THE BEHAVIORAL AND BRAIN SCIENCES (1982) 5, 187-255
Printed in the United States of America

Peer-review practices of psychological journals: The fate of published articles, submitted again

Douglas P. Peters*
Department of Psychology, University of North Dakota, Grand Forks, N.D. 58202

Stephen J. Ceci
Department of Human Development and Family Studies, Cornell University, Ithaca, N.Y. 14853

Abstract: A growing interest in and concern about the adequacy and fairness of modern peer-review practices in publication and funding are apparent across a wide range of scientific disciplines. Although questions about reliability, accountability, reviewer bias, and competence have been raised, there has been very little direct research on these variables.

The present investigation was an attempt to study the peer-review process directly, in the natural setting of actual journal referee evaluations of submitted manuscripts. As test materials we selected 12 already published research articles by investigators from prestigious and highly productive American psychology departments, one article from each of 12 highly regarded and widely read American psychology journals with high rejection rates (80%) and nonblind refereeing practices.

With fictitious names and institutions substituted for the original ones (e.g., Tri-Valley Center for Human Potential), the altered manuscripts were formally resubmitted to the journals that had originally refereed and published them 18 to 32 months earlier. Of the sample of 38 editors and reviewers, only three (8%) detected the resubmissions. This result allowed nine of the 12 articles to continue through the review process to receive an actual evaluation: eight of the nine were rejected. Sixteen of the 18 referees (89%) recommended against publication and the editors concurred. The grounds for rejection were in many cases described as "serious methodological flaws." A number of possible interpretations of these data are reviewed and evaluated.

Keywords: bias; evaluation; journal review system; manuscript review; peer review; publication practices; ratings; refereeing; reliability; science management

Journal articles serve an important function in providing scientists with information about new ideas and discoveries in their areas of interest. Published papers also serve as vehicles for personal advancement, job security, and continued research opportunities. In academic settings the "publication count" is often a factor in determining salary or merit-pay increments, grant funding, promotion, and tenure (Gottfredson 1978; Scott 1974). Getting research published can also have consequences for entire academic departments. Summaries periodically appear in the literature that rank both the overall and the per capita productivity of departments of psychology (e.g., Cox & Catt 1977; Endler, Rushton & Roediger 1978; Roose & Anderson 1970). Such rankings can establish a psychology department's reputation, which can potentially affect the number and quality of graduate students applying for advanced degrees, the awarding of competitive funds, and the pride and self-esteem of individual faculty members.

Although many are undoubtedly content with the peer-review practices employed by modern research journals, a growing number of psychologists have raised important questions about the adequacy of the review system. Moreover, judging from the variety of disciplines represented by those calling for improvements in the review practices of journals, it would appear that criticism of the review process is not limited to one or two areas, but rather extends across many fields of science. (In the social sciences, see Brackbill & Korton 1970; Crane 1967; Gove 1979; McCartney 1973; Revusky 1977; Tobach 1980; Walster & Cleary 1970; in the physical and medical sciences, Cicchetti & Conn 1976; M. D. Gordon 1980; Harnad 1979; Ingelfinger 1974; Jones 1974; McCutchen 1976; Ruderfer 1980; Stumpf 1980; Zuckerman & Merton 1973.)

A major portion of the criticism of the journal review system has concerned the reliability of peer review. Empirical evidence concerning reviewer reliability has, until recently, been rather meager, considering the importance of this topic. Most of the reviewer-reliability literature has been contributed by social scientists, more specifically, by psychologists and sociologists. With a few exceptions (Crandall 1978a; Scarr & Weber 1978), the results of these investigations have not been encouraging. Interrater agreement between the reviewers of a manuscript, measured by a variety of rating scales and statistical analyses, is typically reported as low to moderate, with intraclass correlation coefficients of 0.55 at best (Bowen, Perloff & Jacoby 1972; Cicchetti 1980; Cicchetti

