The major disagreements I have experienced (or witnessed) among reviewers (whether for manuscripts or grants) have occurred primarily because a proper match was *not* made between submitters and reviewers. Although the disagreement can be occasioned by a number of factors, not least among them is a lack of sufficient expertise (or even bias) on the part of one or more of the reviewers.

So, in response to both **Kiesler** and **Bailar**, I would emphasize that the proper selection of reviewers to evaluate a given submission, should, in the long run, increase both the reliability and validity of the peer review process. The sine qua non necessity of obtaining a balanced set of reviews (for both manuscript and grant submissions) is widely accepted by editors, granting officials, reviewers and authors alike. See, for example the additional comments on this important issue by **Adams, Eckberg, Greene, Hargens, Kraemer, Roediger,** and **Stricker,** as well as the recently published work of Fiske and Fogg (1990).

The next major issue I discuss concerns how a given editor or program director uses the information obtained from peer reviews – quite apart from issues of reliability (or validity) – to make publication or funding decisions.

3. Use of peer reviews to improve editorial/funding decisions

3.1. The editor as final arbiter. A general concern is expressed by **Fletcher** about how editors and granting officials use unreliable reviewer recommendations to arrive at publication or funding decisions. **Stricker** talks about the importance of the "active" editor who judiciously "weighs" the information, provided by reviewers, to arrive at a thoughtful publication decision. Similarly, **Rourke** speaks of "fair," "judicious," and "experienced" editorial practices. The reader will also recall **Kiesler**'s concept of the "wise" editor. These descriptions, in turn, are similar to **Bailar**'s informative notion of the editors' integration of their own knowledge with that provided by the additional "wisdom" of members of the editorial board, as well as "special consultants," as required.

Lock (see also Lock 1985) proposes a "hanging committee" to examine and help resolve questions about those manuscripts receiving "gray area" or split-review recommendations. Both **Bailar** and **Crandall** speak of the need for editors to eschew a majority-vote-of-reviewers' option, by exercising their power to override the recommendations of reviewers, whenever required. **Bailar** states that a neglected area of the target article is the realization that an editor's decision is based not only on the overall scientific (or "technical") merit of a given submission but also on such manuscript attributes as "originality," "importance to readership," or "succinctness." I am somewhat puzzled at how an editor accomplishes this important goal, given the known unreliability of such attributes, that is, the data shown in Table 1. Bailar's further elucidation of how he was able to accomplish this objective for *JNCI* would make an important contribution to the field of peer review. **Tyrer** discusses the important role of the editor or chairman of a granting agency in which "executive decision" is used to resolve reviewer disagreements. In a slightly different context, **Armstrong & Hubbard** note that at least some innovative research is published by editors of the *American Psychological Association (APA)* journals, despite the relatively low levels of reviewer agreement on submissions describing such valuable research. **Fuller** expresses concern that such high quality research will frequently be published in relatively few access journals.

In a more general sense, **Bailar** takes the position that the target article did not adequately "pound home" the relatively major role of the editor or granting official in the entire peer review process relative to what he feels is the more minor role played by the reviewers (merely the providers of "relevant information"). Unfortunately, in so doing, the views of Bailar (consistent with those of **Kiesler**) create a purely artificial distinction among authors, reviewers, editors, and consumers of submitted and published papers. As noted earlier (Cicchetti 1982, p. 21), "one of the most persistent problems we still face appears to be the false dichotomy we have tended to create between those who evaluate research and those who are being evaluated. Both derive from the same research species." The provocative and imaginative commentary by **Kraemer** is consistent with this view. She speaks eloquently of the conflicting roles each of us is called on to play ("submittors," "reviewers," "consumers" of scientific papers) and the fundamentally different standards we might invoke, depending on which peer role we are assuming at a given moment. Her emphasis on an "objective," "dispassionate," and "quantitative" approach to the study of peer review as the only hope of identifying and correcting the many shortcomings in the peer review process is to be taken seriously. Toward this important goal, it is almost axiomatic that for science to continue to operate, it requires the imagination and talent of authors, dispassionate and sensitive editors and granting officials, and, finally, interested readers (or consumers) of the published research findings, so that the cycle can continue anew. In this basic and comprehensive sense I disagree with the more narrow-focus views of both Bailar and Kiesler.

Though **Bailar** seems to be unaware of it, data *were* provided in the target article, showing the positive relationship between peer reviewer recommendations and the publication decisions made for more than 1,300 manuscripts submitted between 1973 and 1978 to the general focus *Journal of Abnormal Psychology (JAP)*. It was further noted that the results were consistent with data deriving from reviews of both the *Sociological Review* and the *Physical Review*, which indicate that re-

viewer recommendations are the major factor in the editor's decision to accept or reject (Bakanic et al. 1987; Hargens 1988).

Next is an attempt to broaden our knowledge base in this important area of peer review research. The data is based on the total number of submissions to *JAP* between 1973 and 1977 (1,698 manuscripts). These could be classified as follows: 15 (0.9%) were withdrawn before an editorial decision could be made; 14 (0.8%) were solicited; 384 (22.6%) were reviewed by the editor alone; 175 (10.3%) were reviewed by a single referee (other than the editor); 996 (58.7%) were evaluated by two independent reviewers; 112 (6.6%) received three independent reviews; and 2 (0.1%) received four independent reviews.

As reported in section 2.3 of this Response, 86.7% (333 of 384 manuscripts) were reviewed and summarily rejected by the editor alone on the basis of very poor quality, inappropriateness for the readership of the journal, or on both accounts. All 14 of the solicited manuscripts were accepted for publication. Of the two manuscripts receiving four reviews, one was accepted, the other rejected.

The fate of the remaining manuscripts, namely, those with a single review, two reviews, or three reviews follows.

3.2. The editor's use of single reviews: "Go with the flow." Table 4 shows that of the 175 manuscripts sent to a single reviewer, 58 (33.1%) were accepted by the editor and 117 (66.9%) were rejected.

The full set of data (Part A of Table 4) indicates that, as a general rule, the editor's final decision closely parallels single reviewer recommendations. Part B of the table indicates that when the reviewer recommended that the manuscript be accepted ("as is" or "subject to revision") there was an 85% likelihood of acceptance. Analogously, when the reviewer recommended either resubmission or rejection, there was a 90% probability of rejection. An inspection of discrepancies between reviewer recommendations and editorial decisions indicated that the editor was no more likely to reject manuscripts receiving an "accept/as is" or "accept/revise" recommendation (8/54 = 14.8%) than to accept manuscripts receiving a "resubmit" or "reject" recommendation (12/121 = 9.9%). Applying McNemar's (1947) statistic for correlated proportions produced a chi square(d) value of zero-order significance, that is, 0.45.

3.3. The editor's use of two reviews: "Go with the low." Results are presented in Table 5 for those 996 manuscripts receiving two reviews during the period 1973–1977.

The data can easily be understood if one considers that: (1) Manuscripts receiving a joint reviewer recommendation of "Resubmit" had a 27% probability of being accepted for publication, which is indistinguishable from the base rate journal acceptance rate of 28%; (2) those manuscripts receiving two reviewer votes for acceptance *or* a split between acceptance and resubmission had a 72% probability of being published, as compared to the 72% baseline rejection rate of the journal; and (3) the remaining 635 manuscripts (65%) had only a 5.5% probability of being accepted for publication, a rate more than five times *less* than the journal baseline acceptance rate of 28%.

These findings also have cross-disciplinary implications. Specifically, Lock (1985, pp. 20–21) presented analogous data for 282 articles, or 50% of the 564 articles that were submitted to the medical journal *Thorax* and evaluated independently by two referees. Thirty-eight percent (or 107) of the manuscripts received a unanimous reviewer recommendation for acceptance. All of them were accepted by the editor. The reviewers were in agreement that an additional 38% (or 107) manuscripts should be rejected. The editor rejected all these submissions. Twenty-four percent (or 68) of the submissions received a split-decision, with one reviewer recommending "accept" and the other "reject." The editor accepted

Table 4. *The fate of Journal of Abnormal Psychology submissions receiving a single editorial review (1973–1977)*

A. Considering each reviewer recommendation

Reviewer Recommendation	Number of Manuscripts	Editorial decision		Percentage Accepted
		Accept	Reject	
1 = Accept/As Is	24	22	2	91.7
2 = Accept/Revise	30	24	6	80.0
3 = Resubmit	26	11	15	42.3
4 = Reject	95	1	94	01.1
Total	175	58	117	33.1

B. Considering 1–2 = Accept; 3–4 = Reject

Reviewer Recommendation	Number of Manuscripts	Editorial decision		Percentage Accepted
		Accept	Reject	
(1–2) = Accept	54	46	8	85.2
(3–4) = Reject	121	12	109	90.1
Total	175	58	117	

Table 5. *The fate of Journal of Abnormal Psychology submissions receiving two editorial reviews (1973–1977)*

A. Considering "Accept/As Is" and "Accept/Revise" as "Accept"

Reviewer Recommendation	Number of Manuscripts	Editorial Decision		Percentage Accepted
		Accept	Reject	
Accept-Accept	181	159	22	87.8
Accept-Resubmit	150	79	71	52.7
Resubmit-Resubmit	30	8	22	26.7
Accept-Reject	195	28	167	14.4
Resubmit-Reject	162	6	156	03.7
Reject-Reject	278	1	277	00.4
Total	996	281	715	28.2

B. Considering 1 = Accept; 2–3 = Reject

Reviewer Recommendation	Number of Manuscripts	Editorial decision		Percentage	
		Accept	Reject	Agree	Disagree
Accept	353	212	141	60.1%	39.9%
Reject	643	69	574	89.3%	10.7%
Total	996	281	715		

2% (or 6) such manuscripts and rejected the remaining 22% (or 62) manuscripts. Thus, there was a 62/6 or more than tenfold probability that an editor would reject rather than accept a manuscript receiving mixed reviews.

In summary, at least for the general focus journals examined thus far, when editors are faced with split-decisions, they tend overwhelmingly to go with the lower of the two reviewer recommendations.

3.4. The editor's use of three reviews: "Go with the mode." Of the 1698 (6.6%) submissions to *JAP* (1973–1977), 112 received three reviews. As shown in Table 6, the disposition of these manuscripts is again closely related to reviewer recommendations.

Thus, all 5 manuscripts receiving unanimous acceptance votes were accepted for publication. Analogously, the 9 submissions receiving unanimous rejection votes were rejected. The disposition of the remaining 98 (or 87.5%) manuscripts is best understood by the editor's general adoption of majority rule or applying the formula "go with the mode." Thus, 24 of 33 (or 73%) of those submissions receiving 2 "accept" recommendations were accepted, whereas 80% (or 52/65) submissions receiving two "reject" votes were rejected. Application of the McNemar test of correlated proportions indicated no significant difference favoring either the editor's acceptance of these articles (20%) with majority rejection votes or his rejection of those articles (17%) with majority acceptance votes (McNemar's chi square(d), corrected, 1 df = .41, or of zero order significance).

I am unaware of comparable studies on manuscripts submitted to medical journals. **Bailar** noted that it was not unusual, in his role as editor of *JNCI* (1974–1980), however, either to reject manuscripts receiving three positive reviews or to publish submissions receiving three negative reviews. Given the importance of this phenomenon, I would invite Bailar to publish these data, because they contrast so sharply with what I have presented here for a prestigious behavioral science journal. It would be important to know: (a) precisely how frequently the phenomenon occurred; and (b) in what important respects the targeted manuscripts were dissimilar from those that were less problematic. Perhaps Bailar could provide this information in a *BBS* Continuing Commentary.

In the field of physics, recall that journal editors in the more specific focus areas use the "single initial reviewer system" (e.g., Hargens & Herting 1990; Lock 1985, p. 20) and so would tend not to have much data on the fate of manuscripts receiving three reviews. Although it is equally clear that the more general focus areas of physics often receive three reviews (or more), the extent to which an editor uses this information to make specific publication decisions is unknown. Despite this lack of specific information, there are some sparse data, deriving from the field of physics, that bear on the broader issue of how editors use information gained from referees to improve the quality of the editorial decision-making process. These data appeared in the *Physical Review (PR)* and *Physical Review Letters (PRL)* (1987, p. 7) *Annual Report* for the previous year, 1986. The statements pertain to a change in editorial policy for manuscripts submitted from the community of particle theorists to the subfield section, Elementary Particles, which represents one of 10 PRL areas of specialization. (The remaining nine subfields, or PRL content areas, are: General Physics, Cross-Disciplinary Physics, Astrophysics and Geophysics, Condensed Matter (CM), Electricity, CM Mechanics, Plasma Physics, Optics and Fluids, Nuclear Physics, and Atoms and Molecules.) The statement of journal policy change bears quoting:

> In March, 1985, a new system of handling papers in the theory of particles and fields was introduced. The divisional Associate Editors were enlisted to work closely on the processing of these papers, with the

Table 6. *The fate of 112 Journal of Abnormal Psychology manuscripts receiving three reviews (1973–1977)*

Reviewer Recommendation	Number of Manuscripts	Editorial decision		Percentage Accepted
		Accept	Reject	
3 "Accept"	5	5	0	100.0
2 "Accept," 1 "Reject"	33	24	9	72.7
1 "Accept," 2 "Reject"	65	13	52	20.0
3 "Reject"	9	0	9	00.0
Total	112	42	70	37.5

Note. "Accept" = "Accept/As Is" or "Accept/Revise"; "Reject" = "Resubmit" or "Reject"

objective of securing more expert and even-handed reviews and better serving our readers by attracting a larger share of the important papers in this area. To date, the experiment has been working smoothly, and we sense an improvement in relations with the community of particle theorists. This is reflected in increased submissions and publication in this area: Submissions for 1985 and 1986 were about 35% above the 1984 level and the numbers of published papers were about 45% above the 1984 level (reflecting also a moderately increased acceptance rate).

4. Improving the reliability and validity of peer review

4.1. Rationale, unifying concepts. The thoughtful, thought-provoking, and conceptually sound ideas of **Kraemer** seem especially valuable at this point in the exposition. Her very special talent for accurately relating pure mathematical reasoning to the flawed world of clinical reality has been achieved once again, and the field of peer review will be the richer for it. Hers is a well-reasoned commentary on the subtle interplay between issues of reliability and validity. I share her view on the importance of improving reliability but not at the expense of validity and, conversely, of improving validity without compromising reliability; I agree that both can be accomplished. **Kraemer**'s most valuable contribution is a theoretically and empirically sound framework, in which more specific ideas for improving the quality of peer review can be better classified, integrated, and examined critically.

Take **Kraemer**'s first comment about editors' basic need to select reviewers with varying degrees of expertise to achieve a balanced and comprehensive review. Her conclusion that "maximal validity" is achieved when errors are independent and the editor uses as many reliable reviewers as possible is correct. Moreover, if one evaluates this proposition in a cross-disciplinary or cross-specialty sense, its meaning can be further elucidated. For example, in the general subspecialty fields of sociology, psychology, medicine, and physics, it would be essential to select reviewers for their area of expertise (e.g., content specialist, biostatistician, biochemist) to increase the validity of the review. However, this careful selection of reviewers (weeding out the biased and nondiscriminating) should also increase significantly the reliability of the peer review process, at least for the overall reviewer recommendation. (See also the commentaries of **Kraemer** and **Lock**.)

Now if one considers the same issue for specific specialties of sociology, psychology, medicine, or physics, the community of peers may be so well defined that one could select reviewers randomly. This might obviate the nonrandom and unwitting selection of potentially biased reviewers. In summary, a selection process that would be a disaster for a general area may be just what is required in a more specific subspecialty area.

A second, unifying matter that **Kraemer** suggests may not be aimed at improving the reliability or the validity of peer review is the delicate and sensitive problem of striking a meaningful balance between committing Type I (alpha) errors (accepting flawed submissions) and Type II (beta) errors (rejecting nonflawed submissions). Kraemer and I agree that Type II error needs to be reduced but not at the expense of magnifying a Type I error. Again, if Kraemer's broad framework is extended to specific as well as more general areas of inquiry, her ideas dovetail nicely with those of **Cole**, whose theoretically meaty commentary uses the Type I–Type II distinction to explain differences in acceptance and rejection rates for the social and natural sciences.

Cole argues that behavioral scientists prefer to make Type II errors, whereas social scientists prefer to make Type I errors. There is certainly evidence for this hypothesis. Thus, the data presented in section 3 show quite clearly that when editors of major general journals in psychology are faced with a split review, they overwhelmingly opt for rejection. Moreover, Lock (1985) shows that the same phenomenon is at work for editors of general medical journals. The much higher rejection rates for general areas in physics itself, however, seems to indicate that the phenomenon is even broader than Cole suggests. Perhaps the idea needs to be amended: Editors in general areas across and within fields of inquiry desire to avoid Type I errors, whereas their specialized counterparts try to avoid Type II errors. Consistent with this philosophy, **Roediger**, a former editor of a psychology specialty journal, recommends the "when in doubt, accept" philosophy for editorial decisions on manuscripts receiving split reviews.

There appears to be a concerted effort on the part of general focus scientists to set criteria for deciding on just how to achieve the best balance between Type I and Type II errors. It is especially important to accomplish this because it is a well-known biostatistical fact that one can avoid a Type II error by simply increasing sufficiently the

size of N. What this means is that any difference between study samples, *no matter how trivial*, will produce statistical significance, or a so-called "positive" result. Cohen (1988) addresses this issue directly by suggesting, in the broader context of power analysis, that: (a) as a general rule, one adopt a Type II error rate of .20, thereby producing power of .80 (i.e., power = [1 − beta error]. Because Type I (alpha error rate) is often set at the nominal or conventional p level of .05, the adoption of this strategy means, literally, that one considers a Type I error to be of the order of four times as serious as a Type II error; and (b) the substantive, practical, or clinical meaning of the group difference, above and beyond its level of statistical significance should interpret .15 as a SMALL effect, .30 as a MEDIUM effect, and .50 as a LARGE effect. (See section 1.2 of this Response to **Eckberg** for the adoption of analogous guidelines for interpreting the practical significance of kappa, R_I, and R values.)

In the remaining parts of this section of the Response, I examine commentators' reactions to a number of further specific suggestions for improving the quality of peer reviews for both manuscript and grant submissions. These include: the role of multiple reviewers; using author anonymity or "blind" review; revealing reviewer identity; author review of referees; rewarding referee contributions; allowing authors multiple manuscript submissions; developing peer review appeals systems; and training reviewers.

4.2. The role of multiple reviewers. This strategy received widespread endorsement from those commentators who voiced an opinion (i.e., see specifically **Cohen, Crandall, Greene, Hargens, Kraemer, Marsh & Ball, Stricker** and **Zentall**).

On the other hand, **Colman** was opposed to the use of multiple reviewers, which he felt would: (a) produce "social loafing," or a lessening in reviewer effort; (b) encourage a "diffusion of responsibility"; and (c) increase substantially the workload of already overburdened referees. I believe that the "social loafing" phenomenon, to the extent that it might exist among multiple peer reviewers, can be greatly attenuated or even eliminated by appropriately rewarding useful and thoughtful reviews, while eliminating reviewers who consistently produce biased and poorly reasoned reviews. I am not so concerned about the increase in reviewer workload for the reasons given in section 7.2 of the target article, namely, that there are pools of potential referees large enough to make multiple reviews possible across disciplines (see also the relevant remarks of **Fletcher** in this regard).

4.3. Using author anonymity or "blind" review. A preference is expressed by **Kraemer** and **Lock** for author anonymity (blind reviews) as a strategy for improving the overall quality of peer review. **Lock, Colman,** and **Greene** mention the nonfeasibility of the practice for peer review of grants because it would minimize or eliminate the important role of the author's research "track record" in deciding on the merits of the proposed research. **Zentall** feels that voluntary blinding might defeat its own purpose because those most likely to benefit from their past record of research accomplishments would be the least likely to use the process. **Bailar** claims "substantial anecdotal evidence" to support the notion that reviewers evaluate more accurately the strong and weak qualities of a given manuscript submission when they are not blinded to authors. **Lock** presents some recent empirical data on this issue, which conflicts with Bailar's conjectures. Specifically, McNutt et al.'s (1990) report that blinding proved successful for 76% of reviewers, and that it also resulted in a 21% improvement in the overall quality of the reviews. On a 5 point scale for assessing the quality of peer review (in which 1 = very poor and 5 = excellent) the mean "summary grade" was significantly higher ($p = .007$) for the blinded over the nonblinded evaluations of the same manuscripts. Moreover, the difference in median grade was a full point. Blind reviews had a median quality of review of 4, whereas the nonblind reviews of the same submissions showed a median value of 3. Finally, based on an intriguing study by Garfunkel et al. (1990) **Lock** notes the urgency of studying the effects of blinding on editors themselves.

4.4. Revealing reviewer identity. It is argued by **Adams** that forcing reviewers to sign their evaluations would result in more constructive criticism of an author's work. **Rourke** concurs and predicts that the freedom of information movement will ultimately force journals to adopt the policy of signed reviews.

On the opposite side of the ledger, **Kraemer** is opposed to signed reviews. Consistent with the position I endorsed (sect. 7.4, target article), she opts instead for a voluntary decision for whether or not to sign. **Zentall** and **Greene** are also opposed to signed reviews, and for somewhat similar reasons. Zentall predicts that the fear of retribution would significantly reduce the likelihood of individual participation in the peer review process. Greene argues for the continued need to prevent the peer review of grants from becoming "personalized."

4.5. Author review of referees. The responsibility of detecting and discarding poor quality peer reviews is placed directly on the editor by **Kraemer**. Similarly, **Colman** enjoins the editors periodically to "solicit" authors' evaluations of reviewers' criticisms, reviewers' replies to authors' criticisms (as required) and possibly to invoke the aid of an independent judge (or "arbiter") until a "fair" editorial decision occurs. Consistent with these views, **Kiesler**'s "wise" editor would identify and ignore biased reviews. With respect to peer review of grants, **Greene** finds author review a useful screening procedure whose implementation in the Department of Veterans Affairs (DVA) will be considered as an adjunct to the present policy of inviting authors to submit the names of acceptable and nonacceptable reviewers. I agree with these recommendations, which complement those suggested in section 7.4 of the target article.

4.6. Rewarding referee contributions. It is suggested by **Adams** that referees be rewarded for providing quality peer reviews, without creating "undue bias" or impinging on the "freedom of scientists." I strongly agree. I would also add that whenever possible reviewers who consistently provide high quality evaluations (well reasoned, well documented, balanced) should be invited to serve as consulting editors, members of the boards of journals for which they review, or even associate or full editors, as appropriate. They should also be asked to

serve on grant review panels for such funding agencies as NIH, DVA, and NSF. In this manner scientists would be rewarded on the basis of their own scholarly contributions to peer review rather than on the basis of potentially more subjective criteria.

4.7. Allowing authors multiple manuscript submission. Agreement is voiced by **Mahoney** with the position taken in section 7.7 of the target article that, for a number of cogent reasons, the option of multiple manuscript submissions is not a viable one. His citation of Epstein (1990) adds confirming empirical support for the position. In my informal discussions with colleagues, I have yet to find one who would endorse such a practice.

4.8. Developing peer review appeals systems. Whereas an argument for a more formal appeals system for rejected manuscripts is made by **Zentall, Cole,** citing the finding by Stinchcombe and Ofshe (1969) that many acceptable articles are falsely rejected, opts for editors gradually to increase publication rates for submitted articles even at the risk of levying page charges on authors. Although this is an interesting suggestion, I am not sure what specific criteria editors would apply to justify increasing their acceptance rates. Commentators were in agreement, however, that the unfair disapproval of a grant submission is far more serious in its consequences than the unfair rejection of a journal article (e.g., **Adams, Cole, Kiesler, Mahoney, Salzinger,** Zentall). The problem is especially serious for funding in the social and behavioral sciences. Thus, as Mahoney notes, the National Research Council (1988) emphasized the need for a 30% increase in funding in these areas where funding has dropped 25% between 1972 and 1987, whereas it has increased by 36% in other areas of science during the same funding period.

Greene advocates strongly the need for an appeals system for any funding agency, because the peer review system is an imperfect one. He notes that the DVA has had an "effective" system for more than a decade. He also admits that appeal is a "sensitive" and "complex" phenomenon, so the ground rules on which it is based require periodic assessment. I agree with Greene's position. Rather than a formal appeal process per se, **Cole** recommends that granting foundations admit publicly that many of their rejected proposals are as fundable as many that are approved. He advocates specifically that the approval of such previously declined proposals should be undertaken even at the expense of reducing funding levels for the ensuing round of new grant proposals.

My concern with Cole's recommendation is that once a grant proposal receives the official federal stamp of "disapproval," it becomes more and more difficult to convince such lay persons as members of Congress that the submission should really have been funded in the first place. My solution would be to assign high priorities (no number attached) to the best considered proposals quite independently of whether there is funding available to support them. One could then request from Congress whatever additional funds may be required to support *all* the high priority grants. I believe that the way the system works today – assigning arbitrary funding cutoffs based on arbitrary numbers – creates the dilemma of funding a proposal with a priority score of, say, 112 and declining one with a score of 113 when *in fact* no reasonable peer reviewer can be expected to make a reliable differentiation of this minute degree of magnitude. To paraphrase **Delcomyn**'s analogy, the task that grant reviewers face is one of being asked to measure the dimensions of a nerve cell with a yardstick. My recommendation is intended to help obviate that measurement problem.

4.9. Training reviewers. The important issue of training reviewers was mentioned, in varying degree, by several commentators (i.e., **Adams, Crandall, Delcomyn, Kiesler, Rourke,** and **Zentall**).

Adams describes the typical "haphazard" and "uncertain" manner in which reviewers eventually learn to become "constructive" evaluators. Adam's previously mentioned support of reviewers disclosing their identity to authors is one way of producing such constructive reviewer reports. With a somewhat similar purpose in mind, **Zentall** proposes that editors send to reviewers a list of recommended guidelines for avoiding potential biases in the evaluation of a given submission. The same general strategy can be used with grant proposals.

Crandall, Delcomyn, and **Rourke** write of the importance of reviewers sharing others' reviews of the same manuscripts. Unfortunately, some granting agencies (e.g., NSF) have policy forbidding such a learning experience. The DVN, on the other hand, does provide this valuable service to its reviewers.

Crandall and **Delcomyn** note that the ability to write a useful review improves with experience. Crandall laments the fact that this experience is often gained at authors' expense. To help remedy the situation, Delcomyn provides a useful set of guidelines for reviewers that, though it derives from physiology, is pitched at a level general enough to be of cross-disciplinary use. The advantage of his guidelines over many others I have examined is that they contain within them the message that it is neither the task of reviewers nor editors to settle differing points of view in a given area of inquiry. Thus, if important questions raised in the introduction are answered through carefully controlled, well executed experiments, and the conclusions spring from the data, then the article should be accepted quite apart from whose particular theory or hypothesis is or is not being supported.

Crandall addresses more formally the notion of training reviewers by introducing the provocative idea of using prototype "ideal" reviews as guides. Filling in some of the required details, I would imagine that editors could locate in their files appropriate prototypic reviews that could be reliably rated as evidence for: "Accept/as is"; "Accept/Revise"; "Reject/Revise/Resubmit"; and "Reject/Unconditionally." With the necessary identifying information removed and the content disguised, these can be sent to authors to use in the same general manner that, for example, prototypic stages of cataract have been used to train ophthalmologists to classify cataract stages (i.e., Cicchetti et al. 1982).

The need for such formal training of reviewers may have been implicit though it appeared via a different route to **Nelson.** In a thoughtful commentary, she raises the issue of the specific process by which reviewers use information to arrive at publication or funding recommendations. She is right that very little is known about this process in peer review. Some findings reported a few

years back (Cicchetti & Eron 1979) were described in more detail in a subsequent *BBS* commentary (Cicchetti 1982). We found that although there were high correlations between what reviewers perceived as "important" and "well-designed" studies and their tendency to recommend publication (rs between .62 [research design] and .73 [importance]), the reliability of these ratings was appreciably lower (.19 and .28, respectively, as given in Table 1 of the target article). Although reviewers were asked to use specific rating forms describing "importance," "research design," and other manuscript attributes, it is entirely possible that they first read the manuscript, decided on their recommendation, and then filled out the form to be consistent with that recommendation (e.g., if one thinks the article is worth publishing then it must be important, well-enough designed, of appropriate reader interest).

Regardless of what specific interpretations may be appropriate, the more formal training of reviewers would probably enable them to use the same set of specific evaluative criteria more consistently. Then the process that reviewers use to arrive at a recommendation will have become standard, reliable, and if applied appropriately (e.g., using prototypic reviews as standards), valid.

5. Concluding comments

A number of commentators have suggested further investigations to place the area of peer review on an ever more solid scientific foundation. Given the interdisciplinary nature of science – my strongest appeal is that the cross-disciplinary approach taken in this target article should be further encouraged in future investigations. I would simply refer the interested reader to the specific commentaries of **Bornstein, Cohen, Gorman, Hargens, Lock, Marsh & Ball, Nelson,** and **Salzinger.**

In my opinion, the training of reviewers, as well as editors, authors, and consumers of research, is pivotal in increasing both the reliability and the validity of peer review.

I have recently come across the first study of which I am aware that broaches this topic directly. Oxman et al. (in press) were able to train successfully three classes of referees ("experts in research methodology," "MDs with research training," and "research assistants" – three in each group) to assess the overall scientific quality and other evaluation attributes of 36 review articles published in a wide range of journals in medicine (e.g., *New England Journal of Medicine*), psychiatry (e.g., American Journal of Psychiatry), and psychology (e.g., Psychological Bulletin):

Following specific training (or practice) on review articles and an additional one hour training session, the 36 articles were evaluated independently by the nine reviewers. For level of "overall scientific quality," the intraclass R, across the nine examiners, was .71; R_i values for each of the three groups of reviewers, separately, were, as follows: "Experts in research methodology" – R_i = .77 (EXCELLENT); "MDs with research training" – R_i = .74 (GOOD); and "Research Assistants" – R_i = .62 (GOOD).

Nine additional evaluative attributes, were measured on 7-point ordinal scales with four anchorage points provided for the scoring of each attribute (e.g., see Cicchetti et al. 1987). The nine attributes and their average R_i values, across the nine judges, concerned the extent to which: (1) search methods were reported (R_i > .8, or EXCELLENT); (2) a comprehensive search of the literature was conducted (R_i > .6 or GOOD); (3) inclusion criteria were reported (R_i > .8 or EXCELLENT); (4) selection biases were avoided (R_i > .6, or GOOD); (5) validity criteria were reported (R_i > .6, or GOOD); (6) validity data were reported (R_i > .6, or GOOD); (7) findings were combined appropriately (R_i = .5 or FAIR; (8) methods for combining the data were reported (R_i > .6, or GOOD); and (9) conclusions were supported by the data (R_i = .40, or FAIR). Although there were somewhat lower levels of agreement among the research assistants than within the other two groups of reviewers, for eight of the 10 evaluations (items 1–7 and the overall evaluation of scientific quality) the differences in R_i values were small. The average lower R_i values of the remaining two attributes, however, were due to the very low R_i levels achieved by "research assistants" relative to the other two groups of evaluators. For rating the extent to which "the findings were combined appropriately," the R_i for "experts" was at .6 (GOOD) and for "MDs with research training," it was > .9 (EXCELLENT). The corresponding R_i for "research assistants," however, was in the very "POOR" range, at > .2. Similarly, the extent to which "conclusions were supported by the data," the R_i's for "experts" and "MDs with research training" were beyond .6 (GOOD), whereas the corresponding R_i for "research assistants" was again in the very POOR range, or barely beyond the .1 level.

These results to my knowledge are the first to demonstrate that reviewers of different levels of experience, can be taught to evaluate reliably the same scientific documents. It is hoped that additional investigations of this kind will be undertaken across a broad range of research topics both within and across disciplines. Following the lead of commentator **Lock,** I would also hope that the important issue of training peer reviewers will be discussed at the 1992 Second World Conference on Peer Review.

Finally, in the open forum of "creative disagreement," I would extend a special invitation to those two commentators (**Bailar** and **Kiesler**), who were the most dubious about the need to study further the reliability and validity of the peer review process. I hope the panorama of ideas expressed by commentators across disciplines will convince them of the need to turn some of their own anecdotal experiences into further valuable research in this area. As editors of prestigious journals in behavioral science and medicine, their future insights and empirical investigations can make major contributions to the further understanding of the vicissitudes of peer review.