 0140-525X/82/020187-69/$09.00/0

& Eron 1979; Gottfredson 1978; Hendrick 1977; Mahoney 1977; McCartney 1973; McReynolds 1971; Scott 1974). For instance, Scott (1974) reported that the degree of agreement between reviewers of manuscripts submitted to the *Journal of Personality and Social Psychology* was only 0.26, and Watkins (1979), using Kappa (K), the statistic that shows the degree of reviewer agreement remaining after correction for chance, found a near total lack of interrater agreement for reviews of manuscripts submitted to the *Personality and Social Psychology Bulletin*. McCartney (1973) examined 1,000 reviews of 500 manuscripts submitted to the *Sociological Quarterly* and found that one-third of the reviewer pairs for a manuscript were in complete agreement (i.e., both evaluated the manuscripts identically on a 5-point rating scale). Another one-third of the reviewers were in proximal agreement (i.e., they designated adjacent points on the rating scale). In the remaining one-third of the cases, the reviewers disagreed. While McCartney found some comfort in these data, it should be noted that the overall level of rater agreement is low after correcting for chance. Ingelfinger (1974) reported figures that showed that the degree of interreferee agreement for 496 manuscripts reviewed for the *New England Journal of Medicine* was below 0.30. Data of this type can certainly erode satisfaction with and confidence in the peer-review system. It is not unuusal to hear researchers express the belief that *chance* (e.g., reviewer idiosyncracies, or the editor's choice of reviewers) had played a major role in determining the fate of a submitted manuscript.

The possibility of response bias[1] in the peer-review process (e.g., institutional affiliation, paradigm confirmation or theory support, editor-author friendship, old-boy networks) has been another area of concern to scientists trying to publish the results of their research efforts. Although claims of reviewer bias have been made, and case histories revealing review prejudice can be found (e.g., the rejection of Garcia's original taste-aversion data by several leading journals - see Revusky 1977), little in the way of direct experimental testing of bias has actually been undertaken. Several of the published reports do, however, provide some support for the assertion of reviewer bias (Bowen et al., 1972; Cicchetti & Eron 1979; Crane 1967; M. D. Gordon 1980; Merton 1968b; Oromaner 1977; Yotopoulos 1961; Zuckerman 1970; Zuckerman & Merton 1973). In a recent investigation, Gordon (1980) analyzed 2,572 referee reports received over a six-year period from several prestigious physical science journals. Each set of reviews and manuscripts was coded in terms of the institutional identity of its authors and referees. The results revealed that major university referees evaluate papers from major universities significantly more favorably than papers from minor, less prestigious universities. This bias was not found for minor university referees; there was little difference in their evaluation of manuscripts from high- or low-status universities. Minor university authors were more frequently evaluated postively by minor university reviewers than by major university reviewers, while major university authors were more often rated favorably by major university reviewers. In considering these studies one must bear in mind Mahoney's (1977) point that most of the existing reports of reviewer bias involve post hoc correlational analyses, which limits the conclusions one can draw. However, in the rare instances in which the suspected sources of bias have been manipulated experimentally, evidence of reviewer bias has likewise been observed (e.g., Goodstein & Brazis 1970; Mahoney 1977).

In the present study we attempted to examine the issue of reliability and response bias in journal reviews, but instead of using indirect, correlational approaches, we decided there would be value in studying the review process directly, as it occurred in its natural setting. Recently published articles from mainstream psychology (research) journals were resubmitted to the journals that had originally published them. By adopting this procedure we hoped to be able (1) to assess the reviewers' familiarity with the author's field (which is often presupposed, but has not been tested), (2) to provide an ecologically valid study of the journal review system, (3) to examine reviewer reliability, and (4) to study response bias among journal reviewers.

Method

Characteristics of journals selected. Thirteen psychology journals publishing research articles were originally selected as the sources for the previously published reports. However, we learned during the study that one of the journal's publication criteria had changed with its new editor, and therefore we did not include the article taken from this journal in the analyses (Journal M; see Table 3). Of the 12 journals studied, only two pairs overlapped in their respective specialty areas. Thus, our sample of psychological journals was fairly broad, with coverage including 10 distinct areas each of which has separate divisional representation in the American Psychological Association (APA). The overall rejection rates for manuscripts submitted to these 12 journals was near 80% at the time the articles we selected were originally accepted for publication (Markle & Rinn 1977). These journals were also considered solid, mainstream purveyors of research by those working in the area, and are among the most prestigious journals in psychology in terms of where psychologists want to publish and where they expect important results to appear (Endler et al. 1978; Koulack & Keselman 1975; White & White 1977). An analysis of the overall "impact" these journals had on the field of psychology (i.e., the mean citation rate of a journal's article with respect to citations by psychology journals; see Garfield 1979a) revealed that 10 of our journals were ranked in the top 20 for impact out of a list of 77 source journals of psychology. Furthermore, the average annual citation frequency per article per journal in the *Journal Citation Report* (Garfield 1979b) for our 12 journals (one and two years after publication) was 1.15, which places these journals in the 75th percentile of all psychology journals listed in Garfield (1979a). All the journals we selected employ a standard practice of nonblind reviewing; approximately half are published by the national organization, APA, and half are not. In addition, our sample (n = 12) represented approximately 70% of all prestigious ("prestige" criteria are presented below), nonblind psychology journals.

Procedure. One article was randomly selected from each of the 12 journals. The only constraint on the selection was that the article had to conform to our "prestige" criteria; that is, at least one of the original authors of each article had to have been affiliated with an institution with a high-ranking department of psychology in terms of prestige ratings, productivity, and faculty citations.[2] The articles chosen in our sample had the following characteristics: (a) With one exception (Journal H, see Table 3), the original author(s) included someone from 1 of the 10 most prestigious departments of psychology in the United States (Roose & Anderson 1970). (b) Authors were selected from the psychology departments with the 30 highest productivity rankings; 75% of the authors were from the top 10 departments for their particular research area (Cox & Catt 1977). (c) Each article had at least one author from the top 25 institutions in terms of citations of faculty research (White & White 1977), with 50% in the first six (Endler, Rushton & Roediger 1978). (d) The articles were selected from papers published between 18 and 32 months prior to resubmission for this study. (e) The mean annual citation count in the *Social Science Citation Index* for our 12 articles was 1.5 for one as well as for two years following publication. Since, as mentioned, the average number of citations per year for articles in these journals during this period was 1.15, it seems that the reports we selected were above average in quality, using a citation count measure, for their respective journals. (The mean number of citations for all social science articles one decade following publication is 1.4; Garfield 1979b.)

Before we resubmitted the article as if for the first time, several alterations were made. First, the names (but not the sex) and the institutional affiliations of the original authors were changed to fictitious ones without meaning or status in psychology; for example, Dr. Wade M. Johnston (a fictitious name) at the Tri-Valley Center for Human Potential (a fictitious institution).[3] The titles of the original articles, the abstracts, and the beginning paragraphs of the introduction were slightly altered. It was hoped that these few changes would sufficiently disguise the resubmissions in the event an editor or reviewer made use of some mechanical filing system that would result in automatic detection (e.g., title, key words, or abstract files). The alterations to the original articles were always minimal and purely cosmetic (e.g., changing word or sentence order, substituting synonyms for nontechnical words, and the like). The meaning of the original articles' titles, abstracts, and initial paragraphs was never altered. The remainder of the introduction and the entire method, results, and discussion sections were typed exactly as they had appeared in print. To further guard against superficial detection, the format for presenting data was occasionally altered by converting graphs to tables and vice versa.

Each manuscript was prepared in accordance with the corresponding journal's instructions to contributors, and then sent to the journal that had originally published it with a cover letter requesting that it be considered for publication. Editors and reviewers had no advance knowledge of our efforts. They were informed about the nature of the project either when they detected that the manuscript was a resubmission during the review process, or when the study was completed. Since these events occurred at different times for the various journals, we asked all participants to keep the existence of our study confidential until the entire project was finished (i.e., when there were no more resubmitted manuscripts being reviewed).

Results

Reviewer reliability. Nine of the 12 manuscripts were not detected (by editors or reviewers) as having been previously published. Since these articles had recently (within two to three years) received a positive evaluation that had resulted in their original publication, one might perhaps have expected a second review from the same journal to yield similar results, that is, acknowledgment of scholarship and recommendation for publication. However, as can be seen in Table 1, this was the exception rather than the rule. Eight of the nine articles were rejected. Only 13% (4/30) [12% (3/26): figures corrected in proof - see Table 1, *ed.*] of the editors and reviewers combined recommended publication in their journal. We should add that every editor or associate editor included in this sample indicated that he had examined the manuscript and that he concurred with the reviewers' recommendations.[4]

When we examined only the evaluations of the reviewers (n = 18) - which might be more meaningful, since the editors' decisions are not independent of the referees' evaluation - we discovered that an even smaller portion, 11% (2/18), found the papers acceptable for publication. Reviewers for seven of the journals (A-G) made clear, unequivocal recommendations against publishing the manuscripts (e.g., Journal A: "There are too many problems associated with this manuscript to recommend its acceptance for ——.").