NOTE

1. Author is also affiliated with Yale University.

References

Abelson, P. H. (1980) Scientific communication. *Science* 209:60–62. [aDVC]

Abramowitz, S. I., Gomes, B. & Abramowitz, C. V. (1975) Publish or politic:

Referee bias in manuscript review. *Journal of Applied Social Psychology* 5:187–200. [aDVC]

Abt, H. A. (1988) What happens to rejected astronomical papers? *Publications of the Astronomical Society of the Pacific* 100:506–08. [aDVC]

Ad Hoc Working Group for Critical Appraisal of the Medical Literature (1987) A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine* 106:598–604. [SPL]

Adair, R. K. (1981) Anonymous refereeing. *Physics Today* 34:13–15. [aDVC]

(1982) A physics editor comments on Peters & Ceci's peer-review study. *Behavioral and Brain Sciences* 5:196. [aDVC]

Adair, R. K. & Trigg, G. L. (1979) Editorial: Should the character of *Physical Review Letters* be changed? *Physical Review Letters* 43:1969–74. [aDVC]

Allen, E. M. (1960) Why are research grant applications disapproved? *Science* 132:1532–34. [aDVC]

Amabile, T. M. (1983) Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology* 19:146–56. [RC]

American Psychological Association (1983) *Publication manual*, 3rd ed. [aDVC]

American Psychological Association (1985) *Standards for educational and psychological testing*. [RFB]

American Psychologist (1989) Members of underrepresented groups: Reviewers for journal manuscripts wanted. *American Psychologist* 44:1555. [RC]

Anonymous (1987) The publication game: Beyond quality in the search for a lengthy vitae. *Journal of Social Behavior and Personality* 2:3–12 [RC]

Armstrong, J. S. (1980) Unintelligible management research and academic prestige. *Interfaces* 10:80–86. [aDVC]

(1982a) Barriers to scientific contributions: The author's formula. *Behavioral and Brain Sciences* 5:197–99. [aDVC]

(1982b) The ombudsman: Is peer review by peers as fair as it appears? *Interfaces* 12:62–74. [aDVC, JSA]

(1982c) Research on scientific journals: Implications for editors and authors. *Journal of Forecasting* 1:83–104. [aDVC, JSA]

Ballar, J. C., III & Patterson, K. (1985) Journal peer review: The need for a research agenda. *The New England Journal of Medicine* 312:654–57. [aDVC]

Baird, J. C., Green, D. M., and Luce, R. D. (1960) Variability and sequential effects in cross-modality matching of area and loudness. *Journal of Experimental Psychology: Human Perception and Performance* 6:277–89. [DL]

Bakanic, V. McPhail, C. & Simon, R. J. (1987) The manuscript review and decision-making process. *American Sociological Report* 52:631–42. [aDVC, LJS]

Bartko, J. J. (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19:3–11. [aDVC]

(1974) Corrective note to: "The Intraclass Correlation Coefficient as a Measure of Reliability." *Psychological Reports* 34:418. [aDVC]

(1976) On various intraclass correlation reliability coefficients. *Psychological Bulletin* 83:762–65. [aDVC]

Bartko, J. J. & Carpenter, W. T. (1976) On the methods and theory of reliability. *Journal of Nervous and Mental Disease* 163:307–17. [aDVC]

Beck, A. T. (1976) *Cognitive therapy and the emotional disorders*. International Universities Press. [aDVC]

Benwell, R. (1979) Authors anonymous? *Physics Bulletin* 30:288. [aDVC]

Berelson, B. (1960) *Graduate education in the United States*. McGraw-Hill. [aDVC]

Bernstein, G. S. (1984) Scientific rigor, scientific integrity: A comment on Sommer & Sommer. *American Psychologist* 39:1316. [aDVC]

Beyer, J. M. (1978) Editorial policies and practices among leading journals in four scientific fields. *The Sociological Quarterly* 19:68–88. [aDVC]

Blashfield, R. K. (1976) Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin* 83:377–88. [rDVC]

Bloch, D. A. & Kraemer, H. C. (1989) 2X2 kappa coefficients: Measures of agreement or association. *Biometrics* 45:269–87. [HCK]

Boehme, G. (1977) Models for the development of science. In: *Science, technology, and society: A cross-disciplinary perspective*, ed. I. Spiegel-Rosing & D. de Solla Price. Sage. [aDVC]

Boehme, G., van der Daele, W. & Krohn, W. (1976) Finalization of science. *Social Science Information* 15:306–30. [aDVC]

Boor, M. (1986) Suggestions to improve manuscripts submitted to professional journals. *American Psychologist*. 41:721–22. [RC]

Bornstein, R. F. (1990) Manuscript review in psychology: An alternative model. *American Psychologist* 45:672–73. [RFB]

(in press) Publication politics, experimenter bias, and the replication process in social science research. *Journal of Social Behavior and Personality*. [RFB]

Bowen, D. D., Perloff, R. & Jacoby, J. (1972) Improving manuscript evaluation procedures. *American Psychologist* 25:221–25. [aDVC]

Bozarth, M. D. & Roberts, R. R. Jr. (1972) Signifying significant significance. *American Psychologist* 27:774–75. [aDVC]

Bradley, J. V. (1981) Pernicious publication practices. *Bulletin of the Psychonomic Society* 18:31–34. [aDVC]

Braida, L. D. & Durlach, N. T. (1972) Intensity perception. II. Resolution in one interval paradigms. *Journal of the Acoustical Society of America* 51:483–502. [DL]

Brennan, R. L. & Light, R. J. (1974) Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology* 27:154–63. [rDVC]

Broad, W. J. (1988) Science can't keep up with the flood of new journals. *The New York Times*, Feb. 16:C1, C11. [JF]

Brook, R. J. & Stirling, W. D. (1984) Agreement between observers when the categories are not specified in advance. *British Journal of Mathematical and Statistical Psychology* 37:271–82. [rDVC]

Byrne, C. (1980) Tutor marked assessments at the Open University: A question of reliability. *Assessment in Higher Education* 5:104–18. [DL]

Campbell, J. P. (1982) Some remarks from the outgoing editor. *Journal of Applied Psychology* 67:691–700. [LLH]

Carsrud, K. B. (1984) Out of the frying pan: A reply to Sommer & Sommer. *American Psychologist* 31:1317–18. [aDVC]

Ceci, S. J. & Peters, D. (1984) How blind is blind review? *American Psychologist* 39:1491–94. [aDVC]

Chalmers, I. (1990) Underreporting research is scientific misconduct. *Journal of the American Medical Association* 263:1405–08. [SPL]

Chalmers, T. C., Frank, C. S. & Reitman, D. (1990) Minimizing the three stages of publication bias. *Journal of the American Medical Association* 263:1392–95. [SPL]

Chase, J. M. (1970) Normative criteria for scientific publication. *American Sociologist* 5:262–65. [aDVC]

Chubin, D. E. (1982) Reform of peer review. *Science* 215:40. [aDVC]

Cicchetti, D. V. (1976) Assessing interrater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry* 129:452–56. [aDVC]

(1980) Reliability of reviews for the American Psychologist: A biostatistical assessment of the data. *American Psychologist* 35:300–3. [aDVC, LJS]

(1981) Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. *Applied Psychological Measurement* 5:101–04. [arDVC]

(1982) On peer review: "We have met the enemy and he is us." *Behavioral and Brain Sciences* 5:205. [arDVC]

(1985) A critique of Whitehurst's "Interrater agreement for journal manuscript reviews:" De omnibus, disputandum est. *American Psychologist* 40:563–68. [aDVC, MED]

(1988) When diagnostic agreement is high, but reliability is low: Some paradoxes occurring in independent neuropsychological assessments. *Journal of Clinical and Experimental Neuropsychology* 10:605–22. [aDVC]

Cicchetti, D. V. & Conn, H. O. (1976) A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. *Yale Journal of Biology and Medicine* 45:373–83. [aDVC]

(1978) Reviewer evaluation of manuscripts submitted to medical journals. Paper presented to the American Statistical Association Meetings, San Diego, CA. (also abstracted in *Biometrics* [1978] 34:728) [aDVC]

Cicchetti, D. V. & Eron, L. D. (1979) The reliability of manuscript reviewing for the *Journal of Abnormal Psychology*. *Proceedings of the American Statistical Association* (Social Statistics Section) 22:596–600. [aDVC]

Cicchetti, D. V. & Feinstein, A. R. (1990) High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* 43:551–68. [arDVC]

Cicchetti, D. V. & Fleiss, J. L. (1977) Comparison of the null distributions of weighted kappa and the C ordinal statistic. *Applied Psychological Measurement* 1:195–201. [aDVC]

Cicchetti, D. V. & Heavens, R. (1979) RATCAT (Rater Agreement/Categorical Data). *American Statistician* 33:91. [aDVC]

Cicchetti, D. V. & Showalter, D. (1988) A computer program for determining the reliability of dimensionally scaled data when the numbers and specific sets of examiners may vary at each assessment. *Educational and Psychological Measurement* 48:717–20. [aDVC]

Cicchetti, D. V. & Sparrow, S. S. (1981) Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency* 86:127–37. [aDVC]

Cicchetti, D. V. & Tyrer, P. (1988) Reliability and validity of personality assessment. In: *Personality disorders: Diagnosis, management and course*, ed. P. L. Tyrer. Butterworth Scientific Ltd. [rDVC]

Cicchetti, D. V., Aivano, S. L. & Vitale, J. (1976) A computer program for assessing the reliability and systematic bias of individual measurements. *Educational and Psychological Measurement* 36:761–64. [aDVC]

(1977) Computer programs for assessing rater agreement and rater bias for qualitative data. *Educational and Psychological Measurement* 37:195–201. [aDVC]

Cicchetti, D. V., Lee, C., Fontana, A. F. & Dowds, B. N. (1978) A computer program for assessing specific category-rater agreement for qualitative data. *Educational and Psychological Measurement* 38:805–13. [aDVC]

Cicchetti, D. V., Sharma, Y. & Cotlier, E. (1982) Assessment of observer variability in the classification of human cataracts. *Yale Journal of Biology and Medicine* 55:81–88. [rDVC]

Cicchetti, D. V., Showalter, D. & Tyrer, P. (1985) The effect of number of rating-scale categories upon levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement* 9:31–36. [aDVC]

Cleary, F. R. & Edwards, D. J. (1960) The origins of the contributors to the A.E.R. during the fifties. *American Economic Review* 50:1011–14. [aDVC]

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46. [aDVC, RR]

(1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70:213–20. [aDVC]

(1988) *Statistical power analysis for the behavioral sciences*, 2nd ed. Lawrence Erlbaum. [rDVC, RR]

Cole, J. & Cole, S. (1973) *Social stratification in science*. University of Chicago Press. [aDVC]

(1981) *Peer review in the National Science Foundation: Phase II of a study*. National Academy of Sciences. [aDVC]

(1985) Experts' "consensus" and decision-making at the National Science Foundation. In: *Selectivity in information systems: Survival of the fittest*, ed. K. S. Warren. Praeger. [aDVC]

Cole, S. (1978) Scientific reward systems: A comparative analysis. In: *Research in sociology of knowledge, sciences, and art*, ed. R. A. Jones. JAI Press. [aDVC]

(1983) The hierarchy of the sciences. *American Journal of Sociology* 89:111–39. [aDVC], SC

Cole, S., Cole, J. & Simon, G. A. (1981) Chance and consensus in peer review. *Science* 214:881–86. [aDVC, SC, LLH]

Cole, S., Cole, J. & Dietrich, L. (1978) Measuring the cognitive state of scientific disciplines. In: *Toward a metric of science: The advent of science indicators*, ed. Y. Elkana, J. Lederberg, R. K. Merton, A. Thackray, & J. Zuckerman. Wiley. [SC]

Cole, S., Rubin, L. & Cole, J. (1978) *Peer review in the National Science Foundation*. National Academy of Sciences. [aDVC]

Cole, S., Simon, G. & Cole, J. (1988) Do journal rejection rates index scientific consensus? *American Sociological Review* 53:152–56. [SC]

Colman, A. M. (1982a) *Game theory and experimental games: The study of strategic interaction*. Pergamon Press. [AMC]

Colman, A. M. (1982b) Manuscript evaluation by journal referees and editors: Randomness or bias? *Behavioral and Brain Sciences* 5:205–06. [AMC]

Conger, A. J. (1980) Integration and generalization of Kappa for multiple raters. *Psychological Bulletin* 88:322–28. [rDVC]

(1985) Kappa reliabilities for continuous behaviors and events. *Educational and Psychological Measurement* 45:861–68. [rDVC]

Conn, H. O. (1974) An experiment in blind program selection. *Clinical Research* 22:128–34. [aDVC]

Cotlier, E., Fagadau, W. & Cicchetti, D. V. (1982) Methods for evaluation of medical therapy of senile and diabetic cataracts. *Transactions of the Opthalmologic Societies of the United Kingdom* 102:416–22. [rDVC]

Cox, R. (1967) Examinations and higher education: A survey of the literature. *Universities Quarterly* 21:292–340. [DL]

Crandall, R. (1986) Peer review: Improving editorial procedures. *BioScience* 36:607–09. [RC]

(1987a) Gauntlet thrown: Publication procedures are challenged. *Dialogue* (APA Division 8) 1:5. [RC]

'198~b) We need research on what constitutes good journal papers – and good editing – not guesswork on how to improve manuscripts! *American Psychologist* 42:407–08. [RC]

(1990) Improving editorial procedures. *American Psychologist* 45:665–66. [RC]

Crane, D. (1967) The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *American Sociologist* 32:195–201. [aDVC]

(1972) *Invisible colleges*. University of Chicago Press. [DLE]

Cronbach, L. J. (1981) Comment on "Chance and consensus in peer review." *Science* 214:1293. [LLH]

Culliton, B. J. (1964) Fine-tuning peer review. *Science* 226:1401–02. [aDVC, RG]

Darley, J. M. & Latane, B. (1968) Bystander intervention in emergencies. Diffusion of responsibility. *Journal of Personality and Social Psychology* 8:337–83. [AMC]

Darlington, R. (1980) Another peek in the file drawers (unpublished manuscript). [PHS]

Davies, M. & Fleiss, J. L. (1982) Measuring agreement for multinomial data. *Biometrics* 38:1047–51. [rDVC]

DeBakey, L. & DeBakey, S. (1976) Impartial, signed reviews. *New England Journal of Medicine* 294:564. [aDVC]

Delucchi, K. L. (1983) The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin* 94:166–76. [rDVC]

Diamond, J. (1985) Variations on a theme. *Nature* 314:222–23. [aDVC]

Dickersin, K. (1990) The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association* 263:1385–89. [SPL]

Doherty, M. E. & Tweney, R. D. (1988) The role of data and feedback error in inference and prediction. Final report for ARI Contract MDA903-85-K-0193. [MEC]

Eckberg, D. (1982) Theoretical implications of failure to detect prepublished submissions. *Behavioral and Brain Sciences* 5:25–26. [DLE]

Eells, W. C. (1930) Reliability of reported grading of essay type examinations. *Journal of Educational Psychology* 21:48–52. [DL]

Eichorn, D. H. & VandenBos, G. R. (1985) Dissemination of scientific and professional knowledge. *American Psychologist* 40:1301–16. [RFB]

Eight APA journals initiate controversial blind reviewing (1972) *APA Monitor*, pp. 1, 5. [aDVC]

Epstein, W. M. (1990) Confirmatory response bias among social work journals. *Science, Technology and Human Values* 15:9–38. [rDVC, MJM]

Estes, W. K. (1975) Some targets for mathematical psychology. *Journal of Mathematical Psychology* 12:263–82. [PHS]

Evans, J. T., Nadjari, H. I. & Burchell, S. A. (1990) Quotational and reference accuracy in surgical journals: A continuing peer-review problem. *Journal of the American Medical Association* 263:1353–54. [JSA]

Feinstein, A. R. (1987) *Clinimetrics*. Yale University Press. [rDVC]

Feinstein, A. R. & Cicchetti, D. V. (1990) High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43:543–49. [arDVC]

Feynman, R. P. (1985) *Surely you are joking, Mr. Feynman*. Bantam. [PHS]

Finn, R. H. (1970) A note on estimating the reliability of categorical data. *Educational and Psychological Measurement* 30:71–76. [aDVC]

Fisher, A. (1989) Seeing atoms. *Popular Science*:102–07. [JSA]

Fiske, D. W. & Fogg, L. (1990) But the reviewers are making different criticisms of my paper!: Diversity and uniqueness in reviewer comments. *American Psychologist* 45:591–98. [rDVC, JSA]

Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76:378–82. [rDVC]

(1975) Measuring agreement between two judges on the presence or absence of a trait. *Biometrics* 31:651–59. [aDVC]

(1981) *Statistical methods for rates and proportions*, 2nd ed. Wiley. [aDVC, RR]

Fleiss, J. L. & Cicchetti, D. V. (1978) Inference about weighted kappa in the non-null case. *Applied Psychological Measurement* 2:113–17. [rDVC]

Fleiss, J. L. & Cohen, J. (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33:613–19. [aDVC]

Fleiss, J. L. & Cuzick, J. (1979) The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement* 3:537–52. [rDVC]

Fleiss, J. L., Cohen, J. & Everitt, B. S. (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72:323–37. [rDVC]

Fleiss, J. L., Nee, J. C. M. & Landis, J. R. (1979) Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin* 86:974–77. [rDVC]

Freeman, C. & Tyrer, P., eds. (1989) *Research methodology in psychiatry: A beginner's guide*. Royal College of Psychiatrists/Gaskell Books. [PT]

Fuller, S. (1989). *Philosophy of science and its discontents*. Westview Press. [MEC]

Furchtgott, E. (1984) Replicate, again and again. *American Psychologist* 39:1315–16. [aDVC]

Garber, H. L. (1984) On Sommer & Sommer. *American Psychologist* 31:1315. [aDVC]

Garcia, C., Rosenfield, N. S., Markowitz, R. K., Seashore, J. H., Touloukian, R. J. & Cicchetti, D. V. (1987) Appendicitis in children: Accuracy of the barium enema. *American Journal of Diseases of Children* 141:1309–12. [rDVC]

Gardner, M. J., Snee, M. P., Hall, A. J., Powell, C. A., Downes, S. & Terrell, J. D. (1990) Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria. *British Medical Journal* 300:423–29. [SPL]

Garfield, E. (1972) Citation analysis as a tool in journal evaluation. *Science* 178:471–79. [RFB]

Garfunkel, J. M., Ulshen, R. H., Hamrick, H. J. & Lawson, E. E. (1990) Problems identified by secondary review of accepted manuscripts. *Journal of the American Medical Association* 263:1369–71. [rDVC, SPL]

Garner, W. R. (1962) *Uncertainty and structure as psychological concepts.* Wiley. [DL]

Garner, W. R. & McGill, W. J. (1956) The relation between information and variance analyses. *Psychometrika* 21:219–28. [arDVC], JBG

Garvey, W. D., Lin, N. & Nelson, C. E. (1970) Some comparisons of communication activities in the physical and social sciences. In: *Communication among scientists and engineers*, ed. C. E. Nelson & D. K. Pollock. Health. [SC]

(1979) Communication in the physical and social sciences. In: *Communication: The essence of science*, ed. W. D. Garvey. Pergamon Press. [aDVC]

Gholson, B. & Barker, B. (1985) Kuhn, Lakatos, and Laudan: Applications in the history of physics and psychology. *American Psychologist* 40:755–69. [aDVC]

Giere, R. N. (1988). Explaining science: A cognitive approach. *University of Chicago Press.* [MEG]

Gillett, R. (1985) Nominal scale response agreement and rater uncertainty. *British Journal of Mathematical and Statistical Psychology* 38:58–66. [rDVC]

Gilmore, J. B. (1979) Illusory reliability in journal reviewing. *Canadian Psychological Review* 20:157–58. [arDVC, JBG]

Glenn, N. D. (1976) The journal article review process: Some proposals for change. *American Sociologist* 11:179–85. [aDVC]

Goodman, L. A. & Kruskal, W. H. (1954) Measures of association for cross classifications. *Journal of the American Statistical Association* 49:732–64. [aDVC]

Goodrich, D. W. (1945) An analysis of manuscripts received by the editors of the *American Sociological Review* from May 1, 1944, to September 1, 1945. *American Sociological Review* 10:716–25. [aDVC]

Goodstein, L. D. (1982) When will the editors start to edit? *Behavioral and Brain Sciences* 5:212–13. [LJS]

Goodstein, L. D. & Brazis, K. L. (1970) Credibility of psychologists: An empirical study. *Psychological Reports* 27:835–38. [aDVC, JSA]

Gordon, M. D. (1977) Evaluating the evaluators. *New Scientist* 73:342–43. [aDVC]

(1978) A study of the evaluation of papers by primary journals in the U.K. University of Leicester. [LLH]

Gorman, M. E. (1986) How the possibility of error affects falsification on a task that models scientific problem-solving. *British Journal of Psychology* 77:65–79. [MEG]

(1989) Error, falsification and scientific inference: An experimental investigation. *Quarterly Journal of Experimental Psychology*, 41A, 385–412. [MEG]

Gorman, Michael E. & Gorman, Margaret E. (1984) A comparison of disconfirmatory, confirmatory and a control strategy on Wason's 2, 4, 6 task. *Quarterly Journal of Experimental Psychology* 12:129–40. [MEG]

Gottfredson, S. D. (1978) Evaluating psychology research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist* 33:920–34. [aDVC, RFB, JBG]

Green, D. M., Luce, R. D. & Duncan, J. E. (1977) Variability and sequential effects in magnitude production and estimation of auditory intensity. *Perception & Psychophysics* 22:450–56. [DL]

Green, D. M., Luce, R. D. & Smith, A. F. (1980) Individual magnitude estimates for various distributions of signal intensity. *Perception & Psychophysics* 27:483–88. [DL]

Greenwald, A. G. (1975) Consequences of prejudice against the null hypothesis. *Psychological Bulletin* 82:1–20. [aDVC, PHS]

(1976) An editorial. *Journal of Personality and Social Psychology* 33:1–7. [aDVC]

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R. & Baumgardner, M. H. (1986) Under what conditions does theory obstruct research progress? *Psychological Review* 93:216–29. [aDVC]

Gross, S. T. (1986) The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics* 42:883–93. [rDVC]

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B. & Shapiro, R. W. (1981) Reliability studies of psychiatric diagnosis. Theory and practice. *Archives of General Psychiatry* 38:408–13. [rDVC]

Guilford, J. P. (1954) *Psychometric methods*, 2nd ed. McGraw-Hill. [RR]

Gulliksen, H. O. (1950) *Theory of mental tests.* Wiley. [DL, LJS]

Guyatt, G. H., Townsend, M. & Berman, L. (1987) A comparison of Likert and visual analogue scales for measuring change in function. *Journal of Chronic Diseases* 40:1129–33. [rDVC]

Hall, J. A. (1979) Author review of reviewers. *American Psychologist* 34:798. [aDVC]

Hargens, L. L. (1988) Scholarly consensus and journal rejection rates. *American Sociological Review* 53:139–51. [aDVC, SC]

(1990) Variation in journal peer-review systems: Possible causes and consequences. *Journal of the American Medical Association* 263:1348–52. [arDVC, LLH]

Hargens, L. L. & Herting, J. R. (1990a) A new approach to referees' assessments of manuscripts. *Social Science Research* 19:1–16. [arDVC, LLH]

(1990b) Neglected considerations in the analysis of agreement among journal referees. *Scientometrics* 19:91–106. [aDVC, LLH]

Harnad, S. (1979) Creative disagreement. *The Sciences* 19:18–20. [aDVC]

ed. (1983) *Peer commentary on peer review: A case study in scientific quality control.* Cambridge University Press (reprinted from *Behavioral and Brain Sciences*, vol. 5). [aDVC]

(1985) Rational disagreement in peer review. *Science, Technology & Human Values* 10(3):55–62. [aDVC, LJS]

(1986) Policing the paper chase. *Nature* 322:24–25. [aDVC, JBG]

Hartog, P., Rhodes, E. C., and Burt, C. (1936) *The marks of examiners.* Macmillan. [DL]

Heavens, R. H. Jr. & Cicchetti, D. V. (1978) A computer program for calculating rater agreement and bias statistics using contingency table input. *Proceedings of the American Statistical Association* (Statistical Computing Section) 21:366–70. [aDVC]

Hendrick, C. (1976) Editorial comment. *Personality and Social Psychology Bulletin* 2:27–08. [aDVC]

(1977) Editorial comment. *Personality and Social Psychology Bulletin* 3:1–2. [aDVC]

Hensler, D. (1976) Perceptions of the National Science Foundation peer-review process: A report on a survey of NSF reviewers and applicants. NSF publication 77-33. [aDVC]

Heskin, K. (1984) The Milwaukee Project: A cautionary comment. American *Psychologist* 39:1316–17. [aDVC]

Holt, V. E. (1985) Research briefings: Peer-review appeals system established. *American Psychological Association (APA) Monitor* 16:18. [aDVC]

Horrobin, D. F. (1990) The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association* 263:1438–41. [JSA]

Howe, M. J. A. (1982) Peer reviewing: Improve or be rejected. *Behavioral and Brain Sciences* 5:218–19. [aDVC]

Hubbard, R. & Armstrong, J. S. (1990) Replication and the development of marketing science. Marketing Department working paper, The Wharton School, University of Pennsylvania. [JSA]

Hughes, H. M. (1976) Letter to the editor. *American Sociologist* 11:178–79. [aDVC]

Hull, D. L. (1988) *Science as a process.* University of Chicago Press. [LLH]

Hunt, E. (1971) Psychological publications. *American Psychologist* 26:311. [aDVC]

Hunt, K. (1975) Do we really need more replications? *Psychological Reports* 36:587–93. [aDVC]

Ingelfinger, F. J. (1974) Peer review in biomedical publication. *American Journal of Medicine* 56:686–92. [aDVC]

(1975) Charity and peer review in publication. *New England Journal of Medicine* 293:1371–72. [aDVC]

Ison, J. R. (1985) The granting system and healthy research. *Science* 230:376. [aDVC]

Iyengar, S. & Greenhouse, J. B. (1988) Selection model and the file drawer hypothesis. *Statistical Science* 33:109–35. [PHS]

Jesteadt, W., Luce, R. D. & Green, D. M. (1977) Sequential effects in judgment of loudness. *Journal of Experimental Psychology: Human Perception and Performance* 3:92–104. [DL]

Jesteadt, W., Wier, C. C. & Green, D. M. (1977) Intensity discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America* 61:169–77. [DL]

Jonckheere, A. R. (1970) Techniques for ordered contingency tables. In: *Proceedings of the NUFFIC International Summer Session in Science, Het Oude Hof*, ed. J. B. Riemersma & H. C. van der Meer. The Hague. [aDVC]

Jones, R. (1974) Rights, wrongs, and referees. *New Scientist* 61:758–59. [aDVC]
Kahneman, D., Slovic, P. & Tversky, A., eds. (1982) *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press. [HLR]
Kamin, L. J. (1981) *The intelligence controversy*, ed. H. J. Eysenck. Wiley. [PHS]
Kazdin, A. E. (1982) *Single-case research designs: Methods for clinical and applied settings*. Oxford University Press. [aDVC]
Kerr, S., Tolliver, J. & Petree, D. (1977) Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal* 20:132–41. [aDVC]
Klayman, J. & Ha, Y.-W. (1987) Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review* 94:211–28. [MEG]
Koran, L. M. (1975a) The reliability of clinical methods, data, and judgments. *New England Journal of Medicine* 293:642–46. [rDVC]
(1975b) The reliability of clinical methods, data, and judgments. *New England Journal of Medicine* 293:695–701. [rDVC]
Koshland, D. E. Jr. (1985) Peer review of peer review. *Science* 228:1387. [aDVC]
Kraemer, H. C. (1980) Extension of the kappa coefficient. *Biometrics* 36:207–16. [rDVC]
(1982) Estimating false alarms and missed events from interobserver agreement: Comment on Kaye. *Psychological Bulletin* 92:749–54. [rDVC]
(1988) Assessment of 2x2 associations: Generalization or signal-detection methodology. *The American Statistician* 42:37–49. [rDVC]
Kraus, C. A. (1950) The present state of academic research. *Chemical and Engineering News* 28:3203–04. [aDVC]
Krippendorff, K. (1970) Bivariate agreement coefficients for reliability of data. In: *Sociological methodology*, ed. E. G. Borgatta. Jossey-Bass. [aDVC]
Krystal, J., Giller, E. & Cicchetti, D. V. (1986) Assessment of alexithymia in post-traumatic stress disorder and psychosomatic illness: Introduction of a reliable measure. *Psychosomatic Medicine* 48:84–94. [rDVC]
Kuhn, T. (1962) *The structure of scientific revolutions*. University of Chicago Press. [aDVC, LDN]
Lakatos, I. (1972) Falsification and the methodology of scientific research programmes. In: *Criticism and the growth of knowledge*, ed. I. Lakatos & A. Musgrave. Cambridge University Press. [aDVC]
Laming, D. (1984) The relativity of "absolute" judgments. *British Journal of Mathematical and Statistical Psychology* 37:152–83. [DL]
(1990) The reliability of a certain university examination compared with the precision of absolute judgments. *Quarterly Journal of Experimental Psychology* 42A:239–54. [DL]
(in press) Reconciling Fechner and Stevens? *Behavioral and Brain Sciences*. [DL]
Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:1599–74. [rDVC]
Latane, B., Williams, K. & Harkins, S. (1979) Many hands make light work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology* 37:822–32. [AMC]
Laudan, L. (1984) *Science and values: The aims of science and their role in scientific debate*. University of California Press. [aDVC]
Lawlis, G. F. & Lu, E. (1972) Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin* 78:17–20. [aDVC]
Lazarus, D. (1982) Interreferee agreement and acceptance rates in physics. *Behavioral and Brain Sciences* 5:219. [aDVC]
LeLewis, D. & Burke, C. J. (1949) The use and misuse of the chi square test. *Psychological Bulletin* 46:433–89. [rDVC]
Leach, C. (1979) *Introduction to statistics: A nonparametric approach for the social sciences*. John Wiley. [aDVC]
Lindsey, D. (1977) Participation and influence in publication review proceedings. *American Psychologist* 32:379–86. [RFB]
(1978) *The scientific publication system in social science*. Jossey-Bass. [aDVC, RFB]
(1988) Assessing precision in the manuscript review process: A little better than a dice roll. *Scientometrics* 14:75–82. [LLH]
Lock, S. (1985) *A difficult balance: Editorial peer review in medicine*. ISI Press. [aDVC, JF]
Lodahl, J. B. (1970) Paradigm development as a source of consensus in scientific fields (unpublished master's thesis). [aDVC]
Lord, F. N. & Novick, M. R. (1968) *Statistical theories of mental test scores*. Addison-Wesley. [LLH]
Luce, R. D. (1989) *A history of psychology in autobiography*, vol. 8, ed. G. Lindzey. Stanford University Press. [PHS]
Luce, R. D. & Green, D. M. (1978) Two tests of a neural attention hypothesis for auditory psychophysics. *Perception & Psychophysics* 23:363–71. [DL]
Luce, R. D., Nosofsky, R. M., Green, D. M. & Smith, A. F. (1982) The bow and sequential effects in absolute identification. *Perception & Psychophysics* 32:397–408. [DL]
Machol, R. (1981) Letter to the editor. *The Sciences* 21:xxi. [aDVC]
Maher, B. A. (1978) A reader's, writer's, and reviewer's guide to assessing research reports in clinical psychology. *Journal of Consulting and Clinical Psychology* 46:835–38. [aDVC]
Mahoney, M. J. (1976) *Scientist as subject: The psychological imperative*. Ballinger. [LLH]
(1977) Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy Research* 1:161–75. [aDVC, LDN, PHS, SPL, JSA]
(1978) Publish and perish. *Human Behavior* 7:38–41. [aDVC]
(1982) Publication, politics, and scientific progress. *Behavioral and Brain Sciences* 5:220–21. [AMC]
(1985) Open exchange and epistemic progress. *American Psychologist* 40:29–39. [ADVC, JBG, RFB, JF]
(1987) Scientific publication and knowledge politics. *Journal of Social Behavior and Personality* 2:165–76. [RFB]
(1990) Bias, controversy, and abuse in the study of the scientific publication system. *Science, Technology, & Human Values* 15:50–55. [MJM]
Mahoney, M. J., Kazdin, A. E. & Kenigsberg, M. (1978) Getting published. *Cognitive Therapy and Research* 2:69–70. [aDVC]
Margulis, L. (1977) Letter to the editor: Peer review attacked. *The Sciences* 17:5, 31. [aDVC]
Marsh, H. W. & Ball, S. (1981) Interjudgmental reliability of reviews for the *Journal of Educational Psychology*. *Journal of Educational Psychology* 73:872–80. [aDVC], HWM
(1989) The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *Journal of Experimental Education* 57:151–69. [HWM]
McCarthy, P., Sharpe, M. R., Spiesel, S. Z., Dolan, T. F., Forsyth, B. W., DeWitt, T. G., Fink, H. D., Baron, M. A. & Cicchetti, D. V. (1982) Observation scales to identify serious illness in febrile children. *Pediatrics* 70:802–09. [rDVC]
McCarthy, P. L., Sznajderman, S. D., Lustman-Findling, K., Baron, M. A., Fink, H. D., Czarkowski, N., Bauchner, H., Forsyth, B. C. & Cicchetti, D. V. (1990) Mothers' clinical judgment: A randomized trial of the acute illness observation scales. *Journal of Pediatrics* 116:200–06. [rDVC]
McCartney, J. L. (1978) Making sense of reviewers' comments. Paper presented to the Southern Sociological Association Meetings, New Orleans, LA. [aDVC]
McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12:153–57. [rDVC]
(1955) *Psychological statistics*. Wiley. [GSW]
McNutt, R. A., Evans, A. T., Fletcher, R. H. & Fletcher, S. W. (1990) The effects of blinding on the quality of peer review. *Journal of the American Medical Association* 263:137–76. [rDVC, JSA, SPL]
Merton, R. K. (1973) The sociology of science: Theoretical and empirical investigations. University of Chicago Press. [aDVC, DLE]
Meyer, G. S. (1979) Academic labor and the development of science (unpublished doctoral dissertation). State University of New York at Stony Brook. [SC]
Mezzich, J. E., Kraemer, H. C., Worthington, D. R. L. & Coffman, G. A. (1981) Assessment of agreement among several raters formulating multiple diagnoses. *Journal of Psychiatric Research* 16:29–39. [rDVC]
Mitroff, I. I. & Chubin, D. E. (1979) Peer review at the NSF: A dialectical policy and analysis. *Social Studies of Science* 9:199–232. Sage [aDVC]
Mulkay, M. (1977) Sociology of the scientific research community. In: *Science, technology, and society*, ed. I. Spiegel-Rosing & D. de Solla Price. Sage. [aDVC]
Mulkay, M. & William, A. T. (1971) A sociological study of a physics department. *British Journal of Sociology* 22:68–80. [SC]
Murphy, R. J. L. (1978) Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology* 48:196–200. [DL]
(1982) A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology* 52:58–63. [DL]
National Research Council (1988) The behavioral and social sciences: Achievements and opportunities. *National Academy Press*. [MJM]
Nelson, L., Satz, P., Mitrushina, M., Van Gorp, W., Cicchetti, D., Lewis, R. & Van Lancker, D. (1989) Development and validation of the Neuropsychology Behavior and Affect Profile. *Psychological Assessment: A Journal of Consulting and Clinical Psychology* 1:266–72. [rDVC]
Newman, H., Freeman, F. & Holzinger, K. (1937) *Twins. A study of heredity and environment*. University of Chicago Press. [PHS]