Two referees for the eighth journal (H) did not make any explicit statements regarding acceptability for publi-

Table 1. *Evaluations of undetected resubmissions*

	Reviewers only		Editors and reviewers	
Journal	Reject	Accept	Reject	Accept
A	2	0	3	0
B	3	0	5[b]	0
C	1[a]	0	1	0
D	2	0	2	0
E	2	0	4	0
F	2	0	3	0
G	2	0	4[c]	0
H	2	0	4[d]	0
I	0	2	0	4[g]
	16 (89%)	2 (11%)	26[e] (87%)[f]	4[h] (13%)[i]

[a] The only reviewer of this article was the editor, who has therefore been included in this section.

Editorial Note: The figures with superscripts *b*-*i* are the ones the commentators saw. They were subsequently corrected by the authors as follows: *b* = 4, *c* = 3, *d* = 3, *e* = 23, *f* = (88%), *g* = 3, *h* = 3, *i* = (12%).

cation, but their reviews were exclusively negative, perhaps best characterized as lists of errors, weaknesses, and shortcomings. We Xeroxed copies of these two indecisive reviews, excluding journal identification, and asked six psychologists (all of whom serve as editorial consultants or editors) how they would rate the manuscript in question, on the basis of these referees' comments alone. A 5-point scale (McCartney 1973) was used, consisting of (1) major contribution: profound, theoretically important, very well conceived and executed, accept without question; (2) warrants publication: solid, sound contribution, accept with minor revisions; (3) sufficiently sound and important to justify publication if space permits; (4) poorly written - limited value, may be publishable if certain features are improved or extended, requires major revision; (5) insufficiently sound or important to warrant publication. The mean ratings for the two reviews were 4.8 and 5.0. Since these figures support our own interpretation (as well as the editor's) that a clear rejection message was implied, we feel justified in classifying these reviews as "rejections."

Critical comments made by reviewers. Perhaps the most serious objections that reviewers had about the manuscripts were directed toward the studies' designs and statistical analyses. Several referees detected methodological flaws in the papers we resubmitted. For example, Journal B: "A serious problem is the range of difficulty of —— material within groups. No account of this range is given and no control of its possible effects is offered. Similarly, the comparability of material across groups is unknown"; Journal D: "Separate analyses of each test, on the assumption that groups are equivalent and then inferring gains, can be misleading... especially when analyses can be made directly"; Journal E: "I do not think that this paper is suitable for publication; further experimental work would be required"; Journal H: "It is not clear what the results of this study demonstrate, partly because the method and procedures are not described in adequate detail, but mainly because of several methodological defects in the design of the study"; Journal G: "Some other problems were... use of ANOVA for post-tests... loss of 3 children from Exp. 1 to Exp. 2."

Aside from the few cosmetic changes mentioned earlier, the manuscripts we submitted were identical to the original versions that had appeared in print. Several reviewers of the resubmitted papers criticized the authors' writing and communicative ability. For example, Journal G: "Tighten up the theoretical orientation in the introduction. It seems loose and filled with overgeneralizations (or, at least, undocumented conclusions) at the present"; Journal E: "Also, I don't know what it means to say that players had the ability to get different numbers of markers to the goal, since the game as described on page five was such that each player had only one marker. It is all very confusing... I think the entire presentation of the results needs to be planned more carefully and reorganized"; Journal F: "Apparently, this is intended to be a summary. However, the style of writing leaves much to be desired in terms of communicating to the reader. It requires the reader to go from a positive result, to a double negative, to a qualification, to a negation of the results, and finally, to a couple of sentences (next to last in that paragraph) whose interpretations are just not clear."

Recognizing previously published research. When we resubmitted articles to the journals that had originally published them, how many were detected or recognized? The answer can be found in Table 2. The results are examined by (a) including all individuals representing each journal (editor, associate editors, and reviewers) who had some contact with the submitted manuscript, and thus could have detected the deception, and (b) including only the reviewers, since one could reasonably argue that it would be unfair to expect editors to be aware of the literature in every field covered by their journal. The results show that it makes very little difference which group is studied since the findings are essentially identical. To our surprise, the overwhelming majority of editors and reviewers (92%, or 35/38 of the editors and reviewers combined and 87% or 20/23 of the reviewers) failed to recognize the manuscripts as articles that had appeared in the mainstream literature appropriate for that research area during the past 18 to 32 months. It should be noted that only two journals (C and I) had changed editors during this period.

Changes in rejection rates and publication criteria. Because changes in a journal's publication criteria or rejection rates could explain why the articles we resubmitted

Table 2. *Detection of resubmissions*

	Editors and reviewers		Reviewers only	
Journal	Detected	Failed to detect	Detected	Failed to detect
A	0	3	0	2
B	0	5	0	3
C	0	1	0	1
D	0	2	0	2
E	0	4	0	2
F	0	3	0	2
G	0	4	0	2
H	0	4	0	2
I	0	4	0	2
J	1	2	1	1
K	1	2	1	1
L[a]	1	1	1	0
	3 (8%)	35 (92%)	3 (13%)	20 (87%)

Note: A thirteenth journal (M; see Table 3) has been excluded from this analysis. The manuscript sent to the journal was rejected on the basis of not being appropriate. Following debriefing the editor informed us of a recent policy change that precluded publishing the type of article we had submitted.

[a] We attempted to find out whether the editor, an associate editor, or a reviewer had recognized the manuscript. Unfortunately, the editor refused to disclose that information. We are therefore estimating that at least two individuals had seen the manuscript for this journal, since two names appeared in the correspondence we received.

Table 3. *Stability of rejection rates*

Journal	Original rejection rate minus rejection rate at time of manuscript's resubmission
A	16%
B	7%
C	−2%
D	−15%
E	−16%
F	0%
G	15%
H	3%
I	5%
J	5%
K	0%
L	7%
M[a]	−2%

Note: Information on rejection rates for APA journals was obtained by comparing figures listed in the June 1980 archival issue of the *American Psychologist* (American Psychological Association 1980) with those listed in earlier archival issues. Figures for non-APA journals were derived by comparing figures listed by Markle & Rinn (1977) with current figures reported by journal editors (personal communications).

Journals A-I failed to recognize the manuscripts and proceeded to review them.

[a] Excluded from analysis because of change in journal's publication criteria.

were not accepted the second time around, we examined the journals and other information sources to find out what changes, if any, had occurred. Our check of editorial statements and topic coverage in each journal revealed no shifts in research emphasis (with the one explicit exception of M, which was deleted from all analyses, as mentioned earlier). Table 3 displays the net change in rejection rates for each journal from the time the original article appeared to the resubmission date for this study. It is clear that the overall rejection figure for the nine journals that reviewed the resubmitted articles had remained quite constant (and high, averaging around 80%) across the two review periods.

Discussion

The results show that out of the nine journals in question (A - H) only one repeated its earlier positive evaluation of the submitted paper and accepted it for publication. Since the articles we selected were from recognized and prestigious research journals and had originally passed a review system averaging 80% rejections, how does one explain their failure to be accepted a second time by the same journal? There are a number of hypotheses that should be considered in trying to answer this question.

A change in journal policy concerning the type of material considered appropriate or a substantial increase (of, say, 25-30%) in rejection rates, making it appreciably more difficult to get a paper accepted, might explain why the resubmitted articles were rejected. However, as just mentioned, we could find no evidence to suggest that either of these factors had in fact prevailed.

One might argue that the reviewers for the resubmitted papers did not remember the specific design or details of the original study, but could recall the points having been made experimentally. Since these articles had appeared in print some time after the work had been completed, it is possible that their results had already been incorporated into the reviewers' implicit sense of what was already known in the area. Thus, the reviewers might have felt that the work was redundant with what they could recall of the literature and rejected the manuscripts on that basis. We would agree with this account if the referees had indicated as much, and had raised this objection in their criticisms, writing, for example, "This point is uninteresting and old," or "This work adds nothing new to the field," or "Similar findings have been reported by previous investigators." No such statements, or anything resembling them, were ever offered. As stated before, the manuscripts were rejected primarily for reasons of methodology and statistical treatment, not because reviewers judged that the work was not new.

Another hypothesis that might account for our results concerns regression effects. Since we selected only manuscripts that had been published, the only possible direction for change following a resubmission would be downward, that is, less than 100% acceptance, with some accepted articles being rejected. While regression to the mean was not controlled (rejected manuscripts were not resubmitted), it is still possible to ask how much regression would be probable. Even if the original acceptance figure for the articles is taken to be a minimal 67% (i.e., with the editor and at least one of two reviewers recommending publication), and even if the correlation between pairs of reviewers were zero (the most extreme possibility), the expected regression could only be to the base rate, that is, the percentage of reviewers initially recommending acceptance - and this is obviously more than the 11% we see in Table 1. In other words, one would have to hypothesize a large *negative* reliability for regression to account completely for the considerable shift in reviewers' judgments. We think that there are more reasonable ways to explain our findings.