Newman, S. H. (1966) Improving the evaluation of submitted manuscripts. *American Psychologist* 21:980–81. [aDVC]
Nisbett, R., & Ross, L. (1980) *Human inference: Strategies and shortcomings in human judgments*. Prentice-Hall. [HLR]
Nisbett, R. E. & Wilson, T. D. (1977) The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology* 35:250–56. [RFB]
Noble, J. H. (1974) Peer review: Quality control of applied social research. *Science* 185:916–21. [aDVC]
Nunnally, J. C. (1978) *Psychometric theory*, 2nd ed. McGraw-Hill. [aDVC]
Orlinsky D. & Howard, K. (1978) The relation of process to outcome in psychotherapy. In: *Handbook of psychotherapy and behavior change*, ed. S. Garfield & A. Bergin. John Wiley & Sons. [LDN]
Orr, R. H. & Kassab, J. (1965) Peer group judgments on scientific merit: Editorial refereeing. Paper presented to the Congress of the International Federation of Documentation, Washington, D.C. [aDVC]
Over, R. (1982) What is the source of bias in peer review? *Behavioral and Brain Sciences* 5:229–30. [aDVC]
Oxman, A. D., Guyatt, G. H., Singer, J., Goldsmith, C. H., Hutchison, B. G., Milner, R. A. & Streiner, D. L. (1991) Agreement among reviewers of review articles. *Journal of Clinical Epidemiology* 44:91–98. [rDVC]
Patterson, E. H. (1969) Evaluation of manuscripts submitted for publication. *American Psychologist* 24:73. [aDVC]
Patterson, K. & Bailar, J. C. III (1985) A review of journal peer review. In: *Selectivity in information systems: Survival of the fittest*, ed. K. S. Warren. Praeger Scientific. [aDVC]
Peters, C. (1976) Multiple submissions: Why not? *American Sociologist* 11:165–79. [aDVC]
Peters, D. P. & Ceci, S. J. (1982) Peer-review practices of psychological journals: The fate of published articles submitted again. *Behavioral and Brain Sciences* 5:187–255. [aDVC, AMC, DLE]
Pfeffer, J., Leong, A. & Strehl, K. (1977) Paradigm development and particularism: Journal publication in three scientific disciplines. *Social Forces* 55:938–51. [aDVC]
Physical Review & Physical Review Letters (1987) Annual report 1986. [rDVC]
Pollack, I. (1952) The information of elementary auditory displays. *Journal of the Acoustical Society of America* 24:745–49. [DL]
(1953) The information of elementary auditory displays. II *Journal of the Acoustical Society of America* 25:765–69. [DL]
Price, D. de Solla (1963) *Little science, big science*. Columbia University Press. [aDVC]
Reid, L. N., Soley, L. C. & Wimmer, R. D. (1981) Replications in advertising research: 1977, 1978, 1979. *Journal of Advertising* 10:3–13. [aDVC]
Relman, A. S. (1978) Are journals really quality filters? Rockefeller Foundation working papers (conference, May 22–23). Coping with the biomedical literature explosion: A qualitative approach. [aDVC]
Remington, M., Tyrer, P. J., Newson-Smith, J. & Cicchetti, D. V. (1979) Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. *Psychological Medicine* 9:765–70. [aDVC]
Rennie, D. (1986) Guarding the guardians: A conference on editorial peer review. *Journal of the American Medical Association* 256:2391–92. [MJM]
Roberts, W. A. (1976) Failure to replicate visual discrimination learning with a 1-min delay of reward. *Learning and Motivation*, 7, 313–25. [TRZ]
Robertson, P. (1976) Towards open refereeing. *New Scientist* 71:410. [aDVC]
Robinson, W. S. (1957) The statistical measurement of agreement. *American Sociological Review* 22:17–25. [aDVC]
Rodman, H. (1970) The moral responsibility of journal editors and referees. *American Sociologist* 5:351–57. [RC]
Roediger, H. L. III (1987) The role of journal editors in the scientific process. In: *Scientific excellence: Origins and assessment*, ed. D. N. Jackson and J. P. Rushton. Sage. [LLH, HLR]
Rogot, E. & Goldberg, I. D. (1966) A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Diseases* 19:991–1006. [rDVC]
Romanczyk, R. G., Kent, R. N., Diament, C. & O'Leary, K. D. (1973) Measuring the reliability of observational data: A reactive process. *Journal of Applied Analysis* 6:175–84. [JDC]
Rosenfield, N. S., Ablow, R. C., Markowitz, R. I., DiPietro, M., Seashore, J. H., Toulouklan, R. J. & Cicchetti, D. V. (1984) Hirschsprung Disease: Accuracy of the barium enema examination. *Radiology* 150:393–400. [rDVC]
Rosenthal, R. (1979) The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86:638–41. [PHS]
(1984) *Meta-analytic procedures for social research*. Sage. [RR]
(1987) *Judgment studies: Design, analysis, and meta-analysis*. Cambridge University Press. [RR]
Rosenthal, R. & Rosnow, R. L. (1984) *Essentials of behavioral research Methods and data analysis*. McGraw-Hill. [RR]
(1985) *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge University Press. [RR]
Rosenthal, R. & Rubin, D. B. (1978) Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences* 3:377–415. [PHS]
(1982) A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology* 74:166–69. [RR]
Rourke, B. P. & Costa, L. (1979) Editorial policy II. *Journal of Clinical Neuropsychology* 1:93–96. [aDVC]
Rowney, J. A. & Zenisek, T. J. (1980) Manuscript characteristics influencing reviewers' decisions. *Canadian Psychology* 21:17–21. [aDVC]
Roy, R. (1985) Funding science: The real defects of peer review and an alternative to it. *Science, Technology, and Human Values* 10:73–81. [aDVC]
Rubin, D. B. (1982) Rejection, rebuttal, revision: Some flexible features of peer review. *Behavioral and Brain Sciences* 2:236–37. [PHS]
Scarr, S. (1982) Anosmic peer review: A rose by another name is evidently not a rose. *Behavioral and Brain Sciences* 5:237–38. [aDVC]
Scarr, S. & Weber, B. L. R. (1978) The reliability of reviews for the American Psychologist. *American Psychologist* 33:935. [aDVC, LJS]
Schönemann, P. H. (1971) The minimum average correlation between equivalent sets of uncorrelated factors. *Psychometrika* 36:21–30. [PHS]
(1989) New questions about old heritability estimates. *Bulletin of the Psychonomic Society* 27:175–78. [PHS]
Schönemann, P. H. & Wang, M. M. (1972) Some new results on factor indeterminancy. *Psychometrika* 37:61–91. [PHS]
Scott, W. A. (1974) Interreferee agreement on some characteristics of manuscripts submitted to the Journal of Personality and Social Psychology. *American Psychologist* 29:698–702. [aDVC, CAK]
Sharp, D. W. (1990) What can and should be done to reduce publication bias? *Journal of the American Medical Association* 263:1390–91. [SPL]
Shrout, P. E., Spitzer, R. L. & Fleiss, J. L. (1987) Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry* 44:172–77. [rDVC]
Smart, R. (1964) The importance of negative results in psychological research *Canadian Psychologist* 5a:225–32. [aDVC]
Smigel, E. O. & Ross, H. L. (1970) Factors in the editorial decision. *American Sociologist* 5:19–21. [aDVC]
Smith, K. (1977) Letter to the editor: Peer review defended. *The Sciences* 17:5. [aDVC]
Snedecor, G. W. & Cochran, W. G. (1967) *Statistical methods*, 6th ed. Iowa State University Press. [RR]
(1980) *Statistical methods*, 7th ed. Iowa State University Press. [RR]
Solomon, D. L. (1989) Editorial communication. *Biometrics*, June 21. [PHS]
Sommer, R. & Sommer, B. A. (1984) Reply from Sommer & Sommer. *American Psychologist* 39:1318–19. [aDVC]
Soper, H. V., Cicchetti, D. V., Satz, P., Light, R. & Orsini, D. L. (1988) Null hypothesis disrespect in neuropsychology: Dangers of alpha and beta errors. *Journal of Clinical and Experimental Neuropsychology* 10:255–70. [aDVC]
Sparrow, S. S., Balla, D. A. & Cicchetti, D. V. (1984a) The Vineland Adaptive Behavior Scales: A revision of the Vineland Social Maturity Scale by E. A. Doll. I. Survey form. American Guidance Service. [rDVC]
(1984b) The Vineland Adaptive Behavior Scales: A revision of the Vineland Social Maturity Scale by E. A. Doll. II. Expanded form. American Guidance Service. [rDVC]
(1985) The Vineland Adaptive Behavior Scales: A revision of the Vineland Social Maturity Scale by E. A. Doll. III. Classroom edition. American Guidance Service. [rDVC]
Spearman, K. (1927) *The abilities of man*. MacMillan. [PHS]
Spitzer, R. L. & Fleiss, J. L. (1974) A reanalysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry* 125:341–47. [aDVC]
Steiger, J. J. & Schönemann, P. H. (1976) A history of factor indeterminancy. In: *Theory construction and data analysis in the behavioral sciences*, ed. S. Shye. Jossey-Bass. [PHS]
Steinberg, M., Rounsaville, B. & Cicchetti, D. V. (1990) Interview for DSM-III-R dissociative disorders: Preliminary report on a new diagnostic instrument. *American Journal of Psychiatry* 147:76–82. [rDVC]
Sterling, T. D. (1959) Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association* 54:30–34. [aDVC]
Stevens, J. C. & Tulving, E. (1957) Estimations of loudness by a group of untrained observers. *American Journal of Psychology* 70:600–05. [DL]

Stevens, S. S. (1971) Issues in psychophysical measurement. *Psychological Review* 78:426–50. [DL]

Stinchcombe, A. L. & Ofshe, R. (1969) On journal editing as a probabilistic process. *American Sociologist* 4:116–17. [rDVC, SC]

Stumpf, W. E. (1980) Letters: "Peer" review. *Science* 207:822–23. [aDVC]

Summary Report of Journal Operations (1969) *American Psychologist* 44:1070. [aDVC]

Surwillo, W. W. (1986) Anonymous reviewing and the peer review process. *American Psychologist* 41:218. [aDVC]

Thomas, G. J. (1982) Perhaps it was right to reject the resubmitted manuscripts. *Behavioral and Brain Sciences* 5:240. [aDVC]

Thomas, H. (1985) On the "file drawer" problem (unpublished manuscript). [PHS]

Tinsley, H. E. A. & Weiss, D. J. (1975) Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology* 22:358–76. [LLH]

Torgerson, W. S. (1958) *Theory and methods of scaling*. Wiley. [DL]

Tyrer, P., Cicchetti, D. V., Casey, P. R., Fitzpatrick, K., Oliver, R., Balter, A., Giller, E. & Harkness, L. (1984) Cross-national reliability study of a schedule for assessing personality disorders. *The Journal of Nervous and Mental Disease* 172:718–21. [rDVC]

Tyrer, P., Owen, R. & Cicchetti, D. V. (1984) The Brief Scale for Anxiety: A subdivision of the Comprehensive Psychopathological Rating Scale. *Journal of Neurology, Neurosurgery and Psychiatry* 47:970–75. [rDVC]

Tyrer, P., Strauss, J. & Cicchetti, D. V. (1983) Temporal reliability of personality in psychiatric patients. *Psychological Medicine* 13:393–98. [rDVC]

Uebersax, J. S. (1981) GKAPPA: Generalized kappa coefficient. *Applied Psychological Measurement* 5:28. [rDVC]

(1982) A generalized kappa coefficient. *Educational and Psychological Measurement* 42:181–83. [rDVC]

(1989) Latent structure modeling of ordered category rating agreement. Paper presented at the annual meeting of the Psychometric Society, UCLA, Los Angeles (A Rand Rand Corp. Note). [rDVC]

Ubersax, J. & Grove, W. (1989) Latent structure agreement analysis. Rand Corp. (A Rand Note). [rDVC]

Volkmar, F. R., Cicchetti, D. V., Dykens, E., Sparrow, S. S., Leckman, J. F. & Cohen, D. J. (1988) An evaluation of the Autism Behavior Checklist. *Journal of Autism and Developmental Disorders* 18:81–97. [rDVC]

Wason, P. C. (1960) On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129–40. [MEG]

Watkins, M. W. (1979) Chance and interrater agreement on manuscripts. *American Psychologist* 34:796–97. [aDVC]

Whitehurst, G. J. (1983) Interrater agreement for reviews for Developmental Review. *Developmental Review* 3:73–78. [aDVC]

(1984) Interrater agreement for journal manuscript reviews. *American Psychologist* 39:22–28. [aDVC, MED]

Wiener, S. L., Urivetsky, M., Bregman, D., Cohen, J., Eich, R., Gootman, N., Gulotta, S., Taylor, B., Tuttle, R., Webb, W. & Wright, J. (1977) Peer review: Inter-reviewer agreement during evaluation of research grant applications. *Clinical Research* 25:306–11. [aDVC]

Wilson, E. B. (1928) Review of "The Abilities of Man, Their Nature and Measurement," by C. Spearman. *Science* 67:244–48. [PHS]

Wilson, J. D. (1978) Peer review and publication. *Journal of Clinical Investigation* 61:1697–1701. [PT]

Wolff, W. M. (1970) A study of criteria for journal manuscripts. *American Psychologist* 25:36–39. [aDVC]

(1973) Publication problems in psychology and an explicit evaluation schema for manuscripts. *American Psychologist* 28:257–61. [aDVC]

Wright, R. D. (1970) Truth and its keeper. *New Scientist* 45:402–04. [aDVC]

Wyer, R. S., Greenwald, A. G. Bernard, H. R., Crandall, R. & Anon. (1987) Comments on "The publication game." *Journal of Social Behavior and Personality* 2:13–22. [RC]

Yotopoulos, P. A. (1961) Institutional affiliation of the contributors to three professional journals. *American Economic Review* 5:665–70. [aDVC]

Ziman, J. (1968) *Public knowledge: The social dimension of science*. Cambridge University Press. [AMC]

(1976) *The force of knowledge: The scientific dimension of society*. Cambridge University Press. [AMC]

Zuckerman, H. & Merton, R. K. (1971) Patterns of evaluation in science: Institutionalization, structure, and functions of the referee system. *Minerva* 9:66–100. [aDVC, LLH]

IAN I. MITROFF AND DARYL E. CHUBIN:

Peer Review at the NSF: A Dialectical Policy Analysis

Social Studies of Science, **9** (1979) 199-232

The controversy over peer review is viewed as a dialectic. The arguments espoused by advocates and critics of the system wherein research proposals are evaluated by advisors to funding agencies are reviewed, particularly the findings of two recent studies of peer review at the National Science Foundation. These findings seem to establish merit as the primary factor in the recommendations of peer reviewers to fund proposals. The findings also beg several questions as to 'acceptable' definitions of meritoriousness and innovativeness, the links among belief, perception, and evaluation, and the sanctioned operation of particularistic factors in the review process. Future studies, it is suggested, must include psychological variables — especially measurement of applicants' and reviewers' 'cognitive styles' — if data are to narrow gaps in knowledge and inform the debate itself. Finally, three models which undergird views of peer review are discussed and related to key social issues in the debate.

In recent years, scientific controversies have, with growing regularity, attracted public scrutiny and debate. The controversy over the nature and functioning of the peer review system is an outstanding case in point. That this controversy strikes to the heart of science's most sacred and cherished values — institutional and political autonomy vis-a-vis the external society — may account for the intensity of the debate.[1] That the debate did not reverberate through the American scientific community at large until 1975, however, suggests two complementary realities that may have forestalled the definition of peer review as a pressing and researchable problem: (1) the sanguinity of scientists during the halcyon years of growth in federal funding for R&D and graduate training;[2] and (2) the tenacity of certain values which undergird the

very institution of science, precluding systematic investigation of mechanisms by which the institution's autonomy, self-governance, and 'uneasy partnership' with government is maintained.[3]

As sanguinity and tenacity have flagged, a more defensive posture has emerged, resulting in empirical investigation of contemporary peer review. The results of such investigation occasion this paper, the purpose of which is fourfold:

1. to outline the nature of the debate: that is, to present systematically the position of the contending parties;
2. to review critically some of the evidence, particularly that emanating from two studies of peer review at the National Science Foundation, which bears upon the debate;
3. to raise issues not addressed in recent studies, but which bear fundamentally on the debate; and
4. to propose a strategy for future studies that will clarify old and new issues and hasten collection of appropriate data for informing, if not resolving, the debate — a debate which centres on peer review as an evaluative mechanism in the execution of science policy.

Before proceeding, two caveats must be sounded, lest we be misconstrued. Firstly, notwithstanding our commentary, we are neither anti-science not anti-peer review. Indeed, we regard peer review in principle as the best available system; this does not mean that the system in practice cannot be improved. Likewise, and one would think with greater ease, studies of peer review *can* be improved, not merely in terms of measurement and modes of analysis, but in approach. If scientists refuse to be reflexive, sceptical, and probing of their own institution — its organization and management — then can they really decry congressional 'incursions' into their policies and practices? Part of the responsibility of autonomy and self-governance is self-scrutiny. This is a reasonable expectation, yet too few dissenting voices within the scientific community tend to be heard.[4]

The second caveat is closely related to our first, but pertains to the scope of this paper. We regard peer review as a kind of science advice involving select members of the community, all of whom act — to a greater or lesser degree — as gatekeepers. These gatekeepers help to regulate the flows of information and fiscal resources through the community by directing, impeding, and expediting flows based upon judgments of quality and merit, allegiances and biases and, probably, on sheer caprice as well.[5] The point is that

science advising entails the disposition of scholarly work (such as grant proposals and manuscripts) by referees representing, but not representative of, the scientific community.[6] It is the linking of advice with ultimate dispositions (that is, decisions) which endows peer review with a distinctive content. Clearly, we seek to generalize in this paper about content, recognizing that the form of peer review varies. For example, whereas the National Institutes of Health use a system of study sections, NSF uses review panels.[7] Whereas some referees act as ad hoc mail reviewers, others attend periodic meetings as panel members. But what counts is that multiple judgments are solicited and weighed differentially (depending on the source) to reach a decision: to fund or not to fund. What we shall argue is that defensible decisions are not inevitably guaranteed by the peer review mechanism. Indeed, the process can be used to justify any decision. The power vested in the mechanism or process is derived, in large part, from the power (for instance, reputation) of the referee-advisor-gatekeeper, and is rationalized by the system.[8] To reiterate, it is to this content we shall generalize, though our data are of a more limited form. The philosophy underlying the mechanism of peer review (at least in the US) warrants such substantive generalization.

The Debate: A Dialectical Statement of the Issues

The mode of presenting the peer review debate can help to elucidate the substance of the debate itself. Insofar as a debate features arguments sampled from a continuum of opinion, those arguments can be presented in the form of a dialectic.[9] However, since this way of posing policy issues may not be as familiar as other forms, it might be helpful to offer at least a brief exposition of dialectical analysis.

In recent years scholars in those diverse fields now subsumed under the 'social studies of science' rubric — particularly in history, philosophy, sociology, and management science — have advanced a series of potent theoretical arguments for a dialectical treatment of policy issues: indeed, they have argued that the conduct of science is dialectical in its basic structure.[10] The essence of the argument is as follows:

(1) most social issues, and for that matter, topics on the leading edge of the sciences (natural as well as social) are conflictual in nature: that is to say, it is difficult, if not impossible, to secure widespread agreement (at least initially) as to their basic definition, let alone their solution;

(2) the failure to secure agreement is not because such issues inherently defy treatment or analysis, but because various parties, due to their respective social, intellectual, and/or value positions, will perceive the same issue in very different ways: in a word, parties at interest bring fundamentally different background assumptions to the same issue; as a result, they tend to develop various interpretations of the same set of data (observations or 'facts');

(3) by themselves, data or facts may not be sufficient to resolve the dispute between contending parties, but may actually serve to intensify it;[11] therefore, rather than presume and depend upon initial agreement between parties, what is required is a method for identifying the disparate assumptions that parties bring to an issue and its debate.

Table 1 is a dialectical representation[12] of the views of the proponents (pro) and the critics (con) toward the peer review system, as currently used by NSF. A careful reading of the report by the US House of Representatives on National Science Foundation Peer Review,[13] plus related documents by the proponents and critics of the present system,[14] clearly reveals the operation of two distinct sets of assumptions about peer review. This means that in Table 1 *for every assumption or contention we have identified as characteristic of the position of one side, we have identified a counter-assumption which is characteristic of the other side. Not only are the assumptions on each side strongly held by their proponents, but they are maximally opposed as well. For each assumption which is characteristic of the one side, there is an equally strong assumption on the other such that the two assumptions are the diametric (or nearly diametric) opposite of one another.*[15]

This characteristic procedure is largely responsible for making the dialectic a distinctive means of conducting policy analysis.[16] By aligning the positions side by side, the method explicitly contrasts and draws out the implications of each. It not only shows what each position affirms (that is, what it alone entails) but it also shows explicitly that to which it is maximally opposed. Clearly, no position, no matter how internally consistent and comprehensive it is, is ever completely self-contained. As a result, no position can be

TABLE 1
A Dialectical Representation of the Current System of NSF Peer Review

Basic Assumptions and/or Contentions	
PRO the Current System	CON the Current System
1. The current system is open; it is free from substantial bias.	1. The current system is closed; it contains substantial bias ('an incestuous buddy system')
2. The system leans over backwards in favour of the maverick.	2. There is a natural bias against revolutionary and innovative ideas.
3. It is possible for programme managers to manipulate the system to get the review they want but this is not being done.	3. Programme managers do manipulate the system to get the reviews they want.
4. Proposals should not be 'blind reviewed' since it is not only difficult to conceal completely the identity of a proposer but it is 'a significant factor in determining the likelihood of success of a project'.	4. Proposals should be 'blind reviewed' so that 'the reviewer cannot play favourites or be biased by his knowledge or ignorance of the proposer'.
5. Reviewers should not be selected at random because the most knowledgeable persons would thus be eliminated.	5. Reviewers should be selected at because this would 'eliminate the possibility of the programme manager purposefully biasing the review through selecting reviewers whose opinions he can predict'.
6. '. . . the system should be designed on the presumption that programme managers and reviewers are, on the whole, honest and ethical, but that vigilance should be maintained over the system in such a way as to insure that unscrupulous acts are rare.'	6. It is 'best to design decision-making systems defensively, i.e., on the presumption that the proportion of dishonest or unscrupulous people among (NSF) programme managers and reviewers is high enough to cause severe problems if those people have a significant opportunity to turn the system to their advantage.'

TABLE 1 (continued)

Basic Assumptions and/or Contentions	
PRO the Current System	CON the Current System
7. 'Applicants should receive verbatim reviewer comments or requests but should not know the identity of reviewers;' 'reviewers will be more candid on all aspects of the proposal . . . if their [identities] . . . are kept confidential.'	7. 'Applicants should receive signed verbatim peer reviews or requests;' 'openness would result in more responsible and objective reviews.'
8. There should not be formal appeal procedures for rejected applicants; 'formal appeal procedures will introduce adversary relationships into the scientific community that have heretofore fortunately been missing.'	8. There should be formal appeal procedures for rejected applicants; 'a procedure is needed to check peer review and ensure that important innovations are supported.'
9. NSF should fund less research at colleges and less prestigious universities.	9. NSF should fund more research at colleges and less prestigious universities.

Source: Op. cit. note 13; see also note 15.

fully explicated and understood in terms of itself alone; we need to understand, at a minimum, how a position pertains to an extreme counter-position. One main purpose of a dialectical policy analysis is to make as explicit as possible the points of opposition between different views of an issue. Because it is vitally important to understand on which points parties disagree, Table 1 frames the debate and allows us hereafter to take the term 'policy analysis' as synonymous with a dialectical treatment of peer review.

Faced with a profound disparity of views, one may be tempted to trivialize or demean the position of one side or the other. Thus, for example, Gustafson states:

> A few *conservative* members of the House of Representatives have recently attacked the confidentiality of peer review in NSF and have questioned the integrity of its program officers. For example, Representative R. E. Bauman of Maryland denounced the peer review system in *bitter* terms on the floor of the House.[17]

Gustafson then cites an excerpt from Representative Bauman's remarks which may be, by any standard, 'bitter' indeed. However, whether they are bitter or not is tangential to the issue. That Representative Bauman is a 'conservative' is likewise tangential. Central to the issue is a deep and serious division between points of view that cannot be dismissed by attaching labels of liberal/conservative, bitter/favourably predisposed, and the like.

Other exchanges between proponents and critics demonstrate just how serious the division is, and underscore that it is far from unequivocally a case of one side being right, and the other being wrong; rather, the situation involves two distinctly differing points of view, each bolstered by cogent arguments. From its own perspective, each side is 'reasonable':

> *Mr Pressler.* Now, concerning the quality of the meetings being a bit higher if they are held in confidence, I am not sure why that is true in governmental meetings. I wonder if you could give us an example of why that would be true in your (scientific peer review) meetings?
> *Dr Sherman.* It has to do with one particular aspect of human nature. Even though the system is based on project grants, it is nonetheless necessary in the assessment of the project proposals to assess carefully the quality of the individual scientist named on the application. Sometimes, from my own experience with the system, the discussions about the individual's qualifications can be extremely heated. Now, it would seem in terms of the right of privacy of the individual, just because he is applying for funds from the Federal Government, that he should not have to lay out or make open to the public all of the considerations about his particular qualifications. The system can operate without jeopardizing the right of the individual.
> *Mr Pressler.* If there was something being used against that individual that was not entirely true, he should have a chance of rebuttal or to correct any misinformation. Then people would have a way of knowing. *That is the other side of the coin.*[18]
> *Mr Schever.* Recently we [the Congress] have discovered what a cleansing effect openness has. It seems to cure a lot of problems. There may be problems with openness in the scientific decision making process, but we have not anticipated that they would be very serious. Recently, the Congress has moved from secrecy to openness. It used to be that after we heard from people like you [Dr Sherman] in our hearings we would go into executive session and do our markups. When it was suggested that the markups be made in public, with people listening to us talking or negotiating, many feared that the system would break down. Many

> thoughtful members felt that going to public markup sessions was an exercise in sheer idiocy. They feared that the majority and minority would not be able to compromise, and that we could never settle anything or report a bill out of committee. We changed this procedure and I think that everyone agreed that the system has been vastly improved.
>
> You have described the [peer] review system as one based on mores. Mores can change. They can yield to the pressures of changing time and conditions . . . In the past, the lifestyle based on these mores [that is, secrecy] promoted a degree of integrity, decency, and internal fair dealing . . . An awful lot of problems would be solved if we changed our system from one based on confidentiality to one based on openness and fair treatment. I don't think that we are now aware of all the problems connected with an open system, but I am willing to bet that many of these problems would not materialize.[19]

Although these quotations reveal the intensity of the differences between the two sides in the debate, they underestimate the significant propositions on which both sides agree. We have detected at least five points of consensus, namely:

(1) 'No method superior to peer review has been found for judging the scientific competence of proposers. Scientific peers are better able than others to judge the design of proposed work, the importance of proposed work to the scientific field, and the past performance of the proposer. Appropriate peer review procedures generally lead to the support of proposals in a high quality range. Using peer review procedures [NSF] has successfully fostered significant advances in basic science over the past 25 years.'[20]

(2) 'Witnesses agreed overwhelmingly that some form of peer review should continue to be used to assist in the allocation of Federal funds for scientific research. Not a single witness suggested that peer review be abandoned, although several witnesses proposed changes in the decision-making processes of [NSF] — some minor and some major in their potential effects.'[21]

(3) 'While many witnesses avowed that peer review results in the support of high-quality research, some of which is truly innovative, there was not much confidence expressed that peer review consistently leads to the support of innovative research if it challenges the mainstream of scientific thought or if it seems unlikely to succeed. Arguments and the weight of opinion to the contrary were rather persuasive.'[22]

(4) 'The Subcommittee had ample opportunity during the hearings to explore whether Congressional review of individual [NSF] grants should be required in addition to Foundation approval before Foundation's action becomes final . . . Opinion was over-

whelmingly against Congressional review.'[23]

(5) No one ascribed to the extreme position that 'applicants should know nothing about who reviewers were or what they said.'[24]

Such substantial agreement on these propositions makes them no less true or false, however, than the propositions over which there is explicit disagreement. In either case, beliefs may be so deeply held by their proponents that, at least for them, they may be immutable and irrefutable — utterly impervious to evidence. For those who are not so rigidly committed to either side as to dismiss either position a priori, the question is: What evidence, if any, exists or could be produced that bears on the various propositions and could alter the beliefs of proponents and critics alike?

The Evidence

In reviewing the evidence marshalled in support of each side of the debate characterized above, we shall restrict our attention to two studies: one conducted by Deborah Hensler,[25] the other by Stephen Cole and his colleagues.[26] The justification for this restriction is that, based on a review of materials on the NSF peer review system,[27] these two contain the most comprehensive bodies of empirical evidence which speak directly to some of the contentious issues in peer review. The review of each study will consist of describing its focus, data, and principal findings, followed by our assessment of its merits and shortcomings. Finally, we shall discuss the inferences about gaps in knowledge which can be drawn from the current literature on peer review — theoretical and operational gaps which invite further imaginative study to effect closure on key issues in the debate.

The Hensler Study

In 1975-76, Deborah Hensler sent a questionnaire to a 5 percent random sample ($n = 1552$) of approximately 31,000 persons who had served as reviewers of research proposals submitted to NSF during fiscal year 1974. The identical questionnaire was also sent to a random sample of 3256 applicants for NSF grants during the period (a 16 percent sample framed by a population of 20,000 grant applica-

tions). For both the reviewer and the applicant samples, the response rate exceeded 80 percent.

Both the original questionnaire items and the resultant analyses of responses were designed to attempt to address several points of contention in the debate. An initial question concerns the similarity of backgrounds of reviewers and applicants. Are applicants being evaluated by those similar to themselves or by those who are significantly different? In terms of age, geographical location, institutional affiliation, and so on, the major difference Hensler finds between the two populations is that 'applicants are somewhat more likely [at a probability level less than 0.05] than reviewers to belong to a more recent academic generation and to be currently located at a non-PhD granting institution.'[28] In general, applicants *are* being evaluated by their peers, if by 'peer' one means one who is similar in professional and demographic background.

On appraising the peer review process, those participating as ad hoc mail reviewers (45 percent of the total respondents across all divisions [directorates] of NSF) saw the system as 'sound'; half saw it as an 'an acceptable peer review mechanism with some weaknesses', and only 4 percent saw it as 'a questionable peer review mechanism with many weaknesses'.[29] In comparison, of those participating as panel review members, 60 percent saw the system as 'sound'; 34 percent as 'acceptable with some weakness'; and 5 percent as 'questionable with many weaknesses'. Analysis of the scaled and open-ended responses generally supports the contention that 'reviewers' assessments of the peer review process based on their experience during the past two years are largely positive.'[30]

The issue of whether particularistic factors may intervene in the evaluation process and influence the reviewer's recommendation to fund was broached by three variants on a single question in the Hensler survey: Given two equally good proposals except for one marked difference, which proposal did the respondents think had a better chance of receiving peer review recommendation to fund? In the first case, one of the proposals was from a well-known institution; the other from a lesser-known institution. In the second case, one of the proposals was submitted by a young, as-yet not established principal investigator (PI); the other, by an older, well-established PI. In the third case, one of the proposals featured approaches which were consistent 'with the mainstream of thought' in the discipline or research area; the other, a project which challenged the mainstream of thought. This last case, of course, concerns

one of the pivotal issues in the peer review debate: whether or not the system is biased against innovative ideas.[31]

The responses to these three questions were unequivocal: 52 percent of the reviewers and 61 percent of the applicants felt that the proposal from the well-known institution had a better chance of being funded; 29 percent of the reviewers and 16 percent of the applicants felt both had an equal chance; less than 3 percent of each group felt the proposal from the lesser-known institution had a better chance. The responses to the other two questions were similar: the older PI and the 'mainstream' proposal, respectively, were favoured.

It is particularly instructive to compare the responses of applicants from (a) institutions that are among the top 20 in securing federal research funds[32] with (b) those located at other PhD granting institutions. Whereas 28 percent of the applicants from the top 20 believed that proposals from both the lesser and the well-known institution have an equal chance of being funded, only 14 percent of the applicants from the other institutions believe this to be the case. Even more revealing is that 39 percent of the 'top 20' applicants believe the proposal from the well-known institution has a better chance, whereas 63 percent of the applicants from the other institutions believe that the well-known institution fares better in the competition.[33]

Finally, how do the respondents regard the confidentiality or openness of the peer review process? Hensler summarizes her respondents' views as follows:

> A substantial majority of reviewers and applicants approve of NSF's new policy of providing verbatim review comments to applicants. About two-thirds of the applicants surveyed indicate they personally would have found verbatim review comments useful the last time they submited a proposal to NSF. Respondents who favor a policy of providing verbatim reviews to applicants say the reviews help applicants to understand the reasons for reviewers' reactions, permit applicants to judge reviewers' competence and provide a possible basis for rebutting reviews. A minority of reviewers — nineteen percent — would approve of a policy of identifying reviewers to applicants and thirty-five percent say they would refuse to continue as reviewers if such a policy were adopted. But close to one-third of the applicants would approve of such a policy. Among the applicants who have *not* also served as reviewers recently more than forty percent would approve of identifying reviewers. Applicants from more recent academic generations and those affiliated with academic institutions which are *not* among the top twenty in federal research funding are more likely to approve of identifying reviewers to applicants, than others. Applicants with recent or previously

> unsuccessful experience obtaining NSF funds are most likely to approve of identifying reviewers. Those who disapprove of identifying reviewers feel that this would lead to lower quality reviews, more difficulty securing participation of reviewers and cause bad feelings among colleagues in the scientific community, among other results.[34]

Reactions to the Hensler Study

While the Hensler survey was the first to document so extensively the perceptions of scientists who had participated in the NSF peer review process, the study was not without its limitations (as Hensler openly acknowledges). For example, the study was initiated by a NSF committee the composition of which is not specified in the report.[35] Originally, the committee was interested in the views of three groups: NSF peer reviewers, recent applicants for NSF funds, and researchers who had served NSF neither as reviewers nor as recent applicants. It is reported that

> after some consideration, the Committee decided that it was not feasible to survey the latter group. But by drawing from the Foundation's files of reviews and proposal actions, it was possible to select two independent samples of recent NSF reviewers and applicants.[36]

We have no way of ascertaining on what grounds it was decided 'not feasible' to survey the opinions of researchers who have neither served as reviewers for NSF proposals nor applied recently for NSF funds. Insofar as the Hensler study demonstrates dramatically the link between one's experiences with and perceptions of the NSF system, it would seem desirable to pursue this feasibility question. We shall argue later that it would seem not only highly desirable but imperative to survey the 'null' group: those who, for whatever reasons, have chosen not to interact with NSF or whose interaction has not been sought by NSF. What are the demographic characteristics of those scientists not interacting with NSF? What are the reasons they give for not doing so? What reasons would NSF personnel give? A study of the attitudes of this null group would seem necessary before general conclusions about the equity of NSF's peer review system can be drawn.[37]

Because the Hensler survey does not 'tell us all that we would like to know about NSF reviewers' and applicants' experiences with the

NSF review process and their attitudes toward this process',[38] future surveys (or other studies, for that matter) must augment our knowledge claims. To wit, we would like to know more about their experiences with NSF, but we should also know more about them qua working scientists, and in relation to their views of science. *The Hensler study precludes inferences about NSF's peer review process not only by restricting the range of questions asked of the respondents, but also by restricting what we know about the respondents.* Even if one retained the current set of questions asked about the process, other questions about the respondents should be asked.

For example, much is made in Hensler's report of the finding that a substantial proportion of both reviewers and applicants feel that a proposal which is consistent with the mainstream of thought in an area stands a better chance of being funded; indeed, 53 percent of the reviewers and 60 percent of the applicants who have been recently declined for a NSF grant and who, in addition, have been previously unsuccessful in securing a grant, concur with this proposition. Given the data and the contention of Congressmen, among others, that the peer review system is generally unresponsive to new or innovative ideas, it would seem especially desirable to seek out and secure the views and experiences of those who can be identified as 'innovative'.

If the measurement (if not the definition) of 'innovativeness' is fraught with difficulties, then calling for the study of those judged to be particularly innovative might merely seem to exchange one can of worms for another. Nevertheless, there now exists a growing literature in the social psychology of science[39] which makes it possible to identify and to differentiate empirically 'more innovatively' minded from 'less innovatively' minded scientists. It would be germane to the debate to know the views of those scientists who may be classified as innovators. Are they as a group more sensitive to the perception of bias (or its absence) in the system? Are they even more sensitive to the lack of receptivity afforded innovative ideas? Are they less likely to apply for a NSF grant because of their perceptions, rightly or wrongly, of NSF? Or is it rather because of their particular innovativeness that they are able to play grantsmanship — that is, to clothe novel ideas in mundane or conventional terms?[40]

Overall, the most serious deficiency of the Hensler study is its fundamental concern with attitudes — that is to say, with what

reviewers and applicants *believe* to be the operation of the NSF peer review system — and not with more direct evidence of how it actually operates (which the study never purported to establish). Lest our intent be misunderstood, we are not thereby disparaging the value of the study. What scientists believe about an institution that vitally affects them is key information. The fact that so many scientists would oppose the disclosure of names of reviewers is important information in its own right. It can inform those in power that strong opposition awaits if a policy of disclosure were instituted. This finding does not indicate, however, to what extent and what forms such opposition might take, or whether the benefits of a new policy might so outweigh the disadvantages that the change would be worthwhile.

It is not that the Hensler study is merely a survey of beliefs, but that it fails to penetrate to the heart of the debate, and consequently, does not aid in its resolution. To facilitate movement towards resolution, at the very least, other crucial beliefs would have to be exposed; at best, there would have to exist some other method(s) for assessing the actual state of the system. As they stand, *the Hensler data do not prove that the NSF peer review system is either biased or unbiased, but that there are sizeable numbers of scientists whose experiences (that is, reviewers versus successful applicants versus unsuccessful applicants) predispose them to support one side or the other of the debate.*

The Cole, Rubin and Cole Study

The Cole, Rubin and Cole study represents the most ambitious project to date to determine the actual operation of the NSF peer review system, at least in its basic research programmes.[41] Commissioned by NSF on behalf of the National Academy of Sciences, the study by Cole and his colleagues (hereafter referred to as Cole) seeks to provide evidence that is independent of scientists' beliefs or attitudes about the presence or absence of biases in the NSF peer review system. The kind of evidence sought by Cole thus augments that yielded by the Hensler survey. Specifically, Cole conducted

> seventy in-depth interviews with scientists involved at all levels of the peer review system, including program directors, former program directors, mail reviewers, review-panel members and supervisory-level NSF officials. We also scrutinized

> more than 250 specific research proposals, read all of the peer review comments on those proposals and examined all of the correspondence between the applicant and the program director . . . In addition, we conducted a quantitative analysis of 1200 applicants to the NSF in the fiscal year 1975. (Roughly half of the applicants were ultimately awarded grants.) The purpose of the quantitative study was to identify those characteristics that were correlated with the receipt of a grant from the NSF.[42]

The characteristics chosen for analysis (basically a series of multiple regressions) consist of nine 'social stratification' variables, including rank of PhD-granting department, current academic rank, and three measures of publication and citation. These variables were then correlated with the ultimate disposition of a grant proposal measured in the aggregate as 'percentage of applicants receiving grants' and 'ratings received on proposals' (trichotomized as high, medium, and low).[43]

Taken together, analyses of these variables are intended to test the validity of two hypotheses which dominate the peer review debate. The first (the 'old-boy' hypothesis) lacks 'conceptual clarity', according to Cole. Does old-boyism refer to 'investigators with a common view of their field', 'networks of friendships', or to 'social position' ('level of eminence')?[44] The second (the 'rich get richer' hypothesis) stipulates that particularistic factors (that is, those unrelated to the merit of a proposal) result in an unfair advantage (for example, for the more eminent and/or those located in high ranked departments) in gaining grant approval.[45]

Based on their quantitative analyses, Cole interprets the evidence as a refutation of both hypotheses:

> The overall pattern of our data suggests that scientists with an established track record, many scientific publications, a high frequency of citations, a record of having received grants from the NSF and ties to prestigious academic departments have a higher probability of receiving NSF grants than other applicants do. *Nevertheless, the granting process is actually quite open and there is nothing approximating a scientific caste system.*[46]

> Of the variance that can be accounted for in funding decisions, the peer review rating (among the social stratification variables) is by far the best predictor.[47]

> . . . a scientist's past performance as measured by citations of his work and his recent NSF funding record does lead to a very slight accumulative advantage, but his academic affiliation does not appear to give him any advantage.[48]

Not surprisingly, in summarizing the results of their study thus far, Cole suggests that

> the scientific enterprise is an exceedingly equitable, although highly stratified, social institution in which the individuals who produce the work that is most favourably evaluated by their colleagues receive the lion's share of the rewards.[49]

Reactions to the Cole Study

We think the Cole study begs several questions that are vital to the peer review debate. To compound this error, the authors make some definitive-sounding extrapolations that seem unfounded by their data.

First, to conclude that 'the peer review rating is by far the best predictor' of the probability of receiving a grant by no means suggests that this rating is a good predictor. Indeed, '89 percent of the observed ratings is left unexplained by the nine variables.'[50] This would indicate that factors other than 'social stratification' variables are at work. Yet no such factors are either employed in the analyses or conjectured in discussion of those analyses. Thus the finding that 'individuals who produce the work that is most favourably evaluated by their colleagues receive the lion's share of the rewards' circumvents the questions of why the work is favourably evaluated. No measure of its significance or innovativeness is presented; we are simply asked to believe that voluminous citation of articles denotes their high quality.[51] (After all, proposals which seek to extend such widely-recognized work must be of sufficient merit to justify NSF's decision to fund.) If these results accurately describe peer review in the basic research programmes of NSF, then Cole must further show why this system is 'extremely equitable, although highly stratified'. They have not done this, their rhetoric notwithstanding.

Second, what Cole recognizes in the data, but overlooks in the interpretation, is that evaluation of a producer is hopelessly intertwined with the evaluation of his or her product in science. If producer and product cannot be separated analytically, then we must ask: Where in the distribution of peer review ratings does particularism tend to prevail? If high consensus is achieved in both tails of the rating distribution (as the data attest),[52] then the critical region of peer review is in the middle. This is the grey area where particularistic factors (such as the applicant's present affiliation or institution of PhD) colour the perceived quality of the proposed research. Since performance or 'track record' (that is, reputation) is supposed to be an explicit factor in reviewing for NSF, particularism has been institutionalized as a (partial) rationale for making both favourable and unfavourable decisions to fund. The contradiction is legitimate; tension between universalism

and particularism is built into the peer review process. Why deny this fact, as Cole seems to do? And why present no data (for example, from interviews) which might contain clues about the tension — namely, the extent to which perceptions of quality are coloured by particularistic considerations?[53]

Third, if we ponder — as we did in our reaction to the Hensler study — the kind of data needed to advance the peer review debate beyond its present impasse, our thoughts return to the characteristics of reviewer and/or applicant which might influence the final decision to fund or not to fund a proposal.

Recent studies of the cognitive-styles of inquiry of scientists reveals that one of the key dimensions distinguishing various styles and scientists from one another is the ability to make, as well as to appreciate, fine differentiations between people, objects, or institutions.[54] Persons who excel at this ability are called 'high differentiators'. They, in short, have a high tolerance for ambiguity. As Gordon and Morse put it:

> High differentiators perceive their environment as a series of discrete parts while low differentiators see their environment as highly homogeneous . . . The ability to differentiate manifests itself in two related ways depending on the nature of the stimulus, human or inanimate. In interacting with people the high differentiator perceives and reacts to each as a unique individual possessing a combination of capabilities and inabilities. The low differentiator perceives people as being more or less alike and thus tends to suppress or ignore individual capabilities.[55]

The point is that low differentiators would tend to see the personal characteristics of an investigator as irrelevant to a proposed investigation because they would see all investigators in a similar light. High differentiators, on the other hand, tend to see personal characteristics as very relevant. Specifically, then, does the sample studied by Cole contain an overabundance of low differentiators? Does a large sample tend to mask or damp out the effect of high differentiators? Does the institutional or social process of rating proposals induce even a high differentiator to act like a low differentiator? That is, does the social process of rating proposals foster a 'do-unto-others-what-might-be-done-unto-you' approach? In short, if we had a sample of clearly identified high differentiators and another of clearly identified low differentiators and we gave each the same set of proposals to rate, would their ratings be the same? Before one can reject the hypothesis that the characteristics of the rater/reviewer and those of the individual being rated are irrelevant to the ultimate disposition of a proposal, one must at least attempt to construct a kind of experiment to test the hypothesis.

Until this is done, *the Cole data and analysis cannot be used to ignore or deny the relevance of personal or cognitive attributes in the operation and understanding of peer review.*

Finally, we feel compelled to remark that despite Cole's claim of 'complete autonomy from NSF in conducting' the research,[56] the research they have reported betrays a commitment to show that even where peer review is not equitable within NSF, the inequity is for the good of science;[57] put another way, inequity is functional for the maintenance of the system — and peer review is the tool of this handiwork. Though they began with good intentions, Cole and his associates may have done more to defend the status quo than to inform the debate on peer review: their evidence has yet to sustain the weight of their conclusions. We eagerly await their complete results.

Further Reflections on the Debate

Suppose that there existed a method of establishing 'conclusively' whether the peer review system was either biased or unbiased.[58] If we take the Hensler findings at face value, then we must acknowledge that a significant number of scientists believe that the NSF peer review system is biased, while another significant number believe that it is *not* biased. One could then construct, in ideal-typical fashion, the contingencies represented in Table 2, where the rows represent the beliefs or judgments of scientists as to whether the NSF peer review system is perceived as biased or unbiased.

TABLE 2
Beliefs Versus System States
In Peer Review

		STATE OF SYSTEM	
		Unbiased	Biased
BELIEF OF SCIENTIST	Unbiased	I Correct	II Problematic
	Biased	IV Problematic	III Correct

The columns represent the admittedly oversimplified case where the actual state of the system is either biased or unbiased.[59]

Cases I and III represent the supposedly 'true' or 'correct' situations, where the system is either unbiased or biased and the perceptions or beliefs of scientists match the correct state of the system. Cases II and IV represent the more interesting and 'problematic cases':[60] we can make this judgment, and claim that these two belief conditions demand special examination, even if we cannot determine the absolute state of the system. Suppose for a moment that the NSF peer review system is biased. What, if anything, would it take to convince the sizable number of scientists who believe that it is unbiased to think otherwise? A body of social psychological evidence and arguments suggests that on the whole scientists are conservative in their judgments,[61] and that those who select a career in science partially do so because they have an overly developed need to believe in the orderliness of the world, if not in its ultimate rationality.[62] Consider, too, the oversocialization argument: the vast majority of scientists are trained for normal, workaday science and not for great or extraordinary science.[63] They are neither trained or interested in challenging old theories,[64] let alone prepared to invent novel or 'revolutionary' theories. At the same time, since strong evidence and arguments exist that the system of science is strongly élitist in its structure and orientation,[65] Case II cannot be dismissed or ignored.

To state the matter somewhat differently, Cases II and IV represent situations of denial or projection. Case II represents the situation of denying there is a problem when there is; Case IV represents the situation of asserting there is a problem when there is not. Case II entails the classic phenomenon of identifying with the aggressor, where in order to ease the painful admission of being the underdog, the underdog or victim overly identifies with the values of the aggressor. The question is: How many of those scientists saying that the NSF peer review system is unbiased are identifying, consciously or unconsciously, with the values of élite scientists? For analytical purposes, scientists are constantly being grouped into 'élite' versus 'non-élite'. Given the endless jostling for position that goes on in academic life, plus the constant ratings of departments and institutions to which scientists are subjected, we can plausibly assume that scientists themselves are aware of their relative standing.[66] What does it do to the self-esteem of scientists to know they are located in an élite or non-élite department or institution? Can we expect this

to have no effect on the operation of the system — or, at the very least, on their beliefs about the system?

Indeed, the question that now emerges is whether scientists differ systematically by discipline, institution, or research area in their beliefs of the presence or absence of bias in the system. This question fuels speculation on the relation between individual and social (systemic) innovativeness hypothesized earlier: Are individuals whose cognitive-style betrays a high propensity to innovate viewed as such by their colleagues? Does this research in fact reflect their innovativeness? Furthermore, do these innovators communicate more frequently with other innovators, and in this sense nurture one another? Finally, and most importantly, are innovators located in greater or lesser abundance at prestigious institutions?

The evidence that élite scientists tend to associate and communicate with other élite scientists more frequently than they do with non-élites, and that élite scientists tend to be affiliated with élite institutions,[67] would suggest that an examination of the interplay of psychological, intellectual and social factors operating in the peer review process is in order. Above all, if sustained innovativeness and éliteness go hand in hand, the concentration of innovative ideas and high quality research proposals submitted by those in a select pool of institutions would need no remedy. However, the distribution of high innovators in the scientific community (as well as in the subpopulation of applicants for NSF funds) is unknown.

This missing link in the debate — a control variable, if you will — signals a need to measure the cognitive styles and background beliefs of participants in the peer review process before one can interpret either (1) the meaning of responses to a survey such as Hensler's, or (2) a quantitative analysis (such as Cole's) which sacrifices qualitative insights into individual differences for statistical significance. Finding this missing link is all the more necessary given Mitroff's findings from his Apollo moon study.[68] Nearly all of his scientists were extremely sceptical of the conventional portrait of the scientist as a neutral, unbiased, objective observer of nature. Moreover, the overwhelming majority of those interviewed gave revealing reasons for why they thought scientists in their role as scientists should not be entirely unbiased. The majority view was that it was necessary for scientists to act as partisan advocates for their hypotheses and theories lest those theories suffer a premature death. This means not that these scientists

neglected or refused to test critically their theories, but that their actual conduct of science is more complicated than that portrayed in conventional accounts. The relevance of these findings to the present discussion is this: Can we expect scientists' views on the general operation of science as an intellectual and social system not to influence their views regarding peer review? Were the moon scientists' views exceptional (they too, constitute an élite sample) or more the rule than previously thought? These are fittings topics for future studies, and in our opinion, vital for further assessment of the peer review system.

What we now know better about the peer review debate from the Hensler and Cole studies is summarized in Table 3. *In our judgment, these data may be necessary, but are insufficient, to settle the principal issues. Although they enhance our understanding in dialectical terms, the data underscore the serious gaps that exist in our knowledge — gaps which must narrow if the debate is ever to approach closure.*

Conclusions and Recommendations

The principal conclusion of this review is that the current data are inconclusive to resolve the debate represented in Table 1; nevertheless, some data do exist to support contentions on each side of the debate.

Because issues such as those inherent in peer review expose tensions in the workings of science as a social system, they call forth deep divisions of value. Such issues, therefore, may not be amenable to treatment (and hence, to resolution) via conventional methods. The debate instead calls for treatment of the issues from more than one theoretical point of view. Philosophers of science have long recognized that scientific data can neither be collected in the first place, nor analyzed in the second, apart from some prior theoretical point of view.[69] That is, one does not collect data without having presupposed some hypothesis, theory, or model, no matter how implicit, unconscious, or informal it may be.

We would assert that at least three models undergird the peer debate; (1) the Accumulative Advantage Model; (2) the Political Model; and (3) the Merit Model. The Accumulative Advantage Model derives from the 'Matthew effect', as explicated by Merton:[70] one who has developed a good reputation based on past

TABLE 3

Relation of the Hensler and the Cole Studies to the NSF Peer Review Dialectic

Evidence in Support or Denial of Assumptions	Basic Assumptions and/or Contentions (see Table 1)	
	PRO the Current System	CON the Current System
Assumption	1. Current system is open, free from bias	1. Current system is closed, contains bias
Evidence:	Supporting: similarity of characteristics of reviewers and applicants; vast majority of respondents see system as sound; believe they were treated fairly.	Denying: extremely few respondents see current systems as possessing major flaws or believe they were treated unfairly.
Status of evidence	Weak on both sides of argument; insufficiency of beliefs per se to determine actual operation of system.	
Assumption	2. System encourages innovative ideas	2. System blocks innovative ideas
Evidence:	Denial: less than 7 percent of respondents believe a proposal which challenges 'mainstream' has a better chance of being funded; roughly 20 percent believe both have an equal chance.	Support: nearly half of respondents believe a 'mainstream' proposal has a better chance of being funded.
Status	Weak; insufficiency of data to establish operation of system.	

TABLE 3 (continued)

	PRO the Current System	CON the Current System
Assumption	3. Programme managers do not manipulate reviews	3. Programme managers do manipulate reviews
Status	Not addressed explicitly in Hensler study; data collected, not reported by Cole	
Assumption	4. Proposal should not be blind reviewed	4. Proposals should be blind reviewed
Status	Not addressed explicitly by Hensler; institutionalization of particularistic factors into review process recognized by Cole.	
Assumption	5. Reviewers should not be selected at random	5. Reviewers should be selected at random
Evidence:	Support: only 15 percent of respondents at most believe in randomization; nearly 65 percent believe in some form of judgment sampling in conjunction with NSF staff	Support: roughly 31 percent believe in some form of randomization and judgment sampling
Status	Weak; insufficiency of beliefs to warrant procedural change.	
Assumption	6. System should be designed on presumption of honesty	6. System should be designed defensively
Status	Not tested for explicitly; charge of 'old boy-ism' refuted, according to Cole.	

TABLE 3 (continued)

	PRO the Current System	CON the Current System
Assumption	7. Applicants should not know reviewers' identity	7. Applicants should know reviewers' identify
Evidence:	Support: roughly 40-50 percent of respondents feel knowing name would make no difference; 12-20 percent feel comments would be less useful	Support: 30-40 percent feel comments would be more useful if one knew name
Status	Weak; again, insufficiency of beliefs to warrant procedural change.	
Assumption	8. There should not be formal appeal procedures	8. There should be formal appeal procedures
Evidence:	Support: feeling that a formal process will further bureaucratize system	Support: about 73 percent of applicants favour a formal appeals system as a remedy for mistakes
Status	Endorsement of idea by Hensler; no consideration of procedures.	
Assumption	9. NSF should fund less research at colleges and less prestigious universities	9. NSF should fund more research at colleges and less prestigious universities
Status	'Rich get richer' hypothesis supported in part by Cole data: researchers at more prestigious institutions and with track record slightly favoured in the review process.	

work accrues more advantages (that is, disproportionately) than those lacking such a good reputation. This cumulates over time so that the 'rich get richer'. The Political Model stipulates that certain élite scientists at élite institutions have disproportionately more access (a) to other élite scientists and scarce scientific resources (for example, information and research funding) and (b) to governmental agencies such as NSF, where they exert influence on science policy and its implementation in their roles as gatekeeper, advisor, and peer reviewer.[71] Finally, the Merit Model states that the work of a scientist is judged primarily on its merit, that research monies are awarded competitively according to universalistic criteria which favour, above all, the applicant's current ability to perform.[72]

We hasten to add that merit is a component in each of these models, but is differentially weighted. Again, in terms of the universalism-particularism continuum, we predict that particularistic factors (that is, attributes of the scientist) tend to predispose reviewers to favourable evaluation of the scientist's work. This emphasis is typically an outgrowth of prior, and oft-repeated, evaluation of that scientist's other work as meritorious. Unlike the Merit Model, both the Accumulative Advantage and Political Models recognize this 'contamination' of evaluations. Alternatively put, these models treat discrete research products (for example, a new proposal or book) as continuous in time or imbued with the quality (fixed at a certain threshold, it would seem) of its producer. In brief, evaluation of research is highly contingent on its source. Adherents of these two models would insist that supporting such researchers — 'the best' — is functional for the system; therefore, favourable peer review of the research in question should follow suit. Those operating on the Political Model would rely more on particularistic-factors than on merit of the specific proposal in recommending disposition. Reviewers enamoured of Accumulative Advantage would attend somewhat less to the credentials of the researcher and more to the substance of the proposed research. Finally, those utilizing merit as the chief criterion of funding support would resort to proposal details per se, far more than to characteristics of its author. These models, then, capture the tension inherent in the reviewer role — a tension which encompasses both the discharge of the particular reviewing task and the overarching mentality one brings to the task.[73]

How, then, have these models been applied to studies of the peer review process? In Hensler's study, none of the three models

appears to be presupposed. That is, in the design, conduct and analysis of her survey, Hensler was essentially atheoretical. In contrast, the Cole study embraces the Merit Model. Cole exhibits a marked preference for (if not an a priori belief in) the Merit Model. We have no quarrel with this, as Cole makes the preference explicit, claiming later, as we have seen, that the data tend to support the Accumulative Advantage Model. However, this preference is troublesome if one suspects that something so complex as peer review requires simultaneous and explicit examination from a number of diverse and competing theoretical perspectives. Even stronger, the same set of data ought to be examined from the perspective of each model. Because each, in all likelihood, is partially correct, future studies must establish, for example, under what circumstances each model obtains. What we are advocating is a testing of the alternatives — new data collection and analysis — to expand the empirical base that impinges upon and must eventually mediate the debate.

There are other aspects, however, to the peer review process and to the debate which we have not considered. Foremost among these is the role of the public in shaping the institutions which purportedly operate on its behalf. What does the public want from NSF management? Does this differ from what the scientific community wants? Do NSF practices produce the best science, and are they conducive to the optimal long-term development of knowledge?[74] Is it not the responsibility of the scientific community and federal agencies such as NSF to invite interested lay parties to enter the dialogue among experts,[75] especially when some of the most important persons for whom the studies are being conducted are not scientists?

If the crux of the peer review debate is the analysis of negotiations between science and its environing communities and not solely negotiations within the scientific community,[76] then science must promote research that illuminates both negotiating processes. Like our predecessors, we have emphasized the latter in this paper. The former, however, is an equally, if not more, vexing research problem that will not conveniently fade away.

The study of scientific autonomy and self-governance is really the study of the science-government partnership. What we have recommended is that this study begin by linking the social psychology of the protagonists to their respective roles in the conduct of scientific inquiry. Only then will the debate over peer review

fulfill the promise of a dialectical policy analysis; only then will the debate prescribe changes of policy into practice.

• NOTES

A presentation based on an early version of this paper was made at the Second Annual Meeting of the Society for Social Studies of Science, held in Boston, Massachusetts, and Harvard University, 14-16 October 1977. The incisive comments of P. Thomas Carroll on that version were most helpful in rethinking and rewriting.

1. See T. Gustafson, 'The Controversy Over Peer Review', *Science*, Vol. 190 (2 December 1975), 1060-66; D. Shapley, 'House Votes Veto Power on All NSF Research Grants', ibid., Vol. 188 (25 April 1975), 338-41; Shapley, 'NSF Violations of Personnel Code Alleged', ibid. (31 May 1975), 915; J. Walsh, 'NSF and Its Critics in Congress: New Pressures on Peer Review', ibid. (6 June 1975), 999-1001; Walsh, 'NSF House Appropriations Panel Gives Warning Tug on Purse Strings', ibid., Vol. 189 (4 July 1975), 26-28; Walsh, 'NSF Peer Review Hearings: House Panel Starts with Critics', ibid. (8 August 1975), 435-37; Walsh, 'Peer Review: NSF Faces Changes, the Question is How Extensive', ibid., Vol. 190 (17 October 1975), 253-56.

2. S. MacLane, 'Peer Review and the Structure of Science,' *Science*, Vol. 190 (14 November 1975), 617.

3. G. M. Lyons, *The Uneasy Partnership: Social Science and the Federal Government in the Twentieth Century* (New York: Russell Sage, 1969). See also J. Haberer, *Politics and the Community of Science* (New York: Van Nostrand Reinhold, 1969).

4. Our mention of this caveat stems from two sets of experiences: firstly, in response to the first public draft of this paper, where colleagues rhetorically asked, 'You're not against peer review, are you?'; and secondly, our inability to convince NSF programme managers that despite the recent studies of their system, the 'crucial experiment' had yet to be performed. We claimed that the present system had been legitimated, but not accurately appraised because (1) many of the right questions had not been asked, or if they had, the answers were not carefully weighed (for example, could *not* be quantified), or (2) the crucial data were not in the public domain, but with special consent could be made available for analysis. Consequently, the study whose findings could be used to alter policy and strengthen the operation of the system never materialized. Our interpretation may smack of sour grapes, but we think the opportunity for internal self-scrutiny was lost (although congressional critics were pacified — for the moment). We do *not* think our argument fell on deaf ears; rather, no one was willing to handle a hot potato which had temporarily cooled.

5. For a review, see H. Zuckerman and R. K. Merton, 'Age, Aging, and Age Structure in Science', in M. W. Riley, M. Johnson and A. Foner (eds), *A Theory of Age Stratification*, Vol. 3, *Aging and Society* (New York: Russell Sage, 1972), 292-356.

6. In general, referees and advisors are older, more eminent and published, and located at more prestigious institutions than the 'average' member of the scientific community. See N. C. Mullins, 'The Structure of an Elite: The Advisory Structure of the Public Health Service', *Science Studies*, Vol. 2 (1972), 3-29; L. Groeneveld, N. Kolier and N. Mullins, 'The Advisers of the US National Science Foundation', *Social Studies of Science*, Vol. 5 (August 1975), 343-54; M. J. Mulkay, 'The Mediating Role of the Scientific Elite', *Social Studies of Science*, Vol. 6 (1976), 445-70.

7. Studies of NIH peer review include G. M. Carter, 'Peer Review, Citations, and Biomedical Research Policy: NIH Grants to Medical School Faculty' (Santa Monica, Calif.: Rand Corporation Report R-1583-HEW, 1974); C. Henley, 'Peer Review of Research Grant Applications at the National Institutes of Health', *Federation Proceedings*, Vol. 36 (July 1977), 2066-68, 2186-90, 2335-38; NIH Grants Peer Review Study Team, 'Grants Peer Review: Report to the Director, NIH, Phase I, Vols. I-III' (Bethesda, Md.: National Institutes of Health, December 1976).

8. A provocative formation of this relationship can be found in J.-J. Salomon, 'The Mating of Knowledge and Power', *Impact of Science on Society*, Vol. 22 (January-June 1972), 123-32.

9. By 'dialectic' we mean more than a mere polar opposition or conflict between viewpoints. We mean that viewpoints are intensely opposed to one another, though the meaning of one is dependent on the other; that is, either viewpoint is completely self-contained but is defined, if only in part, through the other. Be this as it may, we are more interested at present in the *operational* use of the dialectic as a unique methodology for analyzing issues, rather than in quibbling about the various historical meanings of the term.

10. C. West Churchman, *The Design of Inquiring Systems* (New York: Basic Books, 1971); P. Feyerabend, *Against Method* (London: New Left Books, 1976); G. Holton, *Thematic Origins of Scientific Thought* (Cambridge, Mass.: Harvard University Press, 1973); A. Kantrowitz et al., 'The Science Court Experiment, an Interim Report', *Science*, Vol. 193 (20 August 1976), 653-56; M. Levine, 'Scientific Method and the Adversary Model: Some Preliminary Thoughts', *American Psychologist*, Vol. 29 (September 1974), 661-77; R. O. Mason, 'A Dialectical Approach to Strategic Planning', *Management Science*, Vol. 15 (1969), B-403-14; R. K. Merton, *Sociological Ambivalence* (New York: The Free Press, 1976); D. Nelkin, 'Thoughts on the Proposed Science Court', *Newsletter on Science, Technology, and Human Values*, No. 18 (1977), 20-31.

11. Churchman, Feyerabend and Mason, ops. cit. note 10.

12. Mason, op. cit. note 10.

13. 'National Science Foundation Peer Review, Volume I', A Report of the Subcommittee on Science, Research and Technology of the Committee on Science and Technology, US House of Representatives, Ninety-Fourth Congress, Second Session (January 1976).

14. For example, Gustafson, op. cit. note 1.

15. Verbatim quotes have been purposefully included in Table 1 to indicate the depth and sincerity with which the respective parties hold their views. In all cases, the assumptions are direct or abridged quotes excerpted from the document cited in note 13.

16. Churchman and Mason, ops. cit. note 10.

17. Gustafson, op. cit. note 1, 1060 (our italics).
18. Subcommittee, op. cit. note 13, 571 (our italics).
19. Ibid., 579-80.
20. Ibid., 2.
21. Ibid., 25.
22. Ibid., 27.
23. Ibid., 41.
24. Ibid., 43.
25. D. Hensler, 'Perceptions of the National Science Foundation Peer Review Process: A Report on a Survey of NSF Reviewers and Applicants', Prepared for the Committee on Peer Review, National Science Board, and the Committee on Science and Technology, US House of Representatives (Washington, DC: NSF 77-33, December 1976); R. Abel, 'Applicants' and Reviewers' Assessments of the NSF Peer Review Process', International NSF draft paper (Washington, DC: NSF, November 1976).
26. S. Cole, L. Rubin and J. R. Cole, 'Peer Review and the Support of Science', *Scientific American*, Vol. 237 (October 1977), 34-41. This article is an interim report on the project, which is still in progress (see note 41, below).
27. One of us (I.I.M.) was asked to undertake such a review for the Office of Planning and Policy Analysis, National Science Foundation, under Contract No. OM, Order No. 77-SP-0370.
28. Hensler, op. cit. note 25, 15-18, quote on 17.
29. Ibid., 23.
30. Ibid., iv. Furthermore, as Hensler explains (considering both successful *and* unsuccessful applicants), principal investigators'

> evaluations of the appropriateness of the review procedures used are related to disposition of the proposal. But even among those whose proposals were declined, half feel the procedures were appropriate. A majority of *unsuccessful* applicants feel that the decision to decline was unfair but a substantial proportion — forty-three percent — feel that [even this] decision was fair. About 84 percent of declinees who thought the decision was unfair, say they would have appealed the decision if a formal appeals process had existed. Assessments of appropriateness of procedures and fairness of the funding decision do not appear to be related to academic generation, institutional affiliation or region. However, those who have served as NSF reviewers or who have received NSF grants in the past are more likely to evaluate their most recent experience positively — even if they were turned down — than those with less successful experience dealing with NSF . . . About 73 percent of the PIs, including both grantees and declinees, would favor NSF adopting a formal appeals system. The reason for supporting such a system which is volunteered most frequently is that it would provide a remedy for mistakes and misjudgments; the leading reason for opposing it is that it will further bureaucratize and burden the review process. (ibid., v-vi).

31. As many writers have indicated, what is perceived as an innovative idea is relative to time and place in any research community. The line separating innovation from charlatanism or, in the lexicon of the exemplary 'science studies' literature, the difference between transgressions of cognitive norms and true anomalies, is fine indeed. Our view is that 'excessively' innovative ideas will so challenge the paradigmatic foundations of a research area that the innovators and their ideas

will neither gain ready access to the literature nor approval of proposals to pursue their research programme. Mainstream thought, in short, can sustain only moderate innovation. The issue, in the context of peer review, is whether the agencies which administer the system and its resources are the guardians of the mainstream or a refuge for innovators. Surely they are a little of both; hence, the issue of bias and evidence are elusive at best. For related discussion, see below, plus T. S. Kuhn, 'Second Thoughts on Paradigms', in F. Suppe (ed.), *The Structure of Scientific Theories* (Urbana, Ill.: University of Illinois Press, 1974), 459-82; M. J. Mulkay, *The Social Process of Innovation* (London: Macmillan, 1972); H. M. Collins, 'The TEA Set: Tacit Knowledge and Scientific Networks', *Science Studies,* Vol. 4 (1974), 165-86; D. E. Chubin, 'The Conceptualization of Scientific Specialties', *Sociological Quarterly*, Vol. 17 (Autumn 1976), 448-76, esp. 459-70.

32. While Hensler does not specify these 20 institutions, the following ten US universities have been identified as receiving more than a third of all federal expenditures in universities, producing about a third of all the doctorates, and providing 37 percent of the members of federal review panels in the 1960s: California, Caltech, Chicago, Columbia, Cornell, Harvard, Illinois, MIT, Michigan and Minnesota. See W. Hirsch, *Scientists in American Society* (New York: Random House, 1968), 106.

33. These findings accord with previous perceptions, as Hensler (op. cit. note 25, 50) observes that 'reviewers in general, and applicants who have also served as reviewers, are significantly less likely to perceive bias in the process than other applicants . . . . Applicants who have not been successful in obtaining NSF grants recently or in the past are most likely to think that process is biased.'

34. Hensler, op. cit. note 25, 84.

35. Ibid., 1.

36. Ibid.

37. It could well be the case that those who have either *experienced* the most 'bias' (no matter how it is defined) or who *attribute* bias to the system may be those who have either (a) 'dropped out' of the system *prior to* the sampling period or (b) never 'dropped in' in the first place. Without sampling this group, such conjectures simply cannot be evaluated; yet they cannot be dismissed out of hand. In this respect, Hensler can be criticized for a uniform lack of conjecture; she apparently feels no compulsion to explain *why* her survey generated the responses it did.

38. This quotation appears in an earlier version of Hensler's report (op. cit. note 25), dated September 1976, on p. 80.

39. G. Gordon and E. V. Morse, 'Creative Potential and Organizational Structure', *Academy of Management Journal*, Vol. 12 (1969), 37-49; I. I. Mitroff, *The Subjective Side of Science, A Philosophical Inquiry into the Psychology of the Apollo Moon Scientists* (New York: Elsevier, 1974); E. V. Morse and G. Gordon, 'Cognitive Skills: A Determinant of Scientists' Local-Cosmopolitan Orientation', *Academy of Management Journal*, Vol. 17 (1974), 709-23.

40. A prior question (underlying all of these) concerns the distribution of innovativeness in the community at large, and whether the purposive sampling of scientists for the role of peer reviewer proportionately captures this characteristic. Given the profile of reviewers developed in the works cited in note 6, one would think so. Indeed, one would think that innovators are *over*represented among reviewers; likewise, one would *hope* that innovators are overrepresented among the recipients of research funds. (As far as we can tell, there are no questions in the

Hensler study dealing directly with the topic of 'grantsmanship', although it is implied by a few of her questions.)

41. Cole et al., op. cit. note 26. The longer version of this report had yet to appear when we were preparing the final revision of this paper. It was published while the paper was in press: see S. Cole, L. Rubin and J. R. Cole, *Peer Review in the National Science Foundation: Phase One of a Study* (Washington, DC: National Academy of Sciences, 1978). We have been unable to incorporate and respond to details of the published report. An earlier draft of our paper benefitted from an unpublished preliminary version of the Cole report forwarded to I.I.M. as part of his review (see note 27). However, since it was the only public version available at the time, we have taken special pains to confine our comments here to the published interim *Scientific American* report.

42. Ibid., 36.

43. Ibid., 37-39.

44. Ibid., 37.

45. Ibid., 38.

46. Ibid., 40 (our italics).

47. Ibid.

48. Ibid., 41.

49. Ibid.

50. Ibid., 38.

51. The use of citations as a measure or indicator of the perceived importance or standing of a scientist within his or her domain of research seems non-problematic to Cole, despite the reservations expressed by some investigators regarding the validity of the measure. See D. E. Chubin and S. D. Moitra, 'Content Analysis of References: Adjunct or Alternative to Citation Counting?' *Social Studies of Science,* Vol. 5 (1975), 432-41; N. Kaplan, 'The Norms of Citation Behavior: Prologomena to the Footnote', *American Documentation,* Vol. 16 (1965), 179-84. Do numerous citations to the work of a scientist truly reflect the long-term importance and significance of the work, or merely its short-term popularity? Does one cite a work to support one's own, or for more critical and rhetorical reasons? Even if it is presumed that the reviewer has knowledge of citation performance, without taking into account the reasons why scientists cite others, and why variations in citation behaviour exist across disciplines and research areas, can one confidently use citation counts as a 'common denominator' predictor variable? On this score, Cole finds — interestingly enough — that past citations contribute little to the explained variance of the funding decision, suggesting to us that citation may be an irrelevant criterion of performance because most reviewers are *ignorant* of an applicant's citation 'performance'.

52. Cole et al., op. cit. note 26, 39.

53. This is really an expression of our dismay over Cole's decision to report only the results of quantitative analyses. Such analyses can mask individual differences manifested in anecdotal accounts, such as interviews. For example, it would be instructive, if not indispensable, to know how NSF personnel view the system of science, and the rationality of the enterprise in which they are engaged. In the study of eminent scientists who investigated the moon rocks returned by the Apollo missions (see note 39), it was found that nearly the entire sample of forty-two scoffed at, and in the most derisive of terms, the stereotypical view of the scientist and science itself as the 'open, free, unbiased exchange of pure ideas' that is so

commonly portrayed in college texts and in popular accounts of science. Surely it is important to know whether a respondent holds a conventional or a radical view of the workings of science before one can properly evaluate the respondent's attitudes towards the peer review system. Surely programme directors differ in discharging their duties. Are those harbouring a less conventional view of science more critical and sceptical of the operation of NSF peer review?

54. Gordon and Morse, op. cit. note 39.

55. Ibid., 42. In the typical situation used by Gordon to measure differentiation, a person is asked to rate ten of his most immediate colleagues, friends, associates, and so on, on a ten-point scale with respect to (a) their productivity, (b) their creativity, and (c) how easy it is to get along with the individual being rated. Low differentiators tend to rate all ten persons identically; in other words, low differentiators make use of only a small portion of the total ten-point scale, whereas high differentiators tend to make significantly more use of the whole scale. High differentiators tend to view people as different and unique; low differentiators view them as the same.