Another way of addressing the regression issue would be to evaluate how probable the observed outcome was (eight out of nine manuscripts rejected) given varying levels of initial probability of being recommended for publication. In doing this one would have to conclude that *published* manuscripts in these journals have no more than a 43% conditional probability of acceptance because if the probable acceptance rate for any eventually published manuscript had been higher, the likelihood of our observed outcome of eight out of nine rejections would be less than 5%.

$$\text{probability} = 9(.43)^1(.57)^8 + 1(.43)^0(.57)^9 = .046 \text{ for 43\% initial acceptance rate}$$

In practice, the more likely acceptance rate for "quality" manuscripts that one would have to hypothesize, given our data, would be much less than 43%, as this repre-

sents the extreme of the confidence interval for the binomial distribution.[5]

Since the most frequently mentioned grounds for rejecting the manuscripts were "serious methodological flaws," one might want to know whether these perceived flaws had been revealed by methodological or theoretical innovations that had appeared since the articles were first published. This seems not to have been the case. The criticisms were "old" and basic notions having to do with such matters as confounding, nonrandomization, use of ANOVA with dependent data, subject mortality, and the like. These points were considered flaws a year ago and will probably be considered flaws ten years from now.

Since we could detect no major shift in the qualifications or criteria used by the journals to select reviewers for the two periods, we concluded that perhaps one of the following two possibilities had prevailed:

Somehow, by chance, the initial reviewers were less competent than the average at that time, or less competent than the later reviewers. This possibility cannot be given too much credence, though, on purely statistical grounds. One would hardly expect such sizable systematic variation in reviewer competence to occur by chance from one to the other review period.

The second possibility is that systematic bias was operating to produce the discrepant reviews. The most obvious candidates as sources of bias in this case would be the authors' status and institutional affiliation. As mentioned earlier, these two variables have been cited (e.g., by Bowen et al. 1972; Cole & Cole 1972; Crane 1967; M. D. Gordon 1980; Merton 1968b; Tobach 1980; Zuckerman & Merton 1973) as sources of bias that can influence journal reviews in science. The authors of the articles used in our study were all from recognized, productive, and prestigious institutions with highly ranked psychology departments. For both sets of reviews, the authors' names and institutional affiliations were identified, except that fictitious names and institutions were substituted on the resubmitted manuscripts. The predominantly negative evaluations of the resubmissions may reflect some form of response bias in favor of the original authors as a function of their association with prestigious institutions. These individuals may have received a less critical, more benign evaluation than did our unknown authors from "no-name" institutions.

On the basis of his recent study of reviewer bias in 619 physical sciences articles, M. D. Gordon (1980, p. 275) has concluded:

> It can therefore be argued that biases systematically operate within refereeing systems in such a way as to give advantage to those elements of a research community which supply the largest proportion of the referees used by editors of its journals. The papers of such authors may on occasion be less demandingly evaluated than those of authors outside the group. Hence, access to publication may sometimes be easier for them.

The mechanism underlying this form of bias may be something quite similar to Rosenthal's experimenter expectancy effect (see Rosenthal 1966; Rosenthal & Rubin 1978). Journal reviewers may expect manuscripts (and research) from persons at prestigious institutions to be superior in overall quality to those from individuals working in less distinguished settings; as a consequence, giving more favorable evaluations to high-status individuals may serve as a self-fulfilling prophecy. (One might speculate that if reviewers were less rigorous toward prestige institutions and individuals, the publishing and publicizing of lower-quality work would have a negative feedback effect. Whether this would result in better papers or fewer submissions is not clear.) Another possibility is that when referees examine a manuscript submitted by researchers working at highly respected institutions, they may be more sensitive to making "false negative" evaluations, that is, rejecting papers of quality, whereas the major concern in reviewing papers of individuals from lesser known institutions may be that of avoiding "false positive" errors, that is, accepting flawed work. If reviewers base their evaluative criteria on experience or belief to the effect that papers of quality come mainly from prestigious individuals and institutions, then this could be viewed, in signal-detection terms, as criterion or response bias (as suggested in note 1). The cutoff point for publication acceptability may be lowered or raised as a function of status variables associated with an author. In fact, an institution's or an author's prestige could in principle prove to be a *valid* predictor – which would of course argue for the validity and continued exercise of such response biases. Unfortunately, little has been done to examine the validity of such a decision-making strategy empirically. Does this response bias maximize the ratio of "hits to misses" for a journal? At present this is just a matter of conjecture.