56. Cole et al., op. cit. note 26, 34.

57. To this end, we are surprised that the Cole study was not couched in terms of the Ortega hypothesis which the Coles (J. R. Cole and S. Cole, 'The Ortega Hypothesis', *Science*, Vol. 178 [27 October 1972], 368-75) investigated a few years ago. The Coles' rejection of this hypothesis (that all scientists contribute through their modest research efforts to the incremental progress of science) raises questions as to the concentration of funding support among a small portion of the research community. S. J. Turner and D. E. Chubin, in 'Another Appraisal of Ortega, the Coles and Science Policy: The Ecclesiastes Hypothesis', *Social Science Information*, Vol. 15 (1976), 657-62, argue that to equate the distributions of scientific talent, productivity, and reward is little justification for a science policy that deliberately concentrates resources among the élite that populates one tail of those distributions. Rather, they question the efficiency of a policy that would waste the talents of trained personnel without modifying the organizations that train and employ them (though we realize this is far easier said than done). To sustain the research of more scientists could calculably enhance their contributions. Yet no experiments in the democratization of research allocation have been carried out. Thus the proposition remains untested, and for us at least, the Ortega hypothesis, like the 'old boy' and 'rich get richer' hypotheses to which it is intimately related, has been gratuitously laid to rest by the Coles.

58. An 'unbiased' system would obviously not be one that randomly funds proposals; rather, it would fund primarily according to *merit*, which might be defined as innovative, feasible, relevant, or some combination thereof. In the discussion that follows we assume that a lack of bias is both desirable and attainable.

59. We purposefully use the term 'admittedly oversimplified' because the actual situation may be too complex to admit of the two exclusive categories, 'biased' or 'unbiased'. The actual state of the system may be neither biased nor unbiased, or it may be a condition of both — that is, a complicated mixture of partially 'biased' and 'unbiased' elements. Nevertheless, for the purpose of this analysis, it suffices to consider the 'idealized' cases in Table 2.

60. Notice that we do not say that Cases II and IV necessarily represent 'incorrect cases', since the difficulty in knowing the 'true' state of the actual system also

makes it difficult to know or assess 'error'; the term 'problematic' is more appropriate than such decisive terms as 'truth' or 'error', since complex social systems may not admit of such rigid or precise determinations.

61. P. Feyerabend, op. cit. note 10; A. H. Maslow, *The Psychology of Science* (New York: Harper and Row, 1966); D. C. McClelland, 'On the Dynamics of Creative Physical Scientists', in L. Hudson (ed.), *The Ecology of Human Intelligence* (Harmondsworth, Middx.: Penguin Books, 1970); Mitroff, op. cit. note 39.

62. McClelland, op. cit. note 61.

63. On this issue, Kuhn and Lakatos appear to agree. See S. S. Blume, *Toward a Political Sociology of Science* (New York: Free Press, 1974); T. S. Kuhn, 'The Essential Tension: Tradition and Innovation in Scientific Research', in C. W. Taylor and F. Barron (eds), *Scientific Creativity, Its Recognition and Development* (New York: Wiley, 1963), 341-54.

64. Mitroff, op. cit. note 39.

65. Mulkay, op. cit. note 6.

66. Scientists are not only aware of their relative position (for example, department or institution rank), but they tend to aggrandize their position relative to their perception of other departments and institutions. See T. Caplow and R. J. McGee, *The Academic Marketplace* (New York: Basic Books, 1958).

67. For a review, see Mulkay, op. cit. note 6; M. J. Mulkay, 'The Sociology of the Scientific Research Community', in I. Spiegel-Rösing and D. de S. Price (eds), *Science, Technology and Society: A Cross-Disciplinary Perspective* (Beverly Hills: Sage, 1977), 93-148; and P. Boffey, *The Brain Bank of America* (New York: McGraw-Hill, 1975). For a discussion of the institutional 'halo effect' which blurs the empirical distinction between prestige of institution and scientist's reputation (as a proxy for performance, quality of research, and so on), see H. Zuckerman, 'Stratification in American Science', in E. O. Laumann (ed.), *Social Stratification: Theory and Research for the 1970s* (Indianapolis: Bobbs-Merrill, 1970), 235-57.

68. Mitroff, op. cit. note 39.

69. See, for instance, Churchman and Feyerabend, ops. cit. note 10.

70. R. K. Merton, 'The Matthew Effect in Science', *Science*, Vol. 159 (5 January 1968), 56-63.

71. Mulkay, op. cit. note 67; Boffey, op. cit. note 67; Salomon, op. cit. note 8. See also D. K. Price, *The Scientific Estate* (Cambridge, Mass.: The Belknap Press, 1965; and D. S. Greenberg, *The Politics of Pure Science* (Washington, DC: New American Library, 1967).

72. J. R. Cole and S. Cole, *Social Stratification in Science* (Chicago: The University of Chicago Press, 1973); J. Gaston, *Originality and Competition in Science* (Chicago: The University of Chicago Press, 1973); Gaston, *The Reward System in British and American Science* (New York: Wiley-Interscience, 1978).

73. To quote an anonymous referee (for an earlier version of this paper) on decision-making behaviour:

> My hunch is that, as the uncertainty of peer evaluation increases, more and more of the elements of the dialectic are brought to bear so that in some cases, after 'objective' criteria have been used and are found not to distinguish between pairs of proposals, other more subjective and politically controversial premises are used . . . and I think for quite defensible reasons.

74. We thank an anonymous referee for reminding us to call attention to these questions.

75. This is the spirit of the proposed Science Court — the involvement of various publics in scientific decisions which are 'too important to be left to the scientists'. But another anonymous referee observed:

> Basic to the Science Court concept is the idea of scientific judgment. Its purpose is not simply 'presentation and review' of the issues, but indeed, a verdict that reflects the assessment of a 'scientific judge' . . . I have argued this out with Kantrowitz suggesting that he maintain the 'presentation and review' procedure and minimize the importance of the verdict, but he claims that would change the intention in a fundamental way.

See D. Nelkin, 'The Political Impact of Technical Expertise', *Social Studies of Science*, Vol. 5 (February 1975), 35-54, and op. cit. note 10.

76. As Dorothy Nelkin, reflecting on the recombinant-DNA debate, has recently argued: see her 'Threats and Promises: Negotiating the Control of Research', *Daedalus*, Vol. 107, No. 2 (Spring 1978), 191-209.

Ian Mitroff's latest book is *Methodological Approaches to Social Science* (Jossey-Bass, 1978) coauthored with Ralph H. Kilmann, Graduate School of Business, University of Pittsburgh. *Daryl Chubin* has just completed a book manuscript entitled *Viruses and Cancer: A Social Study of Growth and Specialization in Biomedical Research,* coauthored by Kenneth E. Studer, Department of Sociology and Anthropology, Virginia Commonwealth University. Their most recent publication is 'Knowledge and Structures of Scientific Growth: Measurement of a Cancer Problem Domain', *Scientometrics,* Vol. 1 (January 1979), 171-93.

Authors' addresses (respectively): Interdisciplinary Department of Information Science, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA; Department of Social Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, USA.

RUSTUM ROY:

Alternatives to Review by Peers: A Contribution to the Theory of Scientific Choice

Minerva, 22 (1984) 316-328

IN THE mid-1960s a series of papers by Michael Polanyi, Dr Alvin Weinberg and others opened the discussion of "scientific choice".[1] The term "scientific choice" referred to two rather disparate choices. The first was the choice among subfields of science with respect to allocation of resources. The second was the choice among different possible performers of research as to which should be supported. Since those happy days for science, the former question, possibly the most critical issue of science policy, has virtually disappeared from the range of concerns of the makers of science policy. In 1977, in the wake of President Carter's push for "zero-based budgeting", whereby the allocations made for the next budgetary period would disregard the pattern of allocation of the previous budgetary period, I suggested to Dr Philip Handler, then president of the United States National Academy of Sciences, that the Academy create a committee to reflect on the desirability of a "zero-based budget" for science and engineering. Dr Handler's response was that such a task was impossible since it would arouse bitter conflicts among scientists. Under present practices, the support of American science research is frozen into a more or less fixed pattern of distribution of support for engineering, applied scientific research and pure scientific research, and among different subfields; if the pattern were once appropriate, it appears less and less so every day.[2]

The issues raised in the discussion of scientific choice reappeared in the 1970s in connection with "peer review". Unfortunately, the entire focus of the argument was shifted from the important problem of the best way to distribute the total funds for the support of research, to an analysis of one of the less significant methods of deciding how funds should be allocated. The persons who discussed "scientific choice" seemed to take for granted that allocations would be decided on the basis of review by peers. Some of the criticisms focused on the possible miscarriage of justice to individuals. The debate about the merits and defects of review by peers diverted attention

[1] Polanyi, Michael, "The Republic of Science: Its Political and Economic Theory", *Minerva*, I (Autumn 1962), pp. 54–73; Weinberg, Alvin M., "Criteria for Scientific Choice", *Minerva*, I (Winter 1963), pp. 159–171, and "Criteria for Scientific Choice II: The Two Cultures", *Minerva*, III (Autumn 1964), pp. 3–14; Maddox, John, "Choice and the Scientific Community," *Minerva*, II (Winter 1964), pp. 141–159; Carter, C. F., "The Distribution of Scientific Effort", *Minerva*, I (Winter 1963), pp. 172–190; Toulmin, Stephen, "The Complexity of Scientific Choice: A Stocktaking", *Minerva*, II (Spring 1964), pp. 343–359.

[2] Shapley, Deborah and Roy, Rustum, *Lost at the Frontier: American Science and Technology Policy Adrift* (Philadelphia: ISI Press, 1984).

from the more important issue of what different systems are available for the distribution of research funds.

Defects of the Present System

The many government agencies which use the system of review of proposals by peers, and which have sponsored many attempts to validate their system,[3] have not supported a single study to compare such review with other systems of allocating funds for research by scientists of similar qualifications; there has been no comparison with the "strong-manager" method used by the United States Department of Defense, or with the formula system. No effort has been made to examine the "efficiency" of the system in terms of the costs and time required for each grant, or the efficacy of the system in supporting genuine innovation.[4]

Even without systematically analysed comparative data, the failures of review by peers as a way of deciding which projects and which scientists should receive grants seem to be very evident. Yet virtually no senior official has commented on the glaring deficiencies of these procedures. Let us examine a recent example. In 1983 the Department of Defense started a new programme making available $30,000,000 annually to provide some large items of research equipment to universities. The Department of Defense deviated from the procedures used by many of its own subdivisions, which could presumably have selected for those universities working with the Department the articles of equipment most needed. Instead, in an effort to gain public favour, it issued an invitation for proposals to all universities, whether or not there were research groups with significant support from the Department of Defense in those universities. This resulted in a fiasco. Over 2,200 proposals were received for a sum of $625 million. The success ratio was less than 1:60. The time required for the preparation and submission of each proposal may be estimated at one month's work of one person. Thus, 2,200 scientists spent one month—or nearly 200 years of scientific work were diverted from research. Fortunately, scientific peers were not used to evaluate the process; but any estimate of the total expenditure of time must come to perhaps one more year of one scientist's work for each grant made. If we include an average figure of 100 per cent for overhead costs, the cost therefore was equal to 400 average salaries at $40,000 per scientist per year, and the allocation of $30,000,000 cost about $16,000,000. This example does not include the administrative costs of a typical review by post, or of panels

[3] Cole, Stephen and Cole, Jonathan, "The Ortega Hypothesis: Criterion Analysis suggests that only a Few Scientists contribute to Scientific Progress", *Science*, CLXXVIII (27 October, 1972), pp. 368–374; Cole, S. and Cole, J. R. and Simon, G. A., "Chance and Consensus in Peer Review", *Science*, CCXIV (20 November, 1981), pp. 187–255.

[4] Roy, Rustum, "Peer Review of Proposals: Rationale, Practice and Performance", *Bulletin of Science, Technology and Society*, II, 5 (1982), pp. 405–418; Roy, Rustum, Testimony at Hearings on "NSF Peer Review" before the US House of Representatives Subcommittee on Science and Technology, No. 32 (29 July, 1975), pp. 684–693.

of peers. The entire operation, moreover, further exacerbated the differences between the successful and the unsuccessful applicants, since the unsuccessful universities wasted their efforts, while those with the more successful applicants gained even more.

The failure of the system of review by peers is not that it fails to provide financial support for most good scientists. No system could fail to do that! The unnecessary waste of limited resources of scientific talent is the single most telling failure of the peer-review system.

There is no single system of review by peers. There is an infinity of systems in which the influence of the scientific "peer" varies from almost nothing more than the passive lending of a name for legitimation, to almost complete control. I will consider only four major categories of review by peers: reviews submitted by peers, plus an assessment by an assembled panel, plus a site visit; reviews submitted by post, plus an assembled panel; standardised postal reviews, usually with quantified ratings or preferences within prescribed categories; and postal advice or comments without ratings.

Only the most sanguine advocate unfamiliar with the literature would claim that there is any basis for expecting a correlation between a scientist's ability to present an essay and the actual future production of the "best science".[5] The weak links in a "theoretical" sense are that we have no definition of what constitutes the "best science". With the total confusion between the terms "basic" and "applied" and over the value of relevance, and the very major psychosocial differences in perceptions and values, between—let us say—civil engineers and theoretical physicists, the entire system of review by peers is one of reinforcements of the idiosyncracies or the ruling paradigms of any group which is constituted and supported as a unit.

We have no definition of a peer. According to current practice—strongly departing from the judicial model—a peer is defined as one who works in the same narrow subspeciality of scientific research. For example, proposals in microwave plasma synthesis exclude those working in radio-frequency plasmas, or those working in chemical vapour deposition synthesis. The working definition of peers in the present peer-review system certainly means as narrow a group of specialists as can be found to match the subject defined in the proposal.

But is not a better definition of a peer, a person of equal "rank" and "experience" in science, drawn not only from the narrow speciality, but explicitly including the neighbouring fields? Dr Weinberg's insight that the best science would be that which affected a wide group of fields indicates that peers be explicitly defined to include some from neighbouring fields. Not a single agency of the United States government does this.

In the present system, the simplest precautions against conflict of interest

[5] Peters, D. P. and Ceci, S. J., "Peer Review Practice of Psychological Journals", *The Behavioral and Brain Sciences*, V, 2 (1982), pp. 187–255; Harnad, Stephen, "Peer Commentary on Peer Review"; followed by 56 comments, *ibid.*, pp. 185–186, 196–225.

are ignored. The system flies in the face of the most elementary knowledge of human nature and presupposes a level of objectivity, disinterestedness and honesty, such as never obtained in any human group. Proposals, possibly with the literature thoroughly surveyed, the investigator's best ideas clearly expounded, and experiments specifically laid out, are sent to the set of colleagues who can most adequately evaluate the proposal but who also could use this same information in their own research. Moreover, in the present climate of opinion, a colleague who knows that he or she has the certain power to doom that proposal by a check mark in the "Fair" or "Good" category—even if accompanied, albeit inconsistently, by written praise—might well be inclined to use it. Both applicants and assessors know that this could give the reviewer enough time to perform the same or similar research. Applicants in rapidly developing fields, therefore, often employ the stratagem of applying for funds only for work that is already complete and is nearing submission for publication.

The system of peer review ignores the crucial role of "change" and serendipity in science. Against all historical evidence, the system is based on the idea that genuine discovery can be planned in advance in an essay which meets with the approval of distant colleagues. Obviously, some outstanding work is done by persons who receive grants through peer review of their applications—after all, any system will deliver a good fraction of its funds to the best scientists. This is no vindication of the system.

The defects of the system of review by peers may be summarised as follows: It disregards the multiplicity of systems of assessment and the possibility of combining their best features. It involves an enormous waste of the finite resource of the time of scientists and is inherently unfavourable to innovation. The schedule inherent in the process—often requiring some months to write an elaborate proposal and a waiting period of between six and twelve months—does not correspond with the actual schedule according to which creative scientists work, where the period from gestation of an idea to trying it out is much shorter. The intellectual momentum is thus often lost. The process encourages "competition" instead of co-operation and collaboration as the most effective mode of achieving the best scientific results.

Alternative Systems of Allocating Funds for Research

The only alternatives to the system of review by peers are not lotteries or the granting of equal amounts of financial support to every "qualified" scientist, although such experiments may be worthwhile. Scientists do not seem to wish to consider the more serious alternatives. Scientists who defend the status quo say that the peer-review system is the source of the success of American scientists, as demonstrated by such achievements as the winning of Nobel prizes.[6] Many scientists who are reasonably well sup-

[6] Committee of Scientific Society Presidents, Testimony at Hearings on "NSF Peer Review" before the US House of Representatives, Subcommittee on Science and Technology (July 1975), p. 1,096; also pp. 1,081, 1,088.

ported find it difficult even to consider alternative systems in an objective and open manner. For those entering research after 1950, the system of peer review has become synonymous with support for their own scientific research. Some regard any criticism as a threat to the continued support of their research—hence there has been no scientific inquiry into or discussion about alternative systems. The National Academy of Sciences which has often risen to the defence of the system of peer review has never considered the possible alternatives.

The fact is that scientific research has been supported by an enormous variety of institutional arrangements all over the world; the system of peer review is only one of these. Alternatives do not need to be invented—they already exist in abundance.

In most countries, the existing systems tend to make "block" grants to universities, laboratories or departments. In Great Britain, for example, the University Grants Committee provides grants to run over a number of years for universities; these include sums for the maintenance of departmental research activities such as costs of new equipment and salaries for research. Graduate students are supported again from block grants to departments, from the Science and Engineering Research Council. Finally, the latter body—and other ministries—also makes grants for special research projects on the basis of one-page general proposals. In Japan "laboratory-sized groups" under a senior professor, with some younger teachers and students, are supported for five- to ten-year periods, on the basis of site-visits and review, while students are supported directly by the Ministry of Education. In South Africa, an interesting variant of review by peers focuses exclusively on the most recent research by the individual professor, without requiring any proposal describing the research which the professor intends to carry out in the immediate future.

In the United States itself much more money for research is actually distributed by the "strong-manager" system used by the Department of Defense and other agencies which support "mission-oriented" basic research than by the system of review by peers. In the academic world of the United States, a system for the support of research—review by peers—became firmly established as a result of the involvement of the major universities in military research during the Second World War: with the best of intentions, it turned out to favour those universities. In some disciplines—notably the applied sciences—scientists frequently receive part of the support for their research through the "strong-manager" system and part from agencies which depend on review by peers. Such persons are particularly well placed to compare the systems. Interestingly enough, the fields which are further removed from application, such as theoretical physics and chemistry, radio-astronomy, and many parts of the life sciences, have never experienced any system of support other than the peer reviews employed by the National Institutes of Health and the National Science Foundation. Nevertheless, many leading scientists of the United States are

beginning to be exasperated by the waste of time required by the system of review by peers.

Three alternatives to that system are worthy of consideration. They differ in the extent to which they diverge from the present procedures. All three can be tried simultaneously within a single major grant-awarding body. These alternatives share certain major principles or presuppositions. First, past success is the best basis for the prediction of future performance. Second, the support of small groups or individuals on a continuing basis for the appropriate length of time—let us say, for seven years—increases the probability of success and the efficiency of the system. Third, the most innovative science is done in the context of attaining a broadly defined objective, yet carried out with the minimum of close supervision of the goals, methods or budgetary categories of individual projects. Fourth, co-operation between the very best specialists and the bringing together of their very best experimental capacities is a necessary condition for the advancement of knowledge in fields at the frontier of development. The failure of the universities to make institutional provision for interdisciplinary research is, therefore, seriously damaging to them.[7] Fifth, the advancement of scientific knowledge which goes hand in hand with the advancement of the public interest is the most valuable.

All the systems for the allocation of funds for research must reconcile two apparently contradictory requirements: greater freedom for the investigator and greater accountability, not merely for honesty in expenditures but for the goals of research.

A Formula for Support based on Productivity assessed by Peers

Why should the public support research with no specific mission or goal? Such research carried on at universities has the several functions. It makes new knowledge available to the world at large by publication. It trains students at an advanced level by apprenticeship in research. It establishes a capacity for research—including trained scientists, advanced knowledge and scientific instruments—enabling the university to participate in the more telestic research, i.e., research linked to a purpose or mission and supported by agencies such as the Department of Energy, the Department of Transportation, the Occupational Safety and Health Administration and the Department of Defense. It makes possible research done in support of the private industrial firms which produce a major part of the wealth which goes to the support of research.

In keeping with the view that the best guide to the prediction of success is past achievement, financial support should be proportional to the past productivity of the scientist in these four functions. The grants should not be to individuals but to a group, usually of the size of a department or an

[7] Roy, Rustum, "Interdisciplinary Science on Campus: The Elusive Dream", *Chemical Engineering News*, LV (29 August, 1977), pp. 28–40.

interdisciplinary laboratory. The arrangement would work as follows: a governmental body concerned, for example, with chemistry, would maintain records submitted annually by every institution of higher education conducting research or graduate education or both. The data supplied would include information on the "productivity" of the institution with respect to the four functions, including the number of papers published, mainly in a set of journals agreed by the staff of the department or laboratory. Each member can only be counted as one, although his or her work may be split over two or more units, with no double counting of product permitted. At a later stage, citations might be used as a measure of scientific productivity, but for the first stage, there is little difference between using papers and using citations as a measure of productivity when one aggregates the publications of 20 to 40 persons over three to five years.[8] The university would provide the numbers of advanced degrees granted, in each unit, each year. A simple scheme of weighting different degrees would reflect the amount of effort used to produce them. The measure of the value of the particular scientific capacity of the unit for serving practical or mission-oriented ends is the amount of financial support received from mission-oriented government agencies. The effectiveness in the performance of the fourth is the total financial support for research received from private industry. These data would of course be available for each year; the measures used would be a rolling average of the preceding three or four years.

The combined formula: The actual formula using the data provided would be as follows, where all numbers represent rolling three-year averages:

> Total sum to be granted to unit = A × number of publications + B × weighted number of advanced degrees + C × sum received for research from mission-oriented agencies + D × sum received for research from private industry.

The weighting factors—A, B, C, D—would be adjusted by each agency so that the total of money distributed to all institutions would equal the total budget. The relative values of these factors becomes a flexible device for making policy. The weighting scheme is easily understood by high administrators and legislators. For example, if it were desired to encourage collaboration between university and industry, it would only be necessary for the legislative body to increase the weighting of D; this signal would immediately be perceived and acted upon throughout the country. Likewise, if there were shortages of qualified persons in one area, and a surplus in another, the legislature would simply rule that, for example, factor B would be tripled in the field of shortage and halved in the field in which there was a surplus of trained persons.

[8] Roy, Rustum, Roy, N. R., and Johnson, G. G., "Approximating Total Citation Counts from First Author Counts and from Total Papers". *Scientometrics*, V, 2 (1983), pp. 117–124.

Allocation of academic unit's research allotment to individuals from grants to unit: One of the immediate objections to this scheme by productive scientists is the concern that it will reward the lazy members of the unit. This must be guarded against and it is simple to do so. Many American universities already do what is proposed here in the reallocation of the "return of overhead" to departments and research units. Each unit will be required by the granting body to provide an "acceptable plan" so that individual scientists will benefit from their own productivity. Typically, such a plan might propose that 5 to 10 per cent of the funds will be retained at the college level for major equipment or projects for which any unit or individual can compete. Similarly 10 to 15 per cent of the funds might be retained at the departmental or unit level for commonly used equipment and technicians, etc. The 75 to 85 per cent would then be divided up within the department using the same—or slightly modified—formula so that the most "productive" members of the department would receive the greater part of this genuinely "unearmarked" support.

Summary: This "formula" is, of course, very different from the many other such schemes proposed, in that it takes account of "productivity" in the field in which support is given, and in that it is flexible. Representative George Miller of California, chairman of the House of Representatives Subcommittee on Research, introduced his "formula funding" bill in 1968; his formula did not involve any measures of productivity at all.[9] In contrast, my formula draws on the most intensive and defensible review by peers and hence should immediately command the support of all those who believe in assessment by scientific peers. Review by peers is present through the fact that the vast majority of the publications occur in journals in which the referees are scientific peers. A published paper has undergone peer review which has assessed the quality of completed research. Although peer review of completed research has itself been attacked,[10] it is less vulnerable to criticism than the system of review by peers of proposed research projects. In any case, counting all publications instead of publications reviewed by peers in technical fields would make hardly any difference in the distribution over the country as a whole. The award of advanced degrees, except in some master's degrees, has been subject almost always to collective judgement by the department or the committee of examiners. A successful application for grants for research from "strong managers" in a mission-oriented agency is also conducted after a much more stringent review by highly qualified peers than any postal review of a proposal. Finally, by far the most stringent review by a peer is that which occurs when a research-manager in an industrial laboratory allocates $50,000 to support research at a university.

The formula which I propose here can deal with the situation of the new, usually young member of the unit in several ways. Any unit which appoints

[9] Representative George P. Miller, author of HR 35, Testimony at Hearings before US House of Representatives, 91st Session of Congress (February 1969).

[10] Peters, D. P. and Ceci, S. J., *op. cit.*; Harnad, S., *op. cit.*

new members should receive one extra share for each new member equal to the average allocation for research for each member of the unit. If the newly appointed person is a replacement, there is presumably already enough money to pay the salary and the expenses of the research. This averaging scheme in my view is less perturbing to the pattern of any particular university than a fixed award for young scientists. An award of $50,000 per year for a new assistant professor who has recently been awarded the doctorate might not be unreasonable at the major universities, but it could cause jealousy at some others.

A comparison of the scheme I propose with the existing system of awards on the basis of review by peers shows that the proposed scheme rests on a clear principle. It is a proportional award by the representatives of society for delivering the "products" those representatives desire. Moreover, it is based on recorded, quantifiable performance, not on promises made in an essay. The "best science" is defined as that which honours both scholarship and the public interest, and is "quantified" in the four terms of the formula. There are, moreover, different sets of peers for different evaluations. In most of these, the particular group of peers is much more broadly cognisant of other fields of the national interest, or industry's wants, than in the prevailing system of peer review. In addition, conflict of interest is virtually eliminated. Whereas in the system of peer review, a single negative review by one of the peers can block the entire research of a senior scientist with less than six months' notice, no single person can affect the immediate future of any individual drastically.

This elimination of conflict of interest goes far to remove the corrosion of the integrity of the community caused by the present system. It goes much further than the commendable though mild effort by the National Institutes of Health in the warning statement which its reviewers are asked to note, but not even to sign and return. Because the recipient is not constrained by any proposal, or in any fashion in the National Science Foundation, the National Institutes of Health or among his colleagues, he can follow any unforeseen lead or any unexpected opening. He can chose to do the most risky experiments in his genuinely atelistic work, i.e., work without a particular goal in view, by balancing it with work for a mission-oriented agency and industry. Dr Weinberg's proposal made 20 years ago that atelestic research should be an "overhead" on mission-oriented research will have been realised.[11]

The magnitude of an individual's research grants will vary gradually over time, guaranteeing every scientist a certain proportion of the stable allocation of funds to do some of the difficult work requiring five or more years, which is now shunned for the quick production of papers. The beneficial feature of the formula is that it provides no sharp line of demarcation with respect to quality of individuals or universities. Thus it

corresponds much more accurately to the realities of the distribution of talents. This schema has the merit of rewarding productivity of the same "commodities" wherever they are produced. A wider range of universities will get some money on a steady basis, even if the total sum is not large.

It is not by any means obvious that this funding scheme will by itself increase co-operation. It will, on the other hand, certainly diminish dishonesty, which as Dr Yalow has pointed out may be as harmful as anything else to the scientific community. It will make co-operation a little more likely.[12]

A Simpler Formula for Matching Support

A simpler version of the formula proposed above has been advocated in the various schemes proposed by others for "matching" funds. Thus, Dr Weinberg's idea would provide basic research funds to a unit or an individual in direct proportion to the mission-oriented research done by that unit or individual; in the formula which I propose it would consist only of the third term. Similarly, the report of the commission on industrial innovation, under the chairmanship of Jordan Baruch and appointed by President Carter, proposed that universities should receive five to seven times the amount provided by industry.[13] In effect, this proposal is represented by the fourth term of my formula. I believe a case can be made for using Dr Weinberg's rationale and reducing the formulae to a matching grant proportional to C times total support by mission-oriented agencies plus D times support by industry. This would, however, limit the general applicability of the formula since certain less applied fields would be at a severe disadvantage.

Support Based on Peer-Evaluation of Performer's Past Achievements

The system based on performance can be applied to individuals. In this variant, support would be provided on the basis of the entire, fairly recent achievement in research of the individual. This is the system now being used by the Council for Scientific and Industrial Research of South Africa.

The peers who are selected to evaluate the scientific achievement of any individual must rigorously exclude any possible conflict of interest. In the case of American academic scientists, this would exclude other American academic scientists. Industry and government scientists who cannot possibly be supported from the same source provide a much larger pool of candidates

[12] Yalow, Rosalyn S., "Is Subterfuge Consistent with Good Science", *Bulletin of Science, Technology and Society*, II, 5 (1982), pp. 401–404.

[13] Baruch, Jordan, *Final Report of the Advisory Committee on Industrial Innovation, United States Department of Commerce* (Washington, DC: US Government Printing Office, September 1979).

who are less likely to be biased. The reviewing group should be drawn by random-number selection by computer from a qualified set selected for each subdiscipline, and should include one or two from related disciplines. The qualified sets of reviewers would be revised annually by a committee of the National Research Council. In general, it might be valuable to have small groups of individuals in roughly the same field all evaluated by the same group; this would provide a basis for comparative judgements. The best group of peers is obvious: it is the group of scientists in the same general fields in the many similar institutions in foreign countries.

The data presented to the assessors would be a curriculum vitae, up-to-date bibliography, and perhaps a two-page statement by the individual being assessed summarising his or her recent achievements in research. Again, this kind of judgement is hardly alien to academic traditions. It is made daily in appointive decisions in universities and industry. The peers would rank their evaluation of the quality, quantity and the originality of the research and its value to society. The actual budget allocations could be left to the managers since this would be dependent on the subfield of study.

International Evaluation by Peers of Established Groups

This is a variant of the scheme for individuals applied to departments or other research units. It is in widespread use now in many parts of the world. In its ideal form, unfortunately, it cannot cope with large numbers of institutions. For example, to allocate funds among half a dozen bioengineering groups, each group would prepare a statement of recent achievements and provide the data on its productivity over the preceding five years. The international group of assessors would both evaluate each group's "productivity", compare the productivity of the groups in the set, and point out possible overlaps, gaps, etc. Ideally, the peer group would visit each of the units, assess its capacity "on the ground" and provide a much more precise recommendation regarding the relative merits of each unit working on the subject to the grant-awarding body. Whenever such assessments have been used, they have been regarded as eminently fair.

Clearly it is impractical to have the same group of assessors visit 100 or even 25 institutions. A postal review alone could perhaps be done for 20 or 25 by the same set of assessors, but several sets of assessors may be needed when large sets of institutions must be evaluated. Visits to sites by an international peer group are clearly the best way to evaluate the smaller sets of specialised research institutes, such as water research institutes and materials research laboratories. It is again a mark of the conservatism of the management of American research institutions that this has never been done. It is a pity, since the international nature of the reviewers would very properly reflect the realities of modern science.

An Optimal Review of Proposals by Peers

Since the smaller the change the more likely its adoption, I include for completeness an alternative to the present system, modifying it in simple ways which would eliminate some of the major objections to the present system and which would not be more costly.

From the outset, it should be made clear that the system seeks to evaluate the proposer and the ideas, and that each will be weighted equally or that greater emphasis will be laid on the former. Hence, the proposal should present in detail the recent research achievements of the programme and provide a short general indication of the work to be attempted. The review form must explicitly explain the two different evaluations being requested.

Two simple sets of precautions will avoid conflict of interest by selection of assessors from industry and government, as I have already recommended, for all academic applicants. One can very easily avoid manipulation of the selection of referees by the manager of the programme by using a randomised selection from a qualified set of referees designated by the National Research Council or a similar body. A simple legislative enactment would solve this problem once and for all. For judgements regarding the allocation of public funds, the proportion of peers drawn from industry, governmental and academic laboratories should reflect roughly the national percentages of working scientists in the particular discipline throughout the country.

The instructions to assessors should be very clear—as they are not today. They should be designed by experts in the preparation of questionnaires. The instructions should include the probabilities of the outcome associated with each level of rating. Thus the assessor should be informed that, say, an average rating of B− results only in a 5 per cent probability of support, A− in 75 per cent and an A+ in 95 per cent probability of support. Similarly the exact weighting to be placed on the applicant's achievements in research and on his proposed research should be indicated unambiguously.

Once the assessments are received, they should be sent to the applicant for technical rebuttal, to be completed within exactly two weeks and with only one cycle of such rebuttal permitted. This avoids absurd errors in the referees' judgements.[14] The Dutch government uses this system.

It is clear that we have absolutely no guarantee that a "principal investigator" will actually do the research he has proposed and whether it will be "successful", and if so, whether it will be of any value to science. Hence, it is absurd to have a sharp disjunction between successful and unsuccessful applicants in the distribution of funds. The budgets should be

[14] Recently one of our proposals was classified as border-line and turned down by the National Science Foundation, although one of the reviewers' main reasons for a lower rating was that in an interdisciplinary proposal we had not included Professor X., the best known person at Pennsylvania State University. The reviewer might have had a point had Professor X. not been dead for over two years. The National Science Foundation, while embarrassed, had no way to correct the error because the process by then had moved to the next stage.

reallocated to diffuse the boundary by taking the bottom 15 to 20 per cent of the successful applicants and cutting their budgets by, let us say, 25 per cent and awarding partial support—from 33 to 50 per cent of their requested sum—to 10 to 15 per cent of those just below the present dividing line. Getting something started, or maintaining something already under way, is often invaluable to morale and very economical.

In the United States where different major agencies which support research use different systems, it would be a worthwhile experiment to employ both the schemes within the same agency in different divisions. After some years, the parts of the scientific community which have had experience of both schemes could be surveyed for their preference. Meanwhile, studies of the preferences of scientists who have already been supported by both systems might be undertaken immediately to justify the experiment.

Conclusion

The questions of scientific choice which were left unresolved when the rapid expansion of academic science in the United States began in the early 1960s have come back to trouble the scientific community. There is now widespread dissatisfaction with the process of review by peers as one of the major systems for the allocation of public funds for research. While earlier criticisms had been brushed off by the assertion—unsupported by facts—that no other systems existed, the present situation cannot be so easily dismissed.

A serious examination of other national and international arrangements shows that a wide variety of procedures are in use and there is no research which shows that one system is either more productive scientifically, or more cost-effective in bringing about valuable scientific research. New systems which may be considered should avoid the major defects of the system of peer review as now practised: the enormous waste of scientists' time, the great potential for conflicts of interest, and the inherent bias against innovation.

The principal system which I have proposed here combines the best elements of peer review with the simplicity and efficiency of the use of a formula. Moreover, this formula based on peer review of performance incorporates all the elements for which the academic scientific establishment should be accountable to its patron, which is the public treasury. A final virtue of the proposed system is that it provides simple and convenient procedures through the use of numerical weighting factors for the policy-maker to guide the support of scientific research as a whole.

ALAN L. PORTER AND FREDERICK A. ROSSINI:

Peer Review of Interdisciplinary Research Proposals

Science, Technology & Human Values, 10 (1985) 33-38

The peer review of research results submitted for journal publication raises elementary issues of fairness and reliability.[1] Peer review of proposals to perform research in the future, however, is even more problematic. For many reasons, judging untested ideas is inherently more uncertain than evaluating completed work. Research has a way of evolving in directions that are unchartable in advance—certain data may prove unattainable, new discoveries by the researchers or by others may point to reorientation, or personnel may change. In a 1974 study, Grace Carter found that ratings of National Institutes of Health (NIH) initial grant applications were correlated with independent ratings of their later reapplications by a thin 0.4 (i.e., only 16% of the variance was accounted for by the other rating).[2] While this figure includes unreliability due to an independent rerating, it also reflects changes in the perception of the value of specific projects as research progresses. One National Science Foundation (NSF) proposal reviewer made special note of "exemplary staffing plans" of a proposal he evaluated.[3] Ironically, that same project changed staff repeatedly to the extent that it was not possible even to identify a project leader. In essence, then, peer review of proposals is a difficult business.

Peer review as a process has engendered strong charges and defenses.[4] Rustum Roy, for example, argues that peer review of proposals has no conceptual basis, wastes resources, and impedes innovative research.[5] Setting aside such total objections to the process, Harvey Brooks makes a well-reasoned case that peer review is better in some respects than in others, and for some tasks than for others.[6] Peer review of proposals is better for evaluating within defined fields than across fields; better for collecting expert opinion on what Brooks calls the "truth" dimension (the pursuit of knowledge for its own sake) than differentiating along a "utility" dimension (research to be applied toward a specific end). According to this view, it follows that peer review is less satisfactory for applied or policy research than for basic research. Furthermore, "the broader the intellectual territory covered, the less consensus there will be on the ranking."[7]

These attributes of peer review point to potential difficulty in its use to evaluate crossdisciplinary research proposals. Such proposals involve multiple skills focused on a scientific research problem. Thus, these proposals present substantial difficulties in identifying an appropriate "peer" group. It is likely to be difficult to identify peers whose expertise fully encompasses the proposed crossdisciplinary research. If located, they are apt to have a strong personal stake in the outcome of the evaluation in the event that the number of researchers concentrating in the area is small. If such a peer group cannot be gathered, review by persons not fully familiar with the domain can prove especially perilous. Recognizing such issues, program managers may hesitate to undertake review of proposals that lack an established peer group. Crossdisciplinary proposals may truly "fall between the cracks" of the disciplinary programs.[8]

Professor Porter is in the School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332; Professor Rossini is in the School of Social Sciences, Georgia Institute of Technology, Atlanta, GA 30332.

* *This research was supported by the National Science Foundation, Office of Interdisciplinary Research, Grant OIR–8209893. The views expressed in this paper are those of the authors alone.*

One of the most striking observations about peer review of proposals is the extent to which this process is unquestioningly accepted in the absence of much empirical data on how it performs. Cole et al., in a landmark study, contrasted NSF peer ratings with independent sets of raters.[9] They found that the fate of individual grant applications was about half determined by characteristics of the proposal, about half by "noise" in the review process. We know of no empirical data, however, on such critical issues as the effect of reviewer heterogeneity[10] or the characteristics of the proposed research on expected ratings. The intent of our article, therefore, is to provide such evidence on one sample of crossdisciplinary projects.

Study Design

This study draws on 257 sanitized reviews of 38 projects from five different National Science Foundation programs: Neurobiology, Environmental Geosciences, Archeology, Earthquake Hazard Mitigation—Societal Response, and Science and Technology to Aid the Handicapped. These encompass basic, applied, and policy research. The sampling was purposive rather than statistically representative or random. We worked with program managers to identify projects they perceived as interdisciplinary (and added a few problem-oriented disciplinary projects for comparison).

Roy lists four dimensions basic to considering peer review of proposals[11]:

(1) The candidate set of proposals—(a) in response to an RFP versus (b) unsolicited, with a deadline, versus (c) no "set" at all (a program considers pertinent proposals as they come in);

(2) Reviewers—how many and who selects them;

(3) The review medium—mail (advisory only), mail (binding), panel, site visit, or combinations; and

(4) The degree of interaction with the principal investigator (PI), ranging from none to substantial.

In these terms, the 38 NSF projects for which we had reviews cross all candidate set types. For some, peer evaluation incorporates explicit comparison among a set of proposals; for others, a proposal is considered alone. The evaluations range in number of reviewers from 1 to 17 (with no specific information on their selection); the mean number of reviews per proposal varied from 5.9 to 8.1 for the five programs. These reviews incorporate mail or mail/panel combinations. In a few cases, the review process includes feedback of initial criticisms through the program managers for principal investigator response.

Complementing the proposal reviews, we were able to secure information from the PIs on the nature of the actual research process.[12] In particular, we draw on this to determine how interdisciplinary each project was, based on a weighted function of the PIs' and our own judgments, on the number of disciplines represented on the project team, on the percent of staff from outside the PI's general disciplinary category (i.e., engineering, life sciences, physical sciences, social sciences, or professional fields), and on the range of skills used in the project.

Results

The proposals studied were all funded and thus had high ratings on the NSF scale of one to five (with one being excellent and five being poor). Also, because the projects in the sample were selected for their interdisciplinary character, we originally believed that there would be few differences in the sample in terms of such characteristics as type of project or specialty of reviewer. As Table 1 indicates, we found some rather significant differences, however. These can be summarized by noting that reviewers favored basic scientific research conducted in an academic setting.

The first line of Table 1 condenses information on how ratings differed among the five NSF programs under study. Each of the two Engineering Programs (Earthquake Hazard Mitigation and Science and Technology to Aid the Handicapped) averaged a rating of 2.05. Among the sciences, the single archaeology proposal for which we had numerical ratings scored most favorable at 1.14, followed by the neurobiology proposals at 1.42, and geoscience at 1.68. Engineering proposals received more reviews (mean number of reviews of each = 7.5) than did the science proposals (mean = 6.0).

In a multiple regression of peer rating on the various factors examined, the program variable appeared as the strongest predictor ($p < .001$), dominating other factors, including which proposal was being rated. In other words, more of the variability in rating was accounted for by the program

Table 1. Peer Rating Contrasts

Dimension	More Favorable Type	Mean	(±S.D.)	(N)	Less Favorable Type	Mean	(±S.D.)	(N)	Statistical Significance Level
NSF Program	Scientific	1.50	(±0.69)	(102)	Engineering	2.05	(±0.96)	(149)	0.0001
PI Affiliation	Academic	1.65	(±0.80)	(165)	Non-Academic	2.18	(±0.98)	(86)	0.0001
PI Disciplinary Category	Scientific	1.75	(±0.86)	(189)	Engineering	2.07	(±0.98)	(62)	0.02
Reviewer Disciplinary Category	Scientific	1.75	(±0.87)	(128)	Engineering	2.12	(±0.97)	(70)	0.01
Project Type	Basic Research	1.56	(±0.43)	(17)*	Applied or Policy Research	2.00	(±0.50)	(21)*	0.01
Disciplinary Match Between Reviewer and PI	Same or Similar	1.69	(±0.83)	(199)	Different	2.22	(±0.99)	(69)	0.0001

Note: Scale is the NSF rating from 1 = excellent to 5 = poor.
Scale is naturally compressed as all of these proposals were funded.
* Number of projects on a project-based comparison instead of a review-based one as in the other four contrasts.

to which the proposal happened to be submitted than by perceived differences in merit among individual (funded) proposals. Stated another way, the funding cutoff varies significantly by program. The person submitting a crossdisciplinary proposal might do well to compare the typical peer rating profile for each program that might evaluate the proposals.

Switching to stepwise regression, we considered whether other variables augment the program variable as a predictor of a proposal's peer rating. Two other variables enter the regression significantly—whether or not the PI is academic and the similarity of PI and reviewer d..ciplines. Each of these is discussed below, with academic PI being the strongest predictor. Wh . both of these variables are entered into the equation, program is no longer a significant predictor. The intercorrelation among the predictors is also discussed further.

The second line of Table 1 summarizes information on ratings by organizational affiliation of the Principal Investigator. Given the peer ratings profile that favors basic scientific research conducted in academic settings, we would have suspected that proposals from PIs nested in established disciplines would have rated better than those from centers. Forty-three percent of the ratings pertained to PIs associated with academic departments. These rated quite favorably (mean = 1.73). Interestingly, proposals from academic centers or composite center/department arrangements fared at least as well (mean = 1.50). To simplify further comparison, we combined all academic proposals in Table 1. Four other groups of proposals (from public or quasi-public organizations involved in funding, or in using research; from large or small contract research organizations) each averaged over 2.0. We conclude that either academic PIs know how to prepare better proposals, or reviewers favor academics, or both.

Similar patterns emerge when we consider the disciplinary category of the PI or of the reviewer. "Science" rates better than "engineering." By type of project, basic research (mean rating 1.56 for 17 projects) does better than applied (mean rating 1.98 for 13 projects) or policy research (mean 2.04 for 8 projects). Applied and policy research are combined as a category in Table 1. Differences by reviewer affiliation (i.e., academic or not; researcher or research user) did not reach statistical significance. Nor did review characteristics interact significantly with corresponding PI characteristics. One might have expected academic reviewers to be more favorable toward academic PIs, and non-academic reviewers to favor non-academic PIs. The review data did not show such inclinations.

Not surprisingly, the factors discussed are heavily interrelated. Examination of correlations among these variables finds all of them significant, ranging in magnitude from 0.22 to 0.74. "Program," in particular, associates with whether the PI (r =

0.41) and reviewer (r = 0.42) are academic (i.e., the engineering programs are more likely to engage non-academics); and with whether the PI (r = 0.48) and reviewer (r = 0.52) are scientists or engineers. The key finding is that projects that are "scientific, basic research in academic units" are the highest rated, while those that are "engineering, applied or policy research in non-academic units" are rated lowest. Projects with other combinations of characteristics are rated between those extremes. The key determining factor seems to be whether or not the project has an academic PI. However, in this study the connections demonstrated among the variables influencing peer rating are correlational, not causal. Our study also does not address the issue of whether some proposals are "better" than others in a sense independent of the ratings. Stated another way, someone could argue that the proposals by the academic PIs are rated more favorably than those from non-academics because they are inherently superior, not because of any favoritism.

The case with respect to disciplinary match between PI and reviewer is more sharply defined. As shown on the last line of Table 1, reviewers favor that which is familiar. We considered information on reviewer and PI affiliation to judge whether their disciplines were the same (e.g., both chemists), similar (e.g., physiologist and anatomist), or different (e.g., engineer and sociologist). This line collapses categories of "same" and "similar" (means of 1.68 and 1.71). To confirm this finding, we examined the similarity of general disciplinary category (e.g., social science, engineering) of PI and reviewer using a non-judgmental coding. Everyone was assigned a three-digit code for discipline and these codes were located in general disciplinary categories, (following the National Research Council's groupings used in the study of doctoral scientists and engineers). Where reviewer and PI were affiliated with the same general disciplinary category, peer ratings were better (mean = 1.73); where they differed, peer ratings were significantly worse (mean = 2.08; p = 0.008).

The implications of this favoritism for the familiar are critical for interdisciplinary research (IDR). First, this effect is extremely strong for it to appear in a sample of a constricted range of ratings. Second, this effect will work against support for IDR in that such research is inherently less likely to be reviewed by persons familiar with the full scope of the planned work. Our conclusion should caution against the generally accepted strategy of choosing review teams, each of which is familiar with only one of the aspects of the project. These results imply that such a review strategy will generate poorer ratings, on average, than would come from a review team on which each member was familiar with the whole scope of the proposed research.

Our final inquiry was whether there was a correlation between project interdisciplinarity (based on a factor analysis as noted under "Study Design") and rating. This query is constrained by the sampling of predominantly IDR projects. We did observe a correlation between rating and degree of project interdisciplinarity in the direction of more interdisciplinary projects being downgraded. In a split-sample analysis, this conclusion shows a significant association for one set of 20 projects (r = 0.50, p = 0.02), but non-significant association for the other subsample (r = 0.16, p = 0.26). However, the correlation is significant for the sample as a whole (r = 0.29, p = 0.05). This partial support for the hypothesis that IDR projects tend to rate less favorably suggests further study of a representative sample of both disciplinary and interdisciplinary proposals (preferably including non-funded as well as supported proposals).

Discussion

Researchers engaged in certain areas, such as the neurosciences, appear to meet less resistance from their peers (proposal reviewers and professional reward evaluators) for performing interdisciplinary research than do those in more traditional disciplinary areas. Some research areas appear more "open" to using techniques, and even substantive expertise, not historically wedded to them. Of the comments which we saw, reviewers rarely criticized crossdisciplinary features of proposals. (One economist reviewer did fault a proposal for including non-economic aspects.) More typically, suggestions were made to add a particular skill to the project team. On occasion, reviewers would indicate reluctance to evaluate the proposal other than in their own domain of expertise. Our general sense was that neuroscientists and archeologists, in particular, did not need to justify their inclusion of "outside" skills. The characteristics of the research problems seemed to require certain skills, and the researchers tried to provide them. On the

other hand, in earlier work, we encountered academic departments that presented serious obstacles to their members trying to include skill areas beyond their discipline's expertise in their research.[13]

If these findings generalize across NSF, then some of the fears expressed formally and informally by members of the research community have merit. Brooks reports that peer evaluations in areas other than basic research are much less satisfactory.[14] Roy cites peer review as "fundamentally and ineluctably anti-innovation."[15] Grover Whitehurst notes variation across fields in implementation of peer review.[16] Present data suggest that peer review of NSF proposals favors research that is performed by academics, in the sciences, and that falls completely within the reviewer's own domain of expertise. Well-established research areas are thus favored over nascent ones. Compensatory mechanisms to counterbalance these inclinations may be warranted.

Our most intriguing finding offers a clue as to why interdisciplinary proposals are downgraded. It is reasonable for a reviewer of proposed research to favor that which is more familiar personally. In such a case, one is apt to understand better what is planned; one may know the researchers personally or by reputation, and hence appreciate their expertise; and one can feel more secure in making strong recommendations. One program manager summarized his experience as demonstrating that reviewers often considered proposed work from the "standpoint of only their discipline." The marked tendency of reviewers to rate proposals from PIs from the reviewer's own discipline more favorably suggests that IDR should not be reviewed the same way as disciplinary projects.

In view of this serious concern, consideration should be given to alternative strategies for reviewing IDR proposals. One proposed approach of composing a review team that includes a reviewer who knows one aspect of the proposed research, another who is expert in a second area, and so on, implies a set of reviewers each unfamiliar with much of the work.[17] According to our present findings, that means an expected downgrading of the proposal in comparison to one of equivalent merit, but for which individual reviewers can better grasp the full scope of the research involved.

To give interdisciplinary research proposals the same opportunity as disciplinary proposals for a good rating, one has several choices. One option is to lower rating standards, but that is difficult to justify within programs. NSF's current policy places applied research projects in basic research programs. When basic and applied research proposals fall within one program, peer reviews are likely to favor the basic science proposals. Our results further suggest that interdisciplinary research proposals within such a program will be at a competitive disadvantage. In essence, using peer review to choose among different types or areas of research is ill-advised.

Program managers, on occasion, collaborate to consider funding a proposed project that is too large or broad in scope to be supported by a single NSF program. When this happens, sometimes "turf" issues arise. The obvious solution to such issues—cross-program review—entails additional jeopardy for the proposer (e.g., even more reviewers expert in only part of the proposal, or multiple panel reviews).

Another possibility is to seek reviewers who are expert across the breadth of the proposed work. As Martha Russell notes, "To the extent that reviewers have a holistic perspective of knowledge creation and use, they can offer their best guesses as to the success of the proposed [interdisciplinary] research and thereby assist in screening projects which are likely to be productive."[18] Such reviewers may be difficult to secure for "frontier" research, and there is danger of inbreeding in cases where only a small number of potential reviewers share interests fully.

Panel meetings offer an advantage in allowing discussion to help each of the members understand unfamiliar aspects of a proposal under review. The remarks of some panelist reviewers support such a strategy as they report "changing their mind after discussion." However, panels are costly and must be constituted to address sets of proposals. Novel interdisciplinary research is likely to fall on the fringe of panel expertise and hence to face a poorer expected rating than more mainstream, disciplinary research, with all other conditions being equal.

We suggest that incorporation of feedback in the mail review process could help remedy the problem of restricted reviewer expertise. Roy reports that no substantial feedback mechanism is formalized in any U.S. granting agency.[19] We suggest extending the provision that is sometimes used of providing the PI an opportunity to respond to the concerns of reviewers before a final funding decision is reached. Explicitly, we advocate a

"Delphi" process.[20] Mail reviewers would comment on a given proposal but not provide a numerical rating. These comments would be provided anonymously to the PI and to the other reviewers. The PI and reviewers would be given an opportunity to explain the proposed research more satisfactorily. Only after receiving these comments and amplifications would reviewers make a numerical rating. Such a process should improve reviewers' perspectives on the full project and resolve concerns arising from unfamiliarity with certain aspects.

This study offers empirical evidence that reviewers of research proposals lean toward certain types of research. It documents a tendency to favor that which emanates from one's own discipline. We need to devise ways to avoid discriminating against crossdisciplinary proposals that lack an established peer group. Perhaps the funding agencies should consider the development of new mechanisms other than traditional peer review to support at least some interdisciplinary research. Russell, for example, describes a structure of research advisory and management committees to balance disciplinary and interdisciplinary research activities in the agricultural experiment station context.[21] Institutional grants and formula allotments certainly have problems, but they could be constructed so as to nourish interdisciplinary research. Combinations of peer review and other elements may offer advantages.

Crossdisciplinary and, especially, interdisciplinary research are vital for scientific and technological innovation. For such research to succeed, it must survive the peer review process.

Notes

1. Stevan Harnad, ed., "Peer Commentary on Peer Review" (special issue), *The Behavioral and Brain Sciences*, Volume 5, Number 2 (1982); Grover J. Whitehurst, "Interrater Agreement for Journal Manuscript Reviews," *American Psychologist*, Volume 39 (1984): 22–28.
2. Grace M. Carter, *Peer Review, Citations, and Biomedical Research Policy: NIH Grants to Medical School Faculty* (R–1583–HEW). (Santa Monica, CA: The Rand Corporation, 1974).
3. Alan L. Porter, Frederick A. Rossini, and Daryl E. Chubin, *Interdisciplinary Research (Problem-focussed, Multi-skilled Research)—National Science Foundation Experiences* (Atlanta, GA: Georgia Institute of Technology, 1984).
4. See Gilbert W. Gillespie, Jr., Daryl E. Chubin, and George M. Kurzon, "Experience with NIH Peer Review: Researchers' Cynicism and Desire for Change," *STHV*, Volume 10, Issue 3.
5. Rustum Roy, "Peer Review of Proposals—Rationale, Practice and Performance," *Bulletin of Science, Technology and Society*, Volume 2 (1982): 405–422; also see the article by Rustum Roy, *STHV* Volume 10, Issue 3.
6. Harvey Brooks, "The Problems of Research Priorities," *Daedalus*, Volume 107, Number 2, Spring 1978: 171–190.
7. *Ibid.*
8. Halsey Royden, "'Risky' Investments," *Science*, Volume 209 (11 July 1980): 216; Alan L. Porter, Frederick A. Rossini, Daryl E. Chubin, and Terry Connolly, "Between Disciplines," *Science*, Volume 209 (29 August 1980): 966.
9. Stephen Cole, Jonathan R. Cole, and Gary A. Simon, "Chance and Consensus in Peer Review," *Science*, Volume 214 (20 November 1981): 881–886.
10. Grace M. Carter, *What We Know and Do Not Know About the NIH Peer Review System*, N–1878–RC/NIH (Santa Monica, CA: The Rand Corporation, 1982).
11. Roy, *op. cit.* (1982).
12. Porter, Rossini, and Chubin, *op. cit.*
13. Frederick A. Rossini, Alan L. Porter, Patrick Kelly, and Daryl E. Chubin, "Interdisciplinary Integration within Technology Assessments," *Knowledge*, Volume 2 (1981): 503–528.
14. Brooks, *op. cit.*
15. Roy, *op. cit.*
16. Whitehurst, *op. cit.*
17. *c.f.* Carter, *op. cit.* (1982).
18. Martha G. Russell, "Peer Review in Interdisciplinary Research: Flexibility and Responsiveness," in *Managing Interdisciplinary Research*, Sidney R. Epton, Roy L. Payne, and Alan W. Pearson, eds., (Chichester, England: John Wiley & Sons, 1983), pp. 184–202.
19. Roy, *op. cit.*
20. *c.f.* Harold A. Linstone, and Murray Turoff, eds., *The Delphi Method: Techniques and Applications* (Reading, MA: Addison-Wesley, 1975).
21. Russell, *op. cit.*

ANGELO S. DENISI, W. ALAN RANDOLPH AND ALLYN G. BLENCOE:
Potential Problems with Peer Ratings
Academy of Management Journal, 26 (1983) 457-464

Peer ratings can be defined as the set of evaluations obtained by having each member rate every other member of a work group, using a specific set of rating scales. A considerable amount of research has been conducted on peer ratings and peer evaluations in general (Brief, 1980; Kane & Lawler, 1978, 1980). However, virtually no research has been reported on how people react on learning they have been either poorly or favorably evaluated by their peers. What effects do these ratings, or other types of peer evaluations (especially when they are negative) have on subsequent interactions, feelings, performance, and any future evaluations?

There is a theoretical basis for anticipating problems following an individual's receiving negative evaluations from peers. Balance, or consistency theories (Osgood & Tannenbaum, 1955; Rosenberg, Hovland, McGuire, Abelson, & Brehm, 1960), for example, would suggest that learning someone has evaluated an employee more poorly than he/she would evaluate himself/herself will lead to "source derogation" (Tannenbaum, 1978), or the lowering of the employee's opinion of the rater and any subsequent evaluations of that rater. The notion of reciprocity in social exchange (Adams, 1965) also suggests that one would "repay" a peer for a poor rating by later giving poor ratings to that peer. Finally, consistency theories would lead one to expect other outcomes following source derogation. For example, there is reason to believe that learning of poor peer evaluations will result in those peers becoming less interpersonally attractive (Kiesler & De Salvo, 1976), which tends to reduce cohesiveness among group members

[1]The authors wish to thank Dean McIntosh for extensive comments on an earlier version of this paper.

(Zajonc, 1962). This reduction in cohesiveness can have serious consequences for the performance of interacting groups (Stogdill, 1972).

There also are empirical data to indicate that people retaliate against peers after receiving negative ratings. Koeck and Guthrie (1975), for example, reported that subjects lowered subsequent personality ratings of peers by giving them more negative ratings after learning those peers had rated them negatively. They did not, however, raise subsequent ratings after learning that peers rated them positively. Bernardin (1980) found similar results of retaliation for poor supervisory evaluations, citing a relationship between the ratings a supervisor gives a subordinate and the subordinate's description of the supervisor's leadership style. This potential for retaliation probably is more serious in the case of peer ratings, however, because the ratee may well have the opportunity to repay the rater in kind.

Thus, there is some reason to expect that learning of negative peer ratings will lead to retaliation during subsequent evaluations, lower group cohesiveness, and perhaps will even cause poorer performance for interacting groups. Nonetheless, no research is known that has directly examined any of these possibilities. It also should be noted that people may react to peer ratings in ways much different from those discussed above. Negative ratings, especially if one feels them to be unjustified, could be viewed as an attack on self-esteem. A number of studies reviewed by Korman (1970) suggested that persons might react to such an attack by working even harder to demonstrate their competence. This is, after all, what would be hoped for as a result of negative evaluations. Conversely, it is possible that positive peer evaluations could produce too much cohesiveness, which, if the group was not particularly motivated to perform well, could result in less task oriented interaction and consequently poorer performance (Stogdill, 1972). These outcomes must be acknowledged as possibilities, although they seem less likely, given the research results reviewed above.

The present study is a laboratory experiment that investigated effects of peer evaluations on group behavior and performance. The following general hypothesis is tested:

> *Knowledge of peer ratings will affect ratings of group cohesiveness, satisfaction, group interactions, and group performance (both perceived and actual) on a subsequent task.*

It thus is predicted that there will be positive effects for individuals who learn that their peers have rated them positively (i.e., various rating "scores" will improve) and negative effects for individuals who learn that their peers have evaluated them negatively. Furthermore, following the results reported by Koeck and Guthrie (1975), it is predicted that the effects following negative peer ratings will be relatively stronger than the effects following positive peer ratings.

Method

Subjects and Procedures

A total of 143 undergraduate students (68 female) participated in a laboratory experiment in partial fulfillment of course requirements. All subjects

worked in one of 34 small groups of 3 to 5, depending on the number reporting for a particular session. Preliminary analyses indicated no effects for subject sex, the sex composition of the groups, or any sex × condition interactions. In addition, there were no effects for group size. Thus all results are presented for the total sample.

On reporting, subjects were given a general description of the procedures to be followed and were told that they would be working as a group on two tasks. Following each task they would be asked to complete a questionnaire that included an evaluation of every other member of their group by name. They also were informed that each person would be shown the evaluations, but only the mean rating received, so that individuals could not tell how any one person rated them. (Of course, in smaller groups, a group member receiving negative feedback would know that no peer could have rated him or her as an outstanding performer.) Subjects also were informed that two persons would be observing their group from behind a one-way glass, but the observers' only role was to record group interactions. After questions were answered, subjects were asked to sign informed consent forms. There was no penalty for refusal to participate. Alternatives were available for students to satisfy this portion of their course requirement.

The two group problems were variations of a truck routing task developed by DeNisi and Pritchard (1978), which required subjects to map a cross-country route to maximize, within certain constraints, the value of the cargo trucked. The two variations were pretested to insure equivalence. This particular task was appropriate for the present study because several distinct pieces of information had to be considered, and up to five people each had to assume a real role within the group. The task also required coordination among group members so that members had continuous interaction. It therefore was impossible for any one person to perform the task alone, and it was difficult for any one member to be disassociated from the outcome, which clearly is a group product. Finally, based on extensive pretesting of the original task, subjects had no feel for the number of points (measuring the value of the cargo) that were needed to represent good or poor scores. Thus, there was little potential for task-generated feedback that could interfere with the manipulated peer feedback (discussed below). Peer rating feedback was given after subjects completed all measures following the first task, and all peer feedback was false. No feedback was given following the second task. Instead, after completing the questionnaires and the peer ratings forms a second time, subjects were debriefed and dismissed.

Measures

1. Task and Socioemotional Behavior. Bales' (1950) interpersonal process analysis (IPA) form was used for assessing group interaction during each task. Two independent observers, blind to feedback conditions, rated each group on all 12 of Bales' categories for each task, recording evaluations

on 7-point scales for each category. Individual categories were combined, according to Bales' suggestion, to form two general dimensions of interaction—task and socioemotional behavior. Ratings for these general dimensions were computed by taking the average of each observer's ratings of all 12 categories for both tasks. These were taken to be the total set of observations (24) for that rater. This was done so that any reliability coefficients computed would be based on a reasonable number of observations. A separate reliability coefficient then was computed for each group by correlating the 24 ratings made by the two observers assigned to that group. The range of reliability coefficients across the 34 groups was between .81 and .95, and the average was .92.

2. *Satisfaction.* Group members were asked to assess their satisfaction with the participation and the contribution of the other members, the group solution, and overall satisfaction with group members. Each item was rated on a 7-point scale with higher ratings indicating greater satisfaction, and the responses to the four items were averaged to form a single measure of satisfaction. The internal consistency of this measure (coefficient alpha) was computed to be .84 and .87 for the two administrations, respectively.

3. *Group Cohesiveness.* Subjects completed a 4-item cohesiveness scale similar to that used by Terborg, Castore, and DeNinno (1976). Each item was rated on a 7-point scale with the items averaged to form a single measure of cohesiveness. The internal consistency of the scale (coefficient alpha) was .85 for both administrations.

4. *Perceived Performance.* Subjects rated their perception of the group's performance on a single 7-point scale with higher ratings indicating more effective performance.

5. *Peer Ratings.* Subjects rated the overall task performance of each of the other group members, by name, using a single 7-point rating scale (1 = very low; 7 = very high). The ratings given by the subjects to peers were averaged separately for each task, and the average ratings given by a subject served as a dependent measure.

Ratings of group interaction were obtained from the two independent observers, but all other measures came from the subjects themselves. *Objective task performance* was the actual dollar value of the cargo collected by the group on its route, expressed in points.

Conditions

Groups were randomly assigned to either positive (17 groups, $n = 70$) or negative (17 groups, $n = 73$) peer rating feedback conditions. After subjects completed peer ratings for the first task, an experimenter collected them and left the room, telling the subjects that their ratings would be averaged and returned to them. While the experimenter was absent, subjects were told to complete the other parts of the questionnaire (i.e., the satisfaction, cohesiveness, and perceived performance items) and were instructed *not* to discuss the peer ratings or the questionnaires. They were told that

this was to insure that each subject recorded only his or her impressions in the questionnaires. Observers were asked to watch and listen for any such discussions and to remind subjects of the instructions (using microphones from behind the one-way glass) if anyone did begin to discuss the questionnaires or ratings. No one discussed either with fellow group members.

The experimenter returned to the room and gave each subject the false average peer rating (for overall performance) in a sealed envelope. Subjects were instructed to read but not discuss the feedback and were quickly given instructions for the second task and told to begin work. The sealed envelopes did not contain the subject's true average peer ratings. Instead, all persons in positive feedback groups were informed that they had received an average rating of about 6.0 (7.0 was the highest rating). All those in negative feedback groups were informed that they had received an average rating of about 2.5 (1.0 was the lowest rating).

No feedback was given following the second task, although subjects did complete the peer ratings and the other questionnaire measures. In addition, subjects were asked to recall the average peer rating that they had received following the first task, as a form of manipulation check. All were able to recall their average rating within .1 of a point. Discussions with subjects during debriefing indicated that they did believe the feedback received.

Results

It was hypothesized that knowledge of peer ratings would affect group member interactions, perceptions, performance, and subsequent peer ratings and that the nature of these effects would depend on the sign of the peer ratings. A series of analyses therefore was conducted considering the sign of the peer rating feedback (either positive or negative) and time (after task 1, before feedback; or after task 2, following feedback) as independent variables, and the following dependent variables: task and socioemotional behavior exhibited (rated by observers), cohesiveness, satisfaction, perceived performance, actual performance (total points collected), and average peer rating given. Table 1 presents the means (and standard deviations) for all dependent variables, broken down by time and sign of peer rating feedback.

Because the various dependent variables were correlated, a multivariate analysis of covariance (MANCOVA) was conducted. It revealed a multivariate effect for the sign of peer feedback ($F = 4.53$, $p < .01$). A series of univariate analyses of covariance (ANCOVA) was then conducted on time 2 ratings with time 1 ratings as covariates, following Huck and McLean (1975), after an initial test indicated that common slope could be assumed ($F < 1$). The ANCOVA results indicated significant ($p < .05$, or better) peer feedback effects on perceived performance ($F = 7.32$, $\omega^2 = .11$), cohesiveness ($F = 6.24$, $\omega^2 = .10$), satisfaction ($F = 10.65$, $\omega^2 = .16$), average peer rating given ($F = 16.50$, $\omega^2 = .20$), and rated task behavior ($F = 5.62$,

Table 1
Means and Standard Deviations for All Variables at Time 1 and Time 2 for Positive and Negative Feedback Groups and the Results of Several Comparisons

	Negative Feedback			*Positive Feedback*		
	Time 1	*Time 2*	*Time 1-Time 2*[a]	*Time 1*	*Time 2*	*Time 1-Time 2*[a]
Observer measures						
Socioemotional behavior	2.84 (1.08)[b]	2.56 (.56)	<1	2.75 (.86)	2.75 (.50)	<1
Task behavior	4.76 (.72)	4.52 (.50)	<1	5.20 (.75)	5.28 (.68)	<1
Perceptual measures						
Perceived performance	6.21 (.89)	5.64 (.81)	5.28**	6.24 (1.46)	6.74 (.82)	<1
Cohesiveness	5.39 (.92)	4.95 (.42)	4.18*	5.70 (.79)	5.95 (.77)	1.02
Satisfaction	5.63 (1.02)	5.09 (.94)	3.85	5.77 (1.04)	5.81 (.88)	<1
Average peer rating given	5.33 (1.40)	4.42 (1.09)	6.08**	5.64 (1.13)	6.20 (.81)	7.42**
Actual performance	39.00 (3.62)	37.50 (4.94)	2.81	40.56 (5.24)	40.94 (3.97)	<1

[a]These are the results of simple effects analyses comparing ratings at time 1 and time 2.
[b]Standard deviations are shown in parentheses.
*$p < .05$
**$p < .01$

$\omega^2 = .09$). Thus, knowledge of peer ratings did affect most of the dependent variables of interest, although rated socioemotional behavior and objective performance did not seem to be affected.

To test the exact nature of these effects over time, for subjects learning of positive and negative peer ratings, a series of simple effects analyses, comparing time 1 and time 2 means within feedback groups, was conducted. Winer (1975) suggests that this type of analysis is more appropriate than simple *t*-tests, given the evidence of sign of feedback × time interactions, and the tests were conducted using formulae provided in his book. As predicted, subjects learning of positive peer ratings raised scores on all perceptual measures from task 1 to task 2, collected more points on the second task, and were rated higher on task behavior for task 2. However, only one change was significant (average peer rating given: $F = 7.42$, $p < .01$). Also as predicted, subjects learning of negative peer ratings lowered scores on all perceptual measures from task 1 to task 2, collected fewer points on the second task, and were rated lower on task and socioemotional behavior for task 2. These changes were significant for perceived performance ($F = 5.28$, $p < .05$,) cohesiveness ($F = 4.18$, $p < .05$), and average peer rating given ($F = 6.08$, $p < .01$). Thus, these results support the predicted stronger effect of learning about negative peer ratings.

Discussion

The results from this study seem to raise some doubts about the use of peer ratings for feedback, but the limitations of the present study are

recognized. Perhaps foremost among these is the short time perspective studied. It is suspected that performance following negative peer ratings would deteriorate over time, but there are several other possibilities. Negative ratings conceivably could motivate a worker to try even harder in order to improve the ratings received; thus performance might actually improve over time. It also is possible that, over time, group members might become accustomed to poor peer ratings and ignore them completely. A different limitation of this study arising from the short time perspective is that in existing work groups, cohesiveness and satisfaction generally have had time to develop more fully than in this study, and workers might be less susceptible to changes following negative peer ratings. One could argue, however, that learning that close and trusted peers have given negative ratings might have an even greater debilitating effect on the group. In any event, given more time together, the apparent deterioration in group-member relations might well lead to a decline in actual performance. There is a clear need for further research with a more realistic time horizon.

Other limitations to the external validity of the present study stem from the manipulation of the peer feedback. Negative peer ratings were operationalized here as a mean rating of about 2.5 on a 7-point scale, but rarely would one peer rate another so poorly. Thus, it is possible that the relatively strong effects found for negative peer feedback were due to the extreme nature of the feedback received. Nonetheless, such a reaction would indicate that subjects did believe that such low ratings were possible. Of course, in existing groups, negative ratings probably are defined more by group norms than by scale points, and what seems to be a much more positive rating could actually be viewed as quite severe by group members.

In general, even given these limitations, the results of the present study suggest the need for further research on peer ratings and peer evaluations in general. This research needs to go beyond demonstrating significant relationships between peer evaluations and some criterion measures. This study, performed in a controlled setting, found that knowledge of how one's peers have rated a person had a definite impact on group behavior (especially when those ratings were negative). This finding cannot be ignored simply because the setting was somewhat artificial. Instead, field research needs to be conducted utilizing designs that allow assessment of effects over longer periods of time so as to understand how group interactions and patterns of behavior might be affected by feedback from peers.