We do of course realize that a complete test of the bias hypothesis would call for resubmitting previously rejected articles with prestigious institutional affiliations. Such a procedure would obviously be much more complex and delicate, and unfortunately it was not possible for us to undertake such a project at the time we began.

The near perfect reviewer agreement regarding the unacceptability of the resubmitted manuscripts, coupled with the presumably near perfect agreement among the original reviewers in favor of publishing, provide additional convergent support for the response bias hypothesis. One might expect the initial manuscripts to elicit high positive consensus if the reviewers were impressed by or had high expectations for those with a Harvard or Stanford affiliation. But even the second time around, there are reasons – albeit different ones – to expect the reviewer consensus we found in our study. In the second case the reviewers may have been in agreement that there were serious *flaws* in the articles – perhaps the stereotypic muddled thinking of authors associated with a Tri-Valley Institute of Growth and Understanding. If this type of bias is actually operating to produce differential reviews and publication decisions, then one might predict that the highest levels of reviewer consensus should be found in journals with nonblind review, since an author's identity and affiliation are much more visible. While we do not have enough data to test this idea, the limited data on interrater agreement in the literature are in the predicted direction. Most reports of reviewer reliability showing good interreferee agreement (e.g., Crandall 1978a) have involved nonblind journals (Scarr & Weber 1978 is an exception). Viewed in this context, our finding of very high reviewer agreement with the resubmitted manu-

scripts is not so surprising; and considering the fate of the original manuscripts compared to the resubmissions, the response bias hypothesis seems to be a fair explanation for two sets of reviews showing near perfect, but *opposite* agreement.[6]

We did attempt to obtain the reviews from the *original* submissions to the nine journals in question. Unfortunately, we encountered some resistance to this idea, and the early lack of editorial cooperation indicated that a complete sample was not going to be possible. We therefore had to abandon this effort. We should add that not all editors were uncooperative: There were a few who were very gracious in supplying us with the requested information.

Although we had only three of the original reviews to examine and compare with our own set, we still contend that the reviews of the resubmissions represented a sizable and significant shift. Of the 30 new reviews 26 were clearly rejections [figures corrected in proof - see Table 1, *ed.*], and if we accept the estimates given in the Scarr and Weber (1978) and Scott (1974) papers, then we would assume that the 9 undetected resubmitted manuscripts that generated these 30 reviews had originally received 18 favorable reviews; that is, at a minimum, the editor and one of the two referees were favorable. If we make these conservative figures (one reviewer recommending acceptance, the other rejection, instead of the more liberal unanimous acceptance) the expected frequencies for a chi-square test of the proposition that the resubmission reviews were not different from the original reviews, then the null hypothesis can be rejected ($\chi^2(1) = 18.9$ [21.8, corrected in proof - see Table 1, *ed.*], $p < .001$). From this perspective it would seem that the outcome we observed was quite improbable.

Failing to recognize published research. Often the cover letter an editor sends to a contributor following a review will contain some statement extolling the referees' knowledge and expertise. The findings in our study were no exception, as, for example, Journal H: "consultants who are very knowledgeable about theory and research in the area of concern to you"; Journal G: "both consultants who are experts in this area." Remarkably, though, almost 90% of the reviewers failed to recognize the resubmitted articles in our study. At the outset we thought that the major obstacle in collecting "real" journal reviews for our resubmissions would be the large number of detections by either the editors or the reviewers. This obviously did not happen - why not? No one could claim that the articles we selected were trivial reports that had appeared in minor, seldom read journals. We took articles from prestigious and widely circulated psychology journals having high "impact" ratings (Garfield 1979b). The articles themselves were above average in citations for the journals they appeared in, as well as for all social science articles in the first two years after publication (Garfield 1979b). Furthermore, these reports were resubmitted for review to the very journals and editors (excepting two cases, C and I) that had originally published them, and they had appeared recently (two to three years ago), not five or ten years in the past. Nor were these articles published *too* recently. Someone familiar with the literature would have had sufficient time to read and discuss them. Perhaps because of a large number of intervening reports, the articles had simply been forgotten by the reviewers (n = 20). It is also possible that those making the second set of reviews had nothing to remember. They may never have seen the original articles or heard them referred to or discussed by others working in the field. Under these circumstances, the failure to recognize the resubmissions, rather than being surprising, would be predictable. Although these explanations are certainly plausible, we suspect that few editors or contributors will find them encouraging, especially since the reviewers' reputed knowledge of an author's field can be such a sensitive issue in cases of negative reviews. Moreover, as mentioned, the articles we selected were above average in number of citations over the two-year period following publication.