References

Adams, J. S. Injustice in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2). New York: Academic Press, 1965, 267-300.

Bales, R. F. *Interaction process analysis: A method for the study of small groups.* Cambridge, Mass.: Addison-Wesley, 1950.

Bernardin, H. J. The effect of reciprocal leniency on the relationship between consideration scores from the LBDQ and performance ratings. *Proceedings of the 40th Annual Meetings of the Academy of Management*, Detroit, 1980, 131-136.

Brief, A. P. Peer assessment revisited: A brief comment on Kane and Lawler. *Psychological Bulletin*, 1980, 88, 78-79.

DeNisi, A. S., & Pritchard, R. D. Implicit theories as artifacts in survey research: An extension and replication. *Organizational Behavior and Human Performance*, 1978, 21, 358-366.

Huck, S. W., & McLean, R. A. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 1975, 82, 511-518.

Kane, J. S., & Lawler, E. E. Methods of peer assessment. *Psychological Bulletin*, 1978, 85, 555-586.

Kane, J. S., & Lawler, E. E. In defense of peer assessment: A rebuttal to Brief's critique. *Psychological Bulletin*, 1980, 88, 80-81.

Kiesler, C. A., & DeSalvo, J. The group as an influencing agent in a forced compliance paradigm. *Journal of Experimental Social Psychology*, 1976, 3, 160-171.

Koeck, R., & Guthrie, G. M. Reciprocity in impression formation. *Journal of Social Psychology*, 1975, 54, 31-41.

Korman, A. K. Toward a hypothesis of work behavior. *Journal of Applied Psychology*, 1970, 54, 31-41.

Osgood, C. E., & Tannenbaum, P. H. The principle of congruity in predicting attitude change. *Psychological Review*, 1955, 62, 42-55.

Rosenberg, M. J., Hovland, C. I., McGuire, W. J., Abelson, R. P., & Brehm, J. W. An analysis of cognitive balancing. In C. J. Hovland & M. J. Rosenberg (Eds.), *Attitude organization and change*. New Haven, Conn.: Yale University Press, 1960, 112-163.

Stogdill, R. M. Group productivity, drive, and cohesiveness. *Organizational Behavior and Human Performance*, 1972, 8, 26-43.

Tannenbaum, P. H. The congruity principle revisited: Studies in the reduction, induction, and generalization of persuasion. In L. Berkowitz (Ed.), *Cognitive theories in social psychology*. New York: Academic Press, 1978, 225-252.

Terborg, J. R., Castore, C., & DeNinno, J. A. A longitudinal field investigation of the impact of group composition on group performance and cohesion. *Journal of Personality and Social Psychology*, 1976, 34, 782-790.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1975.

Zajonc, R. B. The effects of feedback and probability of success on individual and group performance. *Human Relations*, 1962, 15, 149-161.

Angelo S. DeNisi is Associate Professor of Management and Organization Behavior, College of Business Administration at the University of South Carolina.

W. Alan Randolph is Associate Professor of Management and Organization Behavior, College of Business Administration at the University of South Carolina.

Allyn G. Blencoe is a Ph.D. Candidate, College of Business Administration at the University of South Carolina.

MARTIN RUDERFER:

The Fallacy of Peer Review:
Judgement without Science and a Case History

Speculations in Science and Technology, 3 (1980) 533-562

Peer review, the process of judging the contributions that make up the archives of science, is not now justifiable as a scientific endeavour. Yet this process, via the archives of science, is a key factor in determining man's ability to cope with the growing global problems of the population explosion stimulated by past scientific progress. There exists an urgency to improve peer review in order to guarantee the technological growth rate vital for long-term survival. However, a science of peer review has thus far been precluded by the secrecy imposed on the primary raw data — review histories. To begin to rectify this, a case history of an erroneous rejection is presented in detail. The rejected paper, which claimed to correct a published dispute involving atomic timekeeping, was published in SST in 1979 along with a follow-up paper confirming and extending it. The case history leads to the hypothesis that the probability of rejection increases with the degree of innovation in a publishable work. This is validated by the follow-up paper which shows the rejected paper to require a paradigm shift to correct a widespread misinterpretation of rotating clock behaviour now erroneously attributed to special relativity. This results in a simple unification of rotating clock behaviour in atomic timekeeping, the Sagnac effect and the Hafele-Keating experiment. The ability of this case history to clearly delineate origins of human error in review processes demonstrates the need for publication of many more and the desirability of stressing erroneous rejection in peer review at least as much as the traditional emphasis on erroneous acceptance. This one case also supports the urgency required for improving the accuracy and speed of peer review and recommends a number of specific means for accomplishing this.

1. INTRODUCTION

The spectacular growth of modern science following the invention of printing attests to the supremacy of the printed word in man's pursuit of knowledge. The archives of science nurtured the rapid rise of technology in the last few centuries.

Demands on reports for research journals — the roots of these archives — are among the most exacting in all the fields of literature: extensive preliminary research, precise explication, maximum objectivity and rigour, absence of bias and error, novel or utilitarian content, all presented with optimum economical clarity. It is consequently not surprising that the decision of what to accept or reject has always been difficult.

For the last three centuries the principal method of judging potential contributions has been peer review, commonly known as the referee system, wherein acceptance or rejection is decided by an editor based on reports of

anonymous referees of ostensible expertise. Yet, although author reports are required to approach the peak of scientific methodology, referee reports which judge them have had no such explicit restraint — maximum secrecy, unprescribed rigour, lack of precise judgement standards, cursory investigation and no guaranteed impartiality. Credibility of referee reports rests mainly in the belief, often just a hope, that the referee is truly a peer for the material being judged and that such peer knowledge has been adequately applied.[1] No matter what justification is cited for peer review, an ironic truth remains: The scientific method is not being rigorously applied in the *process* of selecting those examples of the scientific method deemed worthy of preservation.

If this seems harsh, consider the effect of a referee error that results in a firm rejection. The only recourse for the author is to submit again elsewhere. The chances of subsequent acceptance are best if his thesis is evolutionary, for the error is then more likely to be localised, i.e. peculiar to the referee. If his thesis is revolutionary, the error is probably more widespread, i.e. peculiar to established doctrine, and the probability of publication within his lifetime is minimal. Although uncommon, revolutionary ideas have played a prominent role in the development of science. There is no specified error-correction mechanism in peer review for insuring that any valid development, evolutionary or revolutionary, is not disregarded for an excessive period or is forever overlooked.

Rejection errors have far reaching adverse consequences for the research community. They result in needless additional time and effort caused by re-submissions, increase the total referee load on the journals, reduce the time spent by authors and referees on primary projects and, especially, have disastrous psychological effects on authors. Mayer and Boltzmann, for example, were so depressed by rejection it was a contributing factor in their attempted suicides.[2] Boltzmann succeeded but Mayer only broke his legs, was confined to a mental institution for a while and was finally recognised for his work on energy conservation after a loss to society of about 15 years. At the other extreme is a recent case known to the author of complete abandonment of research because of disillusionment with the review system.[3] Between these extremes lies a spectrum of psychological effects which rob society of the full potential of many of its most creative members. Measurement of the total cost of review errors has been ignored for too long.

Gross inequities from rejection errors are inherent in the open-loop nature of the present review system. This is contrary to the existence of closed-loop negative feedback systems in almost every facet of society for correcting major errors, viz. separation of executive and judicial functions; law enforcement; elections; elaborate court systems for redress; arbitration; open refereeing, as in sports contests; ombudsmen, forums; etc. The lack of any prescribed error-correcting mechanism built into the referee system brands it as one of the most autocratic in society today.

An author may spend months, years and even a lifetime in preparing a manuscript, yet under the present system it may be arbitrarily rejected by a referee after a cursory reading, akin to the flick of the wrist used by some Nazis in deciding acceptance or rejection for survival. Arbitrary action is not precluded by the present system; in science what is not expressly impossible

should be deemed feasible. How often it has actually occurred is immaterial; what is significant is that there is no sure recourse to authors to correct a defective rejection or, in the case of deliberate malpractice[1,4,5], any form of redress.

The absence of negative feedback is further aggravated by the high positive feedback inherent in the dissemination of scientific ideas. When published research in a new area stimulates further interest, this leads to further publications, still further interest, etc. Although such a snowball effect is useful in rapidly exploiting breakthroughs, the net long-term effect of such a happenstance type of growth has never been fully evaluated, especially in regard to the consolidation of tacit unrigorous assumptions that may unknowingly usurp other, more superior, courses of development. Moreover, it is also well known that the overall effect of positive feedback in any system is rapid uncontrolled growth and long-term instability. The incredible short term gains man has achieved in the relatively brief life of modern science are now beginning to spin off numerous long-term problems, especially from the population explosion it has stimulated. The crucial role of the referee system in delaying the response time of the science establishment to meet the challenges to society induced by past technological successes has already been noted.[6] The most damning indictment of the present referee system is that the haphazard growth it has fostered has thus far provided no capacity to reliably ensure a stable final state toward which science, and hence civilization, is headed. *In short, there is no positive proof or assurance that modern science cannot self-destruct.* It is primarily for this reason that it is timely to place the present error-prone open-loop referee system under the microscope.

2. QUANTITATIVE EVALUATION OF PEER REVIEW

The application of the scientific method — reasoning based on experience — to peer review has always been hampered by the lack of adequate raw data. Every established science is based on measurement of some observable object or phenomena in nature. In a science of peer review, the relevant phenomena are the details of the review process itself, specifically the total communications involved — submitted reports, review reports and all other interchange of information affecting the final decision. However the secrecy heretofore imposed on reviews has precluded dissemination of such data for quantitative evaluation. As a result, prior discussions, criticisms and studies of peer review and suggestions for improving it[1,3-13], although often incisive, are anecdotal, subjective and/or limited.

To rectify this a sufficient number of case histories must be fully exposed to view so that the variegated factors contributing to review decisions can first be systematically analysed and measured. It is only then that ways to improve the review process can be realistically instituted with confidence. With this objective in mind, such a case history is presented below. However it is first useful to explore the kinds of information required from such data.

Despite the scattered published criticisms of peer review, the prevailing view is that it works. But to what precision? In an outstanding quantitative study from analysis of the records of 14,512 manuscripts submitted to *The Physical Review* from 1948 to 1956, Zuckerman and Merton[7,8] concluded,

"The referee system here apparently does what it is supposed to do: Sift out the good papers from the bad." In other words, for this "low-rejection journal" the system worked somewhat better than chance. Reassuring as this may be, it is not unexpected. The success of modern science implies that its review system must have thus far been statistically effective to some significant degree. But if the actual rejection error rate on publication decisions in this sample is expressed as 1 in 10^n, what is the precise value of n?

The upper limit to n is log 14,512 = 4.2, but its actual value is not possible to ascertain from such a contemporary study. Some revolutionary ideas have been widely rejected *after* publication for 25 and 50 years, as for Maxwell's electromagnetic theory and Wegener's continental drift theory, respectively, so the possible rejection of just one revolutionary *unpublished* advance in the sample cited cannot be reliably determined for at least a comparable period.

The Zuckerman and Merton analysis provides a lower limit to n of the order of 1 in accord with a 20 (5) percent rejection rate for single (multiple) authors.[7,8] *This and the upper limit of 4.2 may be far too low to insure that key ideas are not passed by for excessive periods.*

Premature rejections belie the vaunted exhaustiveness of science. How many other key innovations essential to our long-term survival, as those of Carnot, Gibbs, Goddard, Mendel and others[6], have long been ignored for the wrong reasons? What is the distribution of number of such rejections versus delay in acceptance? The history of innovation indicates that it must be quite skewed with a tail that may asymptotically approach zero. This is in contradiction with the prevalent view stated, e.g. by Cole and Cole[14] that if a "scientist who makes a discovery had not made it, it would have been only a matter of time — probably a relatively short period — before the discovery would be made by another scientist". This popular notion is actually unverifiable and therefore untenable as a working hypothesis for a study of the review process. There is no evidence that n is ideally infinite over the still uncertain life span of modern science; rather, the life span must be generally assumed to be interrelated with n. Its measure is a vital one for exploring means to insure long-term survival.

The current pressure on technology to solve the problems wrought by the population explosion in this century also demands increasing attention to time factors in the review process. Population growth has heretofore been governed primarily by immutable biophysiological factors; growth of knowledge is governed by unrelated and little understood psycho-social factors. The usual answer to the Malthusian doomsday predictions is the *expectation* of new advances by the "technological optimists".[15] But if the disorganising effects to society of the population explosion reach the stage wherein they physically impair the ability to solve technical problems, a worldwide catastrophic situation develops. The current state of civilization suggests that such a process may have already begun and that the possibility of eventual self-destruction cannot be positively eliminated.

It is consequently fail-safe to insure that the technological growth rate exceeds the growth rate of the disorganising effects of the population explosion. To guarantee the required technological solutions we need to minimise unnecessary delays in the growth of knowledge. What are the causes

of delays in the dissemination process in quantitative terms? To what extent can these be eliminated without a sacrifice in judgement precision?

The psycho-social factors affecting the review process and its consequences for authors also demand measurement. Can good and bad reviewers be predicted in advance by some suitable criteria? If so, would a cadre of professional reviewers, comparable to judges in the legal process, be useful in some way? Just how do authors react to the frustration from bad reviews and the powerlessness to deal with them? To what extent does this affect total scientific output?

It is only by exposing the review process to public view that such questions may begin to be answered properly and a science of peer review established. It is in this vein that the following review history is presented in detail.

3. CASE HISTORY OF A REVIEW REJECTION

Although contemporary cases of improper rejection have been discussed in the literature[4,16] the secrecy imposed on the review process has precluded full disclosure. Since the journals have refrained from publishing complete review histories, the only available source is from authors. Hence it is not unexpected that the following case history derives from a personal experience.

The relevant paper, entitled "One-Way Doppler Effects in Atomic Timekeeping", was submitted to *Science* in February 1976 and was finally rejected in April 1977. In July 1978 it was sent to *Speculations in Science and Technology* (SST) as a contrary example to public statements stimulated by its inception that established journals eventually publish all relevant ideas. The paper was published[17] as it was when finally rejected by *Science*. It is accompanied by a short introductory history[18] and a follow-up study[19] confirming and extending the original paper. All page numbers in parentheses which follow refer to these SST papers. Also see "Errata" to these.[20]

The total information transfer between author and *Science* editor are included in the Appendices A through S at the end of this report and are identified in Table 1 preceding the appendices. The dates of receipt of manuscripts by *Science* were stamped on the original and were properly acknowledged.

The rejected paper was stimulated by a prior paper in *Science*[21] by Cannon and Jensen entitled "Terrestrial Timekeeping and General Relativity — A Discovery". Their discovery consisted of a dramatic equalisation in the rates of six coordinated worldwide atomic clocks from application of a terrestrial clock-velocity correction. Subsequent criticism[22,23] and the failure to affirm their discovery with uncoordinated clocks caused Cannon and Jensen[24] to retract their explanation of the equalisation and to designate it as an artifact of the data. The rejected paper claimed to offer an alternative explanation of the Cannon-Jensen finding which was consistent with prevalent theory but was applicable *only* to coordinated clocks (pp.401-2), extended the effect to the solar frame and clarified three unexplained effects reported by Sadeh and associates.

4. DISCUSSION OF REFEREE REPORTS

There were nine reviews of the paper (Q,S). The two referee reports (C,D) received by the author with the first rejection and the two (M,N) with the third rejection are now discussed. Reports of the other five reviews are not available for analysis except for verbal reports that one rated the paper as "excellent" (F,G) and that another was generally negative (I).

Report C. This referee made two fundamental errors. In his first paragraph he referred to a published criticism[23] relevant *only* to the frequency discrepancy of the RGO clock at Greenwich Observatory which is specifically excluded from the analysis, as detailed in E, and which is irrelevant to the Cannon-Jensen finding based on the other six clocks. In the second paragraph, the "synchronization procedures" refer to the lack of absolute calibration of the six clocks to the international second. This affects only the Cannon-Jensen theory which requires absolute accuracy and not the finding itself which is an experimental result and hence exists apart from any theory. The adjustments to the atomic clocks by the time laboratories to slave them to the master UTC clock maintained by the Bureau International de l'Heure is exactly what is required to test for the proposed explanation, as discussed (p.402) and elaborated in E. This rejection is a graphic confirmation of the lack of scientific rigour in the judgement process.

Report D. In his paragraph (a) the referee alludes to "imprecision and inconsistency" without proper justification. He erred in stating that the product of frequency and time in equations (6) (p.389) indicates that frequency is a constant. Constancy of phase, i.e., $fT = f_iT_i$, where f is frequency and T is one-way travel time between transmitter and receiver, is demanded by the Lorentz transformations, as cited just below equation (8) (p.390). These only require $f/f_i = T/T_i$. The frequency change in equation (11) (p.390) is the well-known (classical) change due to phase modulation for a relatively moving observer, as stated. It is evaluated for an observation of a one-way propagation over a given duration T which therefore merely serves as a boundary condition. Because the referee may have been confused by the brevity of the original derivation it was expanded as noted in E.

The paragraph added to the paper (p.400) to quantitatively evaluate the referee's implication in (b) that the neglect of Sun's gravitational potential falsifies the paper showed this cause for rejection to be unfounded.

In (c) the error of referee C is repeated for the same reason the Cannon-Jensen theory was rejected — inadequate absolute clock accuracy — despite the discussion disclaiming the relevance of absolute accuracy in the rejected paper (pp.401-2). The discussion in E and in the follow-up paper (p.406) may be further clarified as follows: A systematic difference between two clock rates, as from any physical cause, results in an ever-increasing difference in clock readings which must eventually become sensible. However, random drift of clock rates over a long period results in essentially no difference in clock readings. Thus, if the magnitude of a randomly varying rate initially exceeds a systematic rate difference, the latter must still become observable over a sufficiently long period. Randomness of clock rate is maximised by the practice of the time laboratories to compare and adjust the individual UTC_i clocks to the average (UTC) of a large number of free-running clocks. Any long term

systematic differences between the six clock rates must then eventually assert themselves by difference in clock readings and/or the rate corrections they entail. Absolute clock accuracy is not directly involved.

Report M. This referee appears to be the same as referee D. His refusal to discuss the issues despite the reply (E) to his initial objections and the revisions initiated by them, or to provide any attempt at further falsification, is an example of arbitrary rejection.

Report N. This rejection has two main errors: (i) The offered explanation is nowhere based on a one-way anisotropy in the speed of light, as claimed by the referee. Isotropy is assumed to be constant throughout, as discussed (p.389). (ii) His formal theory is incorrect and hence inappropriate, as detailed in P, due to his confusion of one-way *travel* time T as a general time coordinate. This is affirmed by his statement "T = t is the coordinate time in the frame of the stationary clock." The meaning of t is made clear at equation (1) (p.388) early in the paper, contrary to his statement further below, "t is not made clear until p.12" (p.394), at which point he could have ascertained his error. Also see P. The last comment is a misconstrued version of objection (b) of D which was evaluated and inserted in the manuscript, as discussed in P.

5. GENERAL COMMENTS

The immediate consideration, which precedes any final judgement, is *the rejections for the wrong reasons.* The four reviews of the three referees plus the one satisfactory response yield a minimum error rate of 3/4 for the review history. This is intolerably high for a presumed scientific process involving technical matters which are inherently resolvable. Such a large deviation from the above-chance accuracy of the average review process in physics[7,8] indicates that there is something radically wrong with the way manuscripts are now judged. We can of course arbitrarily assign errors in any isolated instance to indifference, laxity, chance, prejudice, politics, author status or even malpractice[1,4,5], but this would involve only errors of similar arbitrariness. The availability of case histories presents the opportunity to seek the root causes of erroneous rejection.

In this case the most prominent element common to the rejections is the pre-occupation with theory. This is undoubtedly due to the unconventional form of the Doppler effect applied since a travel time formalism is not the common text-book explanation. Nevertheless, the travel time form applied, equation (13) (p.391), is justified for constant radial and transverse velocities and is more rigorously affirmed to the required precision in the follow-up paper (p.417). The necessity for an unconventional formalism is obvious in retrospect: (i) The use of a one-way coordination signal automatically prescribes one-way propagation theory. However, this little known area has still not entered the main stream of physics. (ii) The constancy required by conventional theory in the rates of (uncoordinated) worldwide atomic clocks in the geocentric frame independent of Earth's rotation (p.385) obscures the possibility that other effects of clock rotation may exist. These combined with the supposed published resolution of the Cannon-Jensen finding largely account for the high error rate in this case history vis-a-vis the average. This preoccupation with conformance to conventional theory thereby suggests the

following working hypothesis, termed here the Innovation Theorem: *The probable delay in acceptance of an innovation increases with its departure from the conventional norm.*

The history of innovation in science, technology and other areas, as politics and religion, generally support such an hypothesis. The familiar, widespread hesitancy to adopt radical ideas inexorably points to a deeply ingrained property of the average human mind. Resistance to innovation in science has been sporadically noted but was never systematically considered until 1961 by Barber[25]. Applied to the review process it suggests that rejection error rate must increase for manuscripts which are rated on an increasing scale from evolutionary to revolutionary.

The enormous success of the scientific method is primarily due to its ability to negate resistance to change by the test of experience for resolving the conflicts precipitated by dogma, tradition, preconceptions and the like. Only one conclusive experience has often been sufficient to overcome the ingrained resistance to a new approach, e.g. the 1919 eclipse observations led rapidly to the wide acceptance of Einstein's theory; Hertz's experiment quickly overcame the deep 25-year rejection of Maxwell's theory[26]; and the discovery of the mid-Atlantic ridge soon settled the theoretical objection to Wegener's continental displacement mechanism despite the other supporting evidence he had collected[27]. In other cases, preoccupation with theory has delayed consideration of well-founded experiment, e.g. Ohm[2,25].

The referees similarly neglected the observational aspects that may have modified their preoccupation with theory: (i) The equalisation of clock rates was too significant to summarily ignore (pp.396-7). (ii) The clarification of the one-way effects reported by Sadeh et al[28] — a diurnal variation with a superposed sunrise effect and a variation with clock separation — were based only on conventional theory. (iii) The extension to the solar frame for which a laboratory experiment was proposed (p.401) was based on the verified Sagnac effect. This same preoccupation with theory manifested itself in the five SST reviews of the follow-up paper, but the weight of added evidence tempered recommendation of outright rejection.

6. CONFIRMATION OF THE REJECTED PAPER

In the hindsight of the delayed acceptance of innovation, as for Maxwell, Wegener and Ohm, it is difficult to comprehend why their innovations were not at first provisionally accepted instead of being firmly rejected. The follow-up[19] of the rejected paper supports the working hypothesis that initial rejection of valid innovation relates to its departure from accepted ideas.

The present theory of relatively rotating clocks stems from Einstein's 1905 introduction of special relativity. This is widely *assumed* to explain the Hafele-Keating experiment comparing two relatively rotating clocks. Atomic timekeeping and the long known Sagnac effect also involve rotating clocks. (Although the Sagnac effect has been expressly confirmed only with oppositely directed light rays in a rotating mirror system, Ives rigorously showed[29] that the substitution of rotating clocks for mirrors and light rays gives identical results.) The difference between the three techniques is trivial: differential rotation derives from fixed atomic clocks at different latitudes in atomic

timekeeping, from oppositely rotating clocks in the Sagnac experiment and from differential rotation of atomic clocks at the same latitude in the Hafele-Keating experiment. The difference in clock rates then all derive, to second order, from equation (25) (p.399) which has its relativistic origin in equation (2) (p.388):

$$\dot{t}_i - \dot{t}_j = (\phi_i - \phi_j)/c^2 - (v_i^2 - v_j^2)/2c^2 + \dot{t}_{uv} \qquad (1)$$

where i,j are rotating clocks, $\dot{t}$ is clock rate, ϕ is the gravitational potential, c is the speed of light, $v_{i,j}$ is geocentric rotational velocity and u is Earth's orbital velocity. The $v_{i,j}$ (uv) term is the kinematic effect in the geocentric (heliocentric) frame. For oppositely rotating clocks with ground speed v at Earth's surface corresponding to oppositely directed light rays in the usual Sagnac experiment, $v_i = \Omega R - v$ and $v_j = \Omega R + v$, where Ω is angular velocity of Earth and R is distance of clocks to Earth's axis. Then the difference in clock readings due to the geocentric kinematic effect (i.e. neglecting the ϕ and uv terms) after one revolution in time $\Delta t = 2\pi/\Omega$ is, since $\dot{t}_{i,j} \approx \Delta t_{i,j}/\Delta t$,

$$(\dot{t}_i - \dot{t}_j)\Delta t \approx \Delta t_i - \Delta t_j = 4\pi Rv/c^2 \qquad (2)$$

This is a common expression for the first-order Sagnac effect.[29] In the Hafele-Keating experiment, clock j is flown around the world at average ground speed v and clock i is stationary on Earth. Hence $v_i = \Omega R$ and $v_j = \Omega R + v$. Then equation (1) similarly yields, after one revolution of clock i in time $\Delta t_i = \tau_o = 2\pi/\Omega$ at which time clock j registers time $\Delta t_j = \tau$

$$\tau - \tau_o = -(2\Omega Rv + v^2)\tau_o/2c^2 \qquad (3)$$

This is exactly the kinematic relation tested by Hafele and Keating.[30] Moreover, they noted (last paragraph) that a more precise evaluation should yield effects of Moon and Sun beyond their measurement precision. These include gravitational and kinematic effects, the solar contribution to the latter being identifiable with the last (uv) term in equation (1). This is small but is evaluated and listed for reference in the rejected paper as "sidereal corrections" (Table 1, p.395). Over the six year period of the follow-up analysis these became measurable by the relative mean drift of the clocks (Fig.5, p.410) and by the long-term drift of all the clocks combined due to interaction with the geocentric effect (Fig.4, p.409). Furthermore, these are inherently verifiable directly by the proposed double-disc Sagnac experiment (p.401).

Although this simple unification of rotating clock behaviour is obvious in retrospect, the Hafele-Keating experiment was not brought up by the author (except in P) or by the refreees, nor has its straightforward relevance to atomic timekeeping, the Cannon-Jensen finding or the Sagnac effect been heretofore recognised. The reason becomes apparent from the derivation in the follow-up paper (p.417) that classical transverse aberration suffices to explain all the results. To the precision of the data, the kinematic terms in equation (1) are properly explained by classical, *not* relativistic, transverse aberration. In effect, a paradigm shift in the *interpretation* of rotating clock behaviour is required.

This misinterpretation of existing theory resulted in the following: (i) The only available heuristic route to explore the Cannon-Jensen finding became the unconventional consideration of one-way propagation effects. (ii) Because the most advanced one-way theory, that of Ives (pp.388-9), was based on the largely ignored Lorentz-ether formalism, a further unconventionality resulted. (iii) The supporting connection to the Sagnac effect which followed from Ives' little known rigorous analysis[29] added another element of unconventionality. (iv) The findings are not a *result* of Lorentz invariance, as conventionally assumed in the rejected paper, but are ultimately shown in the follow-up paper (p.418) to be properly described as *not in conflict with* Lorentz invariance. (v) The obvious connection of one-way theory to the Hafele-Keating experiment was denied by the latter's premature widespread affirmation of relativistic clock behaviour combined with the neglect of one-way propagation fostered by Einstein's *definition* of simultaneity. (vi) All these coalesced to reduce the probability of acceptance by the referees. Thus the final resolution of rotating clock behaviour stimulated by the Cannon-Jensen finding involved a series of considerations leading inevitably to rectification of a prevalent misinterpretation which, however, served to block the dissemination of the very considerations by which it could be eventually rectified. This required paradigm shift becomes the basic origin of the review rejection and shows the original paper to be more revolutionary than evolutionary.

Extension of equation (1) beyond Moon and Sun to the Galaxy is also found in the follow-up paper to show the absolute motions of Sun and Galaxy to be within reach (Fig.8, p.414). This should allow, in time and by tightening of the UTC coordination process, a more accurate measure of these motions than by the difficult astronomical methods. The small solar gravitational effect (p.400) must, of course, be included in any complete evaluation of annual residuals from ellipticity of Earth's orbit. For a clock on Earth or Sun with respect to a clock at rest in the universal frame, the kinematic effect in equation (1) also explains the relatively large drifts between the various time standards (pp.415-6). This is of practical importance in astronomy where Earth, solar and atomic time scales have up to now been coordinated empirically. The suggestion of an acceleration origin of the still unexplained large quasar emission line-widths (Ref.12, p.420) follows from equation (12) (p.391), which was not otherwise applied in the analyses. The data also allows a long-distance upper limit to departure from the isotropy of c, as assumed by Einstein and in the analyses, which is determined by the residuals in Fig.8 (p.414) after all other annual effects are deducted.

Most important, the ability of second-order aberration to measure absolute motion of relatively rotating clocks is a validation of an absolute reference frame for light propagation (p.418) which provides a direct confirmation of the Lorentz (relativistic) ether (pp.391-2). This is supported by the recent measurement of the anisotropy in the cosmic background radiation which has been claimed to demonstrate existence of a "new ether"[31]. However, there is nothing new about it since there can only be one such ether frame at rest in the universe — the Lorentz ether already known to be identifiable with the cosmic reference frame for acceleration (p.392).

Besides confirming the rejected paper and clarifying clock behaviour, the follow-up paper indicates how applications may directly proliferate from innovation and hence that these are delayed by improper rejection. This case history thereby links the probability of self-destruction of science from a deficient growth rate directly to the mental inertia inherent in the minds of men as summarised by the Innovation Theorem.

7. TIME FACTORS IN THE REVIEW PROCESS

In view of the increasing need for urgency in solving today's technological problems, opportunities for decreasing unnecessary time delays are of interest. Accordingly, the time factors in this review process are summarised in Table 2 (at the end of this paper).

The greatest delays were incurred by the reviews themselves which accounted for 75.5 percent of the total time to reach a final decision. The time for a single review response averaged to 109 days, which includes the two-way transit time between editor and referees. Since the review cycle represents the greatest opportunity for time reduction, it would be useful to know the distribution of "dead time", i.e. time delays not affecting review quality, due to editorial office, referees and transit. Such data are intrinsically determinable by the journals and demand systematic analysis.

The major cause of the 433 days required to reach a decision is the large number of reviews which, in turn, was aggravated by the low review precision. This case history thereby suggests that an increase in review precision presages shorter review times and reduced load on the journals and research community with its attendant financial, psychological and social benefits.

The total time for a review may be small compared to the delay in final publication caused by a rejection. Due to the nature and period of the subject review, there were no plans to repeat such an exasperating experience. Publication of the two papers would have been indefinitely delayed were it not for the circumstances associated with the inception of SST (p.386) and insistence of the editor of SST on further corroboration.

The high delay per response by reviewers suggests a general disregard for urgency in present review practice. Due to the sequential proliferation of discoveries, which depend on prior discoveries dependent on still prior discoveries, etc., the development of science is exponentially stretched out by unnecessary delays. That a marked reduction in total publication time is feasible has been demonstrated by the *Journal of Clinical Psychiatry*[32] and in physics by Azbel[9] in his comparison of *JETP Letters* and *Physical Review Letters*. He attributes the greater speed of JETP to editorial requirements for greater review precision and shorter response time. In reply, the editors of PRL state[10] that publication delay time "can be reduced substantially only by increasing our costs — and our page charges — significantly *and we choose not to do so*" (italics added). If this direct affirmation of a low priority for urgency is merely a matter of cost to the journals, why is it not also evaluated with respect to the inordinate cost of slower technological growth[15] to society? Can we put a price on increasing the probability of civilization's survival and the attendant improvement in the quality of life? In this light the solution is self-evident: since society is the chief beneficiary it should provide the

necessary funds. However, the isolation of the internal workings of the science establishment from public view has thus far obfuscated the fact that return on any investment in speeding the dissemination process can be matched by few, if any, other investments of our resources at present. Society spends large sums for research and development of specific projects initiated by scientists but negligible amounts on the publication bottleneck, the major path for dissemination of the knowledge so obtained which, in essence, determines the return to society on its initial investment. It is the anachronism of our times that, because of cost, many of the technological offsprings of the science process, e.g. modern communications, computers, fast publication techniques, psycho-social advances, opinion surveys, advanced management methods and priority mail, among others, are not being maximally employed to further enhance the process that gave birth to them. The present state of research publication may be generally compared to the proverbial shoemaker without shoes.

8. TOWARD A SCIENCE OF PEER REVIEW

Peer review is such a multi-faceted, strictly human endeavour that the perennial question, "Does peer review work?" invariably results in a dialectic controversy. The most effective known approach to minimize this is an operational one — reasoning based on measurement — which requires the question to be reframed as "precisely how well does peer review work?" This stress on measurement switches the basic emphasis from disagreement to agreement. Once the parameters of peer review are properly measured, ways to improve it become self-evident through tests, further refinement, further tests, etc. — the prototype of the scientific method. This demands that we begin with study of the basic phenomena themselves, review case histories. The one presented here graphically confirms the need for publication of many more to enable a start in this direction. Nonetheless, it is also imperative to determine what we now glean from this one case.

An expressly recognised goal of review is to preclude defective work; a heretofore neglected goal stressed herein is the necessity to also preclude erroneous rejection of publishable work. Erroneous acceptance is minimized by the current practice of *parallel reviewing* and although it is desirable to contain it, acceptance errors are ameliorated by the back-up practice of assigning high priority to correction of published errors. It may be further ameliorated by publishing referee comments, e.g. as by SST when warranted. However, erroneous rejection leads to conflict, which necessitates *sequential reviewing*, but this has no back-up error-correcting mechanism other than the unsatisfactory, lengthy and uncertain one of submission elsewhere. It is erroneous rejection that results in gross inequities in the review process, causes the most friction and dissatisfaction to authors and journals and produces the most serious consequences for society by slowing technological growth rate. By merely up-grading the precision of sequential reviewing a much improved system is obtainable with minimal disturbance to the present system.

The primary parameter of sequential reviewing is the fraction p of contested manuscripts per review cycle (which may include one or more parallel reviews). Let N_s be the number of submitted manuscripts and N be the

number subjected to review. Then N_s - N is the number rejected primarily for nontechnical reasons, as nonconformance to editorial standards. After m sequential review stages the number that remain contested becomes $p_1 p_2 \ldots p_m N$. If $\langle p \rangle$ is a mean value for all stages, the lower limit to n defined above in the introduction is then given by

$$\langle p \rangle^m = 10^{-n} \qquad (4)$$

For $\langle p \rangle = 0.01$, $n = 2m$. For $m = 2$ (3), the occurrence rate of unresolved conflicts is 1 out of 10^4 (10^6) reviewed manuscripts. Such examples clearly define the goal of sequential reviewing: (i) The value of p must be minimized. (ii) Use m stages to obtain any desired maximum rejection error rate, i.e. desired minimum value of n. (iii) Prevent p from degenerating between stages.

The lack of attention to rejection error in present peer review is not conducive to a minimum value of p or its uniformity between stages. The result is a high conflict rate, e.g. as in *Physical Review Letters*[9,10]. An intrinsic cause is suggested by this case history to lie embedded in the properties of the human mind, as summarised by the Innovation Theorem. Since minds are immensely varied and their modes of operation are still unknown, broad generalisations, theories and panaceas for improving peer review do not yet have an adequate operational basis. The alternative is the empirical one of identifying each specific source of human error and devising appropriate means to minimize it. This approach demands intimate knowledge of review details, as afforded by case histories. Heuristic conclusions deriving from this one case history include the following:

1. The decision to reject was not based on a technical resolution of the conflict but on the prevalent criterion of consensus of the referees. This criterion was not fail-safe because the basic issues were by-passed. No decision should ever be forced when technical disputes remain unresolved. Elimination of this major source of rejection error is simply obtained by *formally defining a contested review to be considered complete only when there is agreement between author and reviewers.* Decision to accept or reject is then automatic and anticlimactic.

2. Such a definition shifts the traditional emphasis on erroneous acceptance, which has built-in error correction mechanisms, to at least equal emphasis on erroneous rejection, which has substantially none. Preoccupation with erroneous acceptance stems from the birth of modern science when there were no precedents for publication norms except the empirical need to establish order and rigour.[7] Today the norm of scientific methodology is well entrenched but the old tradition lingers on. Must we wait for a worldwide crisis to realize that the needs of yesterday are reversed by the needs today for a growth of science independent of its own dogma, traditions and preconceptions and a concomitant need to publicly acknowledge all unfalsifiable innovations and dissident views as rapidly as possible? To expedite the desired resolution of author-reviewer disagreement we need only institute appropriate specific measures, such as:

(a) If public recognition and education by the journals for the consequences of erroneous rejection increase general awareness of authors and

reviewers for the necessity to attain agreement, a spontaneous improvement in review precision and speed should ensue.

(b) More circumspect and rigorous argument is encouraged if all reviewer comments are subject to the possibility of publication. The role of anticipation of rejection in enhancing author performance, commonly advanced to support peer review, is thereby extended to reviewers. Occasional publication of selected examples for tutorial purposes may serve to additionally educate the science community. A special journal, as suggested by Commoner[12], may be warranted.

(c) Investigate the possibility of classifying the significance of all manuscripts from evolutionary to revolutionary on a scale of, say, 0 to 10, and assign a suitable weighted mean R derived from author, reviewer and editor estimates. Besides the value of R as a caution flag for properly resolving the more innovative approaches, determination of the potentially useful p-R distribution function may be facilitated.

(d) The inherent role of the editor as adjudicator is too often degraded to a clerical role, as in this case history, so that reviewers are tacitly assigned the dual role of prosecutor and judge. In a dispute the reviewer is no longer a "referee" but a contestant and should be so regarded. If the editor, or a designated impartial arbitrator, does not exercise a supervisory role, e.g. akin to a referee in a sports contest, the result is an increase in the incidence of arbitrariness, as in M, and rejection for the wrong reasons, as in C, D and N.

(e) Because of the express need to resolve a dispute, it is expedient to require reviewers to indicate whether and/or how author errors may be corrected to allow acceptance where possible. Suggestions from reviewers are not the rule, but are often very useful to authors, even for language usage as by N. Besides quickly resolving a dispute they induce respect for peer review[33] and should be made mandatory by editorial dictum.

(f) For a more serious dispute the simple expedient of explicitly delineating and narrowing the boundary of a disagreement provides a useful resolution vector. (Of course, all prior information must be meticulously forwarded to reviewers and authors.) In the case history five reviews were not made available to the author and there was no attempt in the final review stage to relate to the prior reviews. One reviewer (M) reneged and the other (N) went off in a different direction. Delineation of the boundaries of a dispute may be forced by editorial edict, as by demanding authors and reviewers to indicate agreement or disagreement on all segments of reports on suitable forms that must be returned with comments.

(g) In the event of an impasse a "closed-loop" review may be instituted by requiring author and reviewer to communicate directly, with copies sent to or through the editor. Speedier resolution is facilitated by the higher information transfer rate, but close supervision by editor or arbitrator is essential to prevent degeneration, e.g. as for a boxing match vis-a-vis a bar-room brawl. An example of a successful application is mentioned in O. Useful author-reviewer dialogues with maintenance of anonymity have also been instituted by *Chest*[13]. The still higher information transfer rate allowed by telephone suggests a further extension worth investigating.

3. Periodically publish summaries from data gathered from authors and reviewers to enable public analysis of review precision, variation with time and interjournal comparisons. Such data provide the essential feedback to test and further improve peer review. Request essential data on all important time factors from reviewers on properly designed forms and periodically publish analyses of these.

These tentative conclusions from one case history are cost effective for the journals to the extent they reduce the values of p and m. Additional measures for improving peer review and its study are possible which require investment by society. However, these are justifiable by the long-term benefits that may accrue.

The probability that $\langle p \rangle = 0$, and correspondingly that $n = \infty$, is remote so there undoubtedly must remain some contested manuscripts for which there is no author-reviewer agreement within a reasonable period. (The cut-off point that defines "reasonable" must be set by practical considerations, as the value of R, duration, number of reviews, space limitations and/or cost.) Because these contested remnants may involve fundamental issues, it is not mete to relegate them as heretofore, to the oblivion of journal files. Some may be suitable for mandatory publication with comments, but it may be appropriate to form an independent council, akin to an appeals court in jurisprudence, to openly review all remaining unresolved cases. Publication of the council proceedings, perhaps in a special journal, provides a reference that may be indispensable for preservation of unfalsifiable dissident views that occasionally erupt into major paradigm shifts. Such an appeals function also insures that no inequity of the review process need ever be ignored or denied; it provides the missing negative feedback for closing the loop in the process of peer review.

The efficacy of an appeals council is likely to be intimately related to the choice of peers. This addresses the perennial problem of selection of peers in general. The problem is perspicuous from the revelation that a 10 to 20 percent elite group of scientists accounts for 80 to 90 percent of published research.[34,35] It is obviously impractical for the large output of this small group to be reviewed only by itself. For a random assignment of reviewers based only on professional knowledge, as approximated in current practice, how is it then possible to provide a proper peer match for the evident creativity of this indispensable elite group (or any other subset)? The resulting mismatch is undoubtedly responsible for much of the discontent with peer review. For a start, it is already known that general intelligence as measured by IQ is not significantly correlated with success in research[34], that IQ and creativity are not significantly correlated and that the factors involved in IQ (creativity) are mainly determined by left (right) brain function. Since creativity is measurable with a reliability equivalent to that of IQ[36], it appears feasible to begin to understand and investigate at least one important factor involved in peer matching other than professional expertise and, eventually, to extend this to other measurable attributes that may be involved. With development of suitable tests, the ultimate establishment of a core of relatively few properly trained professional reviewers may be the most cost-effective and ideal way to solve a large part of the review problem.

The long ignored psychological effects of rejection and the attitudes and needs of authors and reviewers are now determinable by professionally designed opinion polls. A test of the Innovation Theorem suggested by this case history and the related psychological origins of resistance to innovation is another neglected area; the possibility of rating already published work on a fairly accurate evolutionary/revolutionary scale R allows evaluation of its important distribution function versus acceptance delay to permit measure of n and its significance as an overall measure of resistance to innovation in peer review.

In summary, the basic conclusions from this study of a case history are: (i) The status of peer review as a scientific endeavour is in a very primitive state. (ii) It is imperative and feasible to vastly improve it.

References

1. Christiansen, D., *IEEE Spectrum*, **12** (11), 29 (1975). Burch, G.E., *Amer. Heart J.*, 97, 265 (1979).
2. Asimov, I., *Asimov's Biographical Encyclopedia of Science and Technology*, Avon Books, New York (1976).
3. Examples of known cases are Hermann Grassmann and J.J. Waterston. See Zuckerman, H., and Merton, R.K., *Phys. Today*, 25 (2), 9 (1972).
4. Jones, R., *New Sci.*, 61, 758 (1974).
5. Denim, S., *New Sci.*, 64, 925 (1974).
6. Ruderfer, M., *Spec. Sci. Techn.*, 1, 219 (1978).
7. Zuckerman, H.A. and Merton, R.K., *Minerva*, 9, 66 (1971).
8. Zuckerman, H.A. and Merton, R.K., *Phys. Today*, 24 (7), 28 (1971).
9. Azbel, M., *Phys. Today*, 31 (12), 82 (1978); 32 (10), 96 (1979).
10. Adair, R.K., Trigg, G.L. and Wells, G.L., *Phys. Today*, 31 (12), 82 (1978).
11. Adair, R.K., *Phys. Today*, 32 (10), 101 (1979). Strandberg, M.W.P., *Ibid.*, 98. Gordon, R.A., *Ibid*, 31 (10), 81 (1978); 32 (4), 13 (1979). Kumar, K., *Ibid*, 32 (4), 13 (1979). Stumpf, W.E., *Science*, 207, 822 (1980).
12. Commoner, B., *Hosp. Prac.*, 13 (11), 25 (1978). Curran, G.L., *Ibid.*, 14 (2), 17 (1979).
13. Soffer, A., *Chest*, 75, 295 (1979).
14. Cole, J.R. and Cole. S., *Science*, 178, 368 (1972), p.372.
15. Boyd, R., *Science*, 177, 516 (1972).
16. Cook, R.E., *Science*, 198, 22 (1977). Baker, V.R., *Ibid.*, 202, 1249 (1978). Yalow, R.S., *Ibid.*, 200, 1236 (1978). Wade, N., *Ibid.*, 201, 31 (1978).
17. Ruderfer, M., *Spec. Sci. Techn.*, 2, 387 (1979).
18. Ruderfer, M., *Spec. Sci. Techn.*, 2, 385 (1979).
19. Ruderfer, M., *Spec. Sci. Techn.*, 2, 405 (1979).
20. Ruderfer, M., *Spec. Sci. Techn.*, 3, 231 (1980).
21. Cannon, W.H. and Jensen, O.G., *Science*, 188, 317 (1975).
22. Pound, R.V., and Vetterling, W.T., *Science*, 191, 489 (1976). Allan, D.W., Mungall, A.G. and Wrinkler, G.M.R., *Ibid.*, p.490.
23. Penny, C.J.A., Smith, H.M., and Wilkins, G.A., *Science*, 191, 489 (1976).
24. Cannon, W.H. and Jensen, O.G., *Science*, 191, 490 (1976).
25. Barber, B., *Science*, 134, 596 (1961).
26. Shapiro, I.S., *Sov. Phys. Uspekhi*, 15, 651 (1973); Russian original in *Usp. Fiz. Nauk*, 108, 319 (1972).
27. Hallam, A., *Sci. Am.*, 232 (2), 88 (1975).

28. Sadeh, D.S., and Au, B.D., *Nature*, 224, 1291 (1969). Sadeh, D., Knowles, S., and Au, B., *Science*, 161, 567 (1968).
29. Ives, H.E., *J. Opt. Soc. Amer.*, 28, 296 (1938).
30. Hafele, J.C. and Keating, R.E., *Science*, 177, 166 (1972).
31. Muller, R.A., *Sci. Am.*, 238 (5), 64 (1978).
32. Easson, W.M., *J. Clin. Psychiatry*, 40, 331 (1979).
33. Hanson, H., *Spec. Sci. Techn.*, 2, 472, 467 (1979).
34. Cole, J.R. and Cole, S., *Social Stratification in Science*, Univ. Chicago Press, Chicago (1973).
35. Maugh II, T.H., *Science*, 184, 1273 (1974).
36. For example, the Johnson O'Connor Research Foundation (U.S.) has for decades applied a creativity test having a correlation coefficient ≈ 0.9.

Table 1 — Review Chronology

Date	Appendix	Content	Route*	Time (days) Accumulated	Time (days) Delay
18 Feb. 1976	A	Letter and MS	au to ed	0	
23 Feb.	—	Received by *Science*		5	5
6 May	B	Letter + C + D	ed to au	78	73
	C	Report of referee			
	D	Report of referee			
17 May	E	Letter and MS	au to ed	89	11
21 May	—	Received by *Science*		93	4
20 August	F	Telephone call	au to sec	184	91
1 October	G	Telephone call	au to ed	226	42
15 October	H	Letter of rejection	ed to au	240	14
21 October	I	Telephone call	au to ed	246	6
22 October	J	Letter and MS	au to ed	247	1
9 November	—	Received by *Science*		265	18
11 Feb. 1977	K	Letter	au to ed	359	94
24 February	L	Letter + M + N	ed to au	372	13
	M	Report of referee			
	N	Report of referee			
26 February	O	Letter + P	au to ed	374	2
	P	Reply to N			
11 March	Q	Letter	ed to au	387	13
16 March	R	Letter	au to ed	392	5
26 April	S	Letter	ed to au	433	41

* au = author; ed = editor; sec = editor's secretary

Table 2 — Summary of Correspondence Time Delays

Response	Appendix	Delay (days)	Delay Totals (days)	Fraction of Total (%)	Average delay per response (days)
MS in transit:	A	5			
	E	4			
	J	18	27	6.2	9
Review reports:	A-B*	73			
	E-H*	147			
	J-L*	107	327	75.5	109
Author letters:	B-E	11			
	H-J	7			
	L-O	2			
	Q-R	5	25	5.8	6.3
Editor letters:	O-Q	13			
	R-S	41	54	12.5	27
Totals:		433	433	100.0	

* Less manuscript (MS) transit time

Appendix A

The Editor,
Science.

Dear Sir,

On 2 June 1975 I submitted an MS attempting to correct the article by Cannon and Jensen (188, 317). This was justifiably rejected in your letter of 29 July.

I have since examined their article in more detail, as well as the Technical Comments you recently published, and have discussed with Cannon and Jensen their original article. I find that there is a satisfactory explanation of their findings, including their subsequent negative result, which has been overlooked. This is discussed in the enclosed MS entitled "One-Way Doppler Effects in Atomic Timekeeping", which is hereby submitted for publication in *Science.*

My approach is based on research I have been conducting over the past 15 years on the one-way velocity of light and its interpretation. This is a little-explored area, but it is directly relevant to the findings of Cannon and Jensen. It not only accounts for their work, but also clarifies certain observations by Sadeh and associates which, to my knowledge, have not yet been satisfactorily explained.

Only those knowledgeable in atomic timekeeping would be appropriate as referees, such as the reviewer(s) of the Cannon and Jensen reports. *(Material deleted.)* Of course Cannon or Jensen would be appropriate unless you consider a possible conflict of interest to be an objection.

For the convenience of the reviewers I am also enclosing reprints of cited articles of mine and copies of computational notes.

Sincerely yours,
Martin Ruderfer.

Appendix B

Note: All letters from Science have been paraphrased.

6 May 1976

Dear Dr Ruderfer,

Your paper on "One-Way Doppler Effects in Atomic Timekeeping" has not been accepted and the referee's comments and your manuscript are enclosed.

Yours truly,
Editorial Staff.

Appendix C

Note: By the kind permission of the referee, this report is reproduced with the exact wording of the original.

The author is apparently unaware that Cannon and Jensen handling of the atomic clock data was incorrect. This was pointed out in a letter to *Science* which I reviewed and which was from Greenwich Observatory scientists.

Cannon and Jensen were apparently unaware of synchronization procedures which made their data analysis incorrect. Thus, Ruderfer's explanation of "their findings" (top of page 3) [bottom of p.387 and top of p.388] does not in my opinion deserve publication.

Appendix D

Note: This referee refused permission to reproduce his comments exactly. The following is a paraphrased version of the original report.

The referee rejected the paper for unmentioned objections but, for brevity, cited only these typical instances:

(a) The derivation is so imprecise and inconsistent it becomes without meaning. As an example, on [p.389] phase is indicated to be constant because it is given as frequency X time in equation (6), but in equation (10) "frequency suddenly becomes time-dependent". Also T is treated as time dependent in equation (10) but it is later treated as an "inconsistent constant" in equation (11) upon removal from the integral.

(b) In the extension of the analysis on [p.398] to include the results of Earth's orbital motion, the effect of Sun's gravitational motion is neglected. Essential relativistic rate variations at a clock fixed on Earth are thereby omitted, resulting in a "seriously incomplete" analysis.

(c) There is insufficient accuracy in the experimental data to detect the effects predicted; the author misconstrued "accuracy, precision and stability" as used in timekeeping. Predicted rate effects are of the order of (3 to 7) X X 10^{-13}, but UTC_i cesium clocks have an intrinsic accuracy of only 2×10^{-12} as stated by Allan, et al, *Science*, 191, 490 (1976) in rebutting the Cannon-Jensen report. This is about an order of magnitude short of that necessary to measure the effects of the author. The same rebuttal more seriously notes that the UTC_i time scales are intermittently adjusted in coordinating them to UTC so that each clock rate is not founded "on any fundamental physical process, as required by theory".

Appendix E

The Editor,
Science.

Dear Dr Abelson,

Thank you for including the referees' comments with your 6 May return of my MS "One-Way Doppler Effects in Atomic Timekeeping".

Both referees have improperly rejected the MS. The one submitting the 6-line comment (which I shall refer to as referee [C]) was grossly in error. The other [D] suggests to me the need for additional clarification in the MS. I have therefore revised the MS accordingly, including some cosmetic improvements, and am herewith resubmitting with the following comments. Except for the two revised pages substituted for page 7 [pp.390-1] and the added insert for page 19 [p.400], all other changes are marked in red.

The discrepancy in Cannon and Jensen's data handling to which referee [C] appears to be referring is in regard to one station (RGO). I was aware of this discrepancy, which the referee could have ascertained if he had read the MS through, because I: (1) discuss this on page 14 [pp.401-2]; (2) refer to the Greenwich Observatory report which [C] reviewed (ref. 26); (3) explicitly exclude the RGO data from my analysis; and (4) was appraised of the cause of the discrepancy in my discussions with Cannon and Jensen. The "synchronisation procedures" mentioned by [C] consider only the absolute *calibration* of a station's proper time to the international second (SI); this process *is unrelated to the one-way synchronisation process I discuss*, which

applies only to the synchronisation of any two remote time scales *irrespective* of their absolute calibration. I make this clear in the top paragraph on page 22 [p.402]. I trust that review [C] will not influence any further review by being given weight as a prior rejection.

The comments of referee [D] are more relevant but, unfortunately, reveal a misunderstanding of portions of the MS. The following comments correspond to his headings.

(a) My paper does not "derive a number of well-known effects", but deals with effects that have *not* been systematically exploited heretofore. The only truly "one-way" effects that are well-known are the conventional Doppler shift and aberration relations. The effects of n, dn/dt and the double-disc Sagnac experiment which I discuss have not appeared elsewhere to my knowledge — and I have been specifically searching for such one-way effects in the literature for over 15 years. Perhaps they have been mentioned in some remote place; if so, they certainly are not "well known". The only well-known refractive effect in the literature related to time dilation is the refinement by Lorentz, confirmed by Zeeman, of the Fizeau and related experiments, e.g. D.A. Evans, *Int. J. Theor. Phys.*, 2, 313 (1969). This is a two-way effect. The one-way refractive effects I discuss disappear in a two-way measurement.

The referee appears to be unfamiliar with phase modulation theory and its accompanying concept of instantaneous frequency, which is the physical origin of the frequency shift Δf that I derive. This accounts for his confusion regarding the frequency and travel time variables. Note that I verify the form of the Doppler effect I derive by showing its consistency with the conventional form by the examples I give for recessional and transverse motion. What the referee's remarks do indicate, however, is that the derivation is too succinct. Accordingly, I have revised page 7 and have replaced them with pages 7 (rev.) and 7a (rev.) [pp.390-1].

(b) I neglect the effect of Sun's gravitational potential (as well as those of other solar bodies) because they are negligible here. I have therefore added a paragraph (p.19a) [p.400] to dispose of this objection explicitly.

(c) I have followed Cannon and Jensen's use of clock "reproducibility" which is appropriate in my application. The referee's concern with absolute calibration of proper time to SI has been extensively discussed in your 6 February issue (191, 489-491). Although this may be pertinent to Cannon and Jensen's theory, *it is irrelevant to the testing of the one-way synchronisation effects I discuss.* I spent considerable space explaining this in the Discussion section. Why has [D] overlooked this? I suspect he did not read this far.

To illustrate that the "inherent accuracy" of clocks is irrelevant, consider the synchronisation of any watch to any wall clock. To do this, it is only necessary to periodically reset the time on the watch to that of the wall clock. The time *difference* between the two is then only dependent on the resetting precision and is not directly related to the absolute precision of the two clocks. This is precisely what is done by the time laboratories in synchronising the UTC_i to UTC, as has been noted in ref. (6), in the adjacent article by Allan, et al, and in my MS. *Only the difference between these is required to test the predicted one-way Doppler effects*, as I discussed. Paradoxically, [D] also notes that "the UTC_i time scales are periodically adjusted to coordinate them with

UTC", but wrongly cites this as prejudicial to my findings by stating these time scales are not "based on any fundamental physical process, as required by theory". This may apply to the Cannon-Jensen theory, but emphatically does not to one-way synchronisation, as I explained in the MS. This further confirms that [D] may not have read my Discussion section.

The referees' comments are undoubtedly due at least partly to the heretofore obscurity of one-way theory, its application and its interpretation. Nevertheless these apply directly to the Cannon-Jensen and Sadeh-Au data and to the theory of noninertial flat-space phenomena. Although one-way theory has been around for some time, it has still not entered the mainstream of current thinking. When the paradigms of the establishment are challenged the worst failings of the referee system surface, but the potential for advance is maximum. I therefore hope that any further review will properly consider the matters I have raised.

Since the two copies of the computational notes and reference reprints are still in your possession, I trust that they will be forwarded with the MS in any further review, in addition to the comments above.

Sincerely yours,
Martin Ruderfer.

Appendix F

The author telephoned the secretary to the editor on 20 August 1976 and was informed that the editor is out of the country and would return on 30 August. In reply to a request for status of the paper, the secretary stated one reviewer reported the paper to be "excellent" and a second was negative. The paper was sent to a third reviewer whose report is overdue.

Appendix G

The author telephoned the editor on 1 October 1976 to inquire on the status of the paper. The editor confirmed that one reviewer cited the paper as "excellent" and that the second was equivocal. The paper is still in the hands of a third reviewer who is being prodded to reply.

Appendix H

15 October 1976

Dear Dr Ruderfer,

Your paper, "One-Way Effects in Atomic Timekeeping", is not accepted. The pressure of papers is very large and even when publication is merited, we often decline such papers.

Yours truly, Editorial Staff.

Appendix I

The author telephoned the editor on 21 October 1976 to clarify the letter of rejection [H], which did not contain the comments of the three previous reviews. The editor reported that the third reviewer commented that the paper was too diffuse, rambled, did not really contribute anything new and that he could not decide what was being shown. The author replied that such a rejection was too vague and was additionally inadequate because the paper involved correction of a double error in the record — the original published paper and the published comments on it. The editor responded that it was not certain that there was an error to be corrected, that he had to go by the consensus of referees and that the paper could always be submitted elsewhere. The author contended that this was not a certain solution since if he died before receiving an acceptance, the errors may never be corrected in the literature. The editor then agreed to a further review.

Appendix J

22 October 1976
The Editor,
Science.

Dear Dr Abelson,

Thank you for the opportunity to discuss with you the status of my submission "One-Way Doppler . . ." in my call on October 21. In accordance with our discussion I am enclosing the MS for further review.

As I mentioned, one of the problems in the review of my MS is the heretofore obscurity of one-way theory which I claim is directly relevant to the Cannon-Jensen and the Sadeh-Au findings. This unfamiliarity has interfered with the review — the referees have substituted unsubstantiated opinion for proper technical evaluation. The point they seem to have ignored, and one which is of primary concern to your office, is that if my thesis is correct the published considerations of the Cannon-Jensen report is in serious error and should be forthwith corrected in the printed record. The only other viable alternative is to unequivocally show that my approach is wrong.

If further reviewers are informed of this need, it may stimulate a more careful and diligent review. Since the matter is purely a technical one, the question is resolvable. As I indicated, I am willing to communicate directly with any referee, with copies sent to your office, to reach a mutually agreeable decision in the event there are any questions that require clarification in order to reach such a decision.

Sincerely yours,
Martin Ruderfer

Appendix K

11 February 1977
The Editor,
Science.

Dear Dr Abelson,

It is almost four months since I resubmitted my MS "One-Way Doppler . . ." for further review and one year since my original submission.

Is there some problem with the present review and can I assist in any way?

Because the MS has raised the question of a serious error published in your journal relating to the original Cannon and Jensen report and the further re-enforcement of this error in your Technical Comments section, such delays are deleterious to the best interests of scientific progress.

I trust that this matter can be resolved to our mutual satisfaction soon.

Sincerely yours,
Martin Ruderfer.

Appendix L

24 February 1977

Dear Dr Ruderfer,

Your paper, "One-Way Doppler Effects in Atomic Timekeeping", has not been accepted. The manuscript and referee's comments are enclosed. We hope this may help you in the modification of the paper for resubmission somewhere else.

Yours truly,
Editorial Staff.

Appendix M

Note: This referee refused permission to reproduce his comments exactly. The following is a paraphrased version of the original report.

The referee states that he still cannot find anything to recommend publication and feels that additional resubmission should not be encouraged.

Appendix N

Note: This referee refused permission to reproduce his comments exactly. The following is a paraphrased version of the original report.

The referee recommends against publication, stating the author attempted to explain the Cannon-Jensen discovery to be due to anisotropy in the one-way speed of light despite the Cannon-Jensen retraction of their claims and the three criticisms of their paper published in *Science* showing their discoveries to be without experimental support.

To demonstrate that the submitted paper has neither experimental nor theoretical support the referee offers this theory: For $\phi = 2\pi fT$ as the phase of the stationary clock at its location, the phase at some position x is $2\pi(fT - x/\lambda)$. At the receiving clock at $x_i = x(T)$ the phase is

$$\phi_i = 2\pi[fT - x(T)/\lambda]$$

for propagation in the x-direction. The received variable frequency wave by the moving clock should be $2\pi \int_0^T f_i dT_i$, not $f_i T_i$, since dT_i is "the elapsed proper time on the moving clock". Equation (7) becomes

$$2\pi[fT - x(T)/\lambda] = 2\pi \int_0^T f_i dT_i$$

On substituting the Einstein time dilation relation this becomes, using $f\lambda = c$

$$f[T - x(T)/c] = \int_0^T f_i(T)(1 - v^2/c^2)^{1/2}\, dT$$

where v is relative velocity. Equations (9) and (10) are incorrect because the author's differentiation for an observer travelling with a wave front of constant phase is impossible due to the constant phase wavefront travelling with speed c with respect to the stationary observer, since clocks cannot travel at speed c.

Moreover, the last equation may be differentiated with respect to T to give

$$f_i = f(1 - \dot{x}/c)/(1 - v^2/c^2)^{1/2}$$

to obtain the correct Doppler signal $f_i - f$ received by the moving clock, noting that the "change" df occurs in equation (10). Because the emitted frequency of the stationary clock is fixed, f cannot change so df has no meaning.

Also f/T has been treated as a constant in equation (11) and the received, instead of emitted, frequency is misinterpreted as f. This is illogical, because the integral is over dt and it is not possible to distinguish "between dt and dT. T = t is the coordinate time in the frame of the stationary clock."

The correct Doppler shift replacing equation (11), including longitudinal and transverse cases, should be

$$\Delta f = f_i - f = f_r - f_t = f_t[1 - (1 - \dot{x}/c)/(1 - v^2/c^2)^{1/2}]$$

Equation (13) is thus incorrect, there are a number of errors in the derivation, and T is not rigorously defined, e.g. for an emitted spherical wave and a circling receiver, the source-receiver distance is constant and hence T = 0. No Doppler shift is predicted so the transverse Doppler shift seems to be overlooked.

Because the referee believes equation (13), on which the author bases his discussion, is incorrect and there are a number of errors in the derivation, he is of the opinion that the discussion is not convincing.

The refractive effects are a "subsidiary issue" not useful to the author for establishing existence of one-way effects.

There appears to be no good reason for writing equation (20).

The author's reference on line 2 above equation (16) to t as the "time measured by the stationary clock" is not clear as clocks on Earth's surface are all stationary in Earth's reference frame. "The meaning of t is not made clear until [p.394]."

"Phenomenally" should be replaced by "phenomenologically"; "rigidly" should be replaced by "rigorously"; and T_x' on [p.393] is the systematic error, not the uncertainty in T'.

In the [p.400] discussion, Sun's effect is not locally detectable, by the principle of equivalence, in a freely falling (locally inertial) frame, as Earth. A second-order Doppler effect due to Earth's orbital motion, overlooked by the author, cancels the term $g_s h/c^2$.

Appendix O

26 February 1977
The Editor,
Science.

Dear Dr Abelson,

Thank you for the inclusion of the latest referees' reports with the return of my MS "One-Way Doppler . . ." and for your continued cooperation in trying to resolve the correction of the Cannon-Jensen reports brought up by the MS.

The short review is of no value in settling the technical problem; the referee is apparently reluctant to tackle the issues and is passing the buck. The longer review appears to be an attempt to resolve the matter; however in a technical sense it is inane and ineffective. For your reference I enclose my comments on this review.

This matter would be ludicrous if it were not so serious — the advancement of science and technology is the key to man's future survival. The scientific journals require rigorous adherence to scientific principles in submitted manuscripts, yet they consistently neglect to apply equal rigour in judging them. To editors, reviews by one or two referees without further verification are sufficient cause for arbitrary rejection. But no scientific matter can properly be decided by the political expedient of popular vote. Would you allow any author to justify a theory or conclusion with the statement that it is the most popular? When is something to be done about putting science into the *judgement* of technical submissions?

I offered one way in our telecon and in my letter of October 22, 1976 — a closed-loop review. If the referee had contacted me first I would have pointed

out his error and he would have been forced to respond more accurately. Closed-loop operation is the way nature and man minimize errors in all types of systems. Why not in the judgement of manuscripts, particularly for the few revolutionary ones vis-a-vis the more common evolutionary ones, as in the current case of the possibility of a serious error in the record? I cannot help but wonder why you did not take me up on this instead of repeating the same old error-prone open-loop process. By improperly rejecting the MS you are merely passing the buck. The Cannon-Jensen matter is a double error, it occured in your journal, and it should be rectified in your journal.

Editors are just as culpable as referees for inefficiencies in the review process. Instead of exercising their supervisory role, editors usually concur with referees without question and end up merely as correspondents between authors and reviewers. I have found only about 1 in 10 editors take an active helpful role in ensuring a proper review. In one case an editor insisted on a closed-loop resolution with speedy beneficial results.

Perhaps the matter of atomic timekeeping seems to you and the referees to be too remote from today's problems and therefore not worth the effort to pursue a proper conclusion. But the travesty that has occurred with my MS could equally have occurred in some urgent area, as energy, a field for which you obviously have much concern. How many potential solutions to the energy crisis have been denied publication by errors in the review process? I know of several simple potential solutions which warrant investigation, yet I hesitate to offer these to the scientific journals because of the three P's — Politics, Personality and Prejudice — which too often substitute for technical rigour in evaluating unsolicited manuscripts. The time and aggravation involved just do not seem worth the effort.

Science is the printed record. The referee process, by controlling this record, has a crowbar effect on the development of science. Errors in this process, by delaying or preventing dissemination of useful ideas and results, have a marked depressive effect on technological growth. Improved efficiency in dissemination of key developments would speed solutions to all problems compared to the rate under the existing slow and burdensome review process.

Sincerely,
Martin Ruderfer.

Appendix P

Author's comments on review of "One-Way Doppler Effects . . ."

The referee has confused the one-way travel time T with the proper time t. His initial expression ($2\pi(fT - x/\lambda)$) is meaningless since T is defined by the distance between source and receiver. The correct expression should be $2\pi(ft - x/\lambda)$. The remainder of his "derivation" is thereby useless.

Furthermore the referee's result cannot be correct since it does not agree with equation (8). This equation is the basis for relativistic aberration and should have alerted the referee that something was wrong with his derivation.

Although it is true that the phase wavefront cannot be measured in flight, it is nevertheless observable at the time the wavefront enters the observer's measuring apparatus. Equations (9) and (10) are thereby operational and pertinent.

The referee's confusion between T and t is exemplified by his remark that $T = 0$ for a rotating clock. Since $T = r/c$, it is necessarily > 0 when $r > 0$ as in the context discussed.

There is no uncertainty about the meaning of t — it is clearly defined in Equation (1).

In his last comment the referee has misapplied the principle of equivalence. The gravitational redshift gh/c^2 exists independently of the Doppler shift (e.g. see J.B. Thomas, *Astron. J.*, 80, 407 (1975), right column). For Earth, these give separate contributions to the total frequency shift of a surface clock and were so calculated by Cannon and Jensen. These were also independently evaluated in the Hafele and Keating experiment. An earlier referee rightly pointed out that corresponding contributions from Sun also exist at the clock site. I simply showed on p.19a [p.400] that the daily change in gravitational redshift between the clock positions nearest to and furthest from Sun, $g_s h/c^2$, is negligible. The orbital Doppler shift is included in the analysis comprising Equations (24)—(32).

In sum, the referee bases his rejection on an alleged falsification of Equation (13) which he correctly states is "crucial for the author's contention". However his attempt at falsification is grossly in error.

It should be further noted that Cannon and Jensen did not reject their original findings but only their theory. Without a theory they could not harmonize their original findings with the later one with about 80 clocks and chose the ad hoc explanation of an "artifact" in the data. My MS affirms that all of their data have a common explanation *required* by relativity.

M. Ruderfer

Appendix Q

11 March 1977

Dear Dr Ruderfer,

Your paper, "One-Way Doppler Effects in Atomic Timekeeping", has by this time been examined by nine people of known competence whose opinion is that we should not publish it. We decline to consider this paper any further.

Yours truly,
Editorial Staff.

Appendix R

16 March 1977

The Editor,
Science.

Dear Dr Abelson,

I cannot dispute that your nine reviewers are of "known competence" provided we understand the term to be restricted to what is commonly known. There is no justification for assuming that any man's competence encompasses the unknown. Otherwise every "expert" may be expected to solve any problem that represents a good jump beyond the state of the art which, judging from the plethora of unsolved problems, has not yet come to pass. My report concerns a little explored area — one-way light propagation — for which there is a paucity of experts of known competence. This applies to the nine referees: although one stated the report was excellent, one offered criticism which I answered and the others would not or could not comprehend my contribution. The last attempt was a travesty.

It is precisely in such cases — the revolutionary quantum jump — that significant progress often results as opposed to the smaller evolutionary advances contained in most reports. Both types are essential to scientific progress. However the review system is primarily suited for the latter; it fails miserably for the former and history bears this out. This is my whole point. The inadequate review of my paper attests to a glaring deficiency in the referee system. A thousand more like rejections would still not justify your decision.

Something must be done for those few cases that probe uncharted territory. Man has too many urgent unsolved problems and too little time to continue to indiscriminately squander and squash creative efforts. This is not a matter for authors or referees to resolve — it is clearly a matter of editorial policy and practice; it represents a timely challenge of lasting potential benefit if it can be done.

I can only repeat my offer — as an editorial experiment I am willing to participate in a test for the concept of closed-loop review for the type of dispute represented by my paper. Whether or not it results in acceptance, the experience is bound to be of future utility to you and other editors if properly done.

Sincerely,
Martin Ruderfer.

Appendix S

26 April 1977

Dear Dr Ruderfer,

Due to your actions, the time and effort spent on your paper has been quite large and, although our nine reviewers may be wrong, we have decided to terminate our consideration of this paper.

There are other journals for such papers and, if your paper does have merit, it will be published somewhere.

Yours truly,
Editorial Staff.

Ára: 900.- Ft + ÁFA