Author-reviewer accountability. In addition to reliability and response bias, accountability has been an issue of concern in debates on the journal review system. Given that researchers in the behavioral and social sciences must often face rejection rates of 70% or higher from top journals in their field, and considering the large number of submitted manuscripts that editors and reviewers must evaluate each year, it is not surprising that much has been said, by both defenders and detractors of the journal review system, about the need for accountability. From one perspective, editors and referees, in light of the heavy time commitment called for by journal reviewing, might complain about the "bane of refereeing, the journal shopper" or "the repeated offender. Born of desperation, insensitivity, or stubbornness, a significant portion of a referee's efforts are repeat business - a succession of bad papers from the same individual" (Webb 1979, p. 60). It has been suggested that a central computer bank be created consisting of the titles and authors of rejected papers. "These would be categorized into two groups: (a) rejected, or (b) acceptable but not sufficient or appropriate for the journal. Minimally, these would be distributed to journal editors or ideally they would be placed on accessible tapes" (Webb 1979, p. 60). The assumption underlying this point of view is that the "journal shopper" is someone who basically knows that his manuscript is of questionable worth (perhaps a pilot study only appropriate for his private archives). The journal shopper, so this reasoning goes, hopes that with a lucky break, such as lenient reviewers, the manuscript might be accepted. Thus, repeated submissions occur until either the author "lucks out" (as perhaps most do, given the large number of existing publication outlets - see Garvey & Griffith 1971) or the list of journals is exhausted. The existence of a central computer bank supplying names of rejected authors to prospective editors is, of course, intended to deter "journal shopping." Certainly very few would want the potentially widespread notoriety of being identified as a "publication loser." The opportunity for a negative halo to develop should be obvious.

On the other side of this accountability issue one finds authors who believe that editors and referees should be more accountable and responsible for the quality of reviews they make in the course of reaching publication decisions (e.g., Colman 1979; Jones 1974; McCutchen 1976; Stumpf 1980). It is not unusual to hear authors of

rejected manuscripts argue that their papers were rejected, not because they did not have sufficient merit, but because of poor reviewer reliability (obviously most likely to be claimed in cases of split reviews), reviewer incompetence, or reviewer bias. One recent proposal that addressed the question of reviewer accountability suggested the creation of an "author review" of journal reviewers.

> The journal editor would send to the author, along with the letter of decision and reviews, a postcard questionnaire that would request the author to evaluate each review. I suggest that three dimensions are necessary: Fairness, carefulness, and constructiveness. There should also be a place on the card for comments. The editor would file the returned postcards under the reviewers' names, rating the editorial decision and final disposition of the manuscript. At intervals, perhaps once a year, the editor could examine these questionnaires. If a particular reviewer received repeated complaints, he or she could be terminated as a reviewer or could receive admonishment from the editor. (Hall 1979, p. 798)

It is argued that with such a reviewer evaluation procedure editors would have a basis for weeding out referees who were not current in their knowledge of the literature, consistently rejected articles on the basis of unreasonable or idiosyncratic criteria, or typically favored authors who were from their own institutions or shared similar research traditions. Journal reviewers faced with a record of accumulating author complaints might, one would hope, become more conscientious in their manuscript evaluations.

In our study none of the 18 reviewers was identified to us either by the editors or on the manuscript reviews sent to the authors. We share the opinion of Hall (1979), Stumpf (1980), and others that anonymous peer reviews may be more costly than beneficial. A system that could allow a reviewer to say unreasonable, insulting, irrelevant, and misinformed things about you and your work without being accountable hardly seems equitable. To some degree the reviewer is indeed accountable – to the editor – but the potential for abuse is still too great to be ignored (see Ruderfer 1980, for an excellent example of this problem).

Suggestions for improving the journal review system. However disquieting one finds evidence of reviewer bias, incompetence, or unreliability, it should not be ignored or dismissed as trivial. Scientists concerned about the quality of the review system should be encouraged to study this important area. We also hope that those choosing to do this will be guided by recent calls for hypothesis formation and discovery in naturalistic contexts (e.g., Bronfenbrenner 1977; Herrnstein 1977; McCall 1977; McGuire 1973; Neisser 1976; cf. Gibbs 1979). However, those wishing to study the review process directly by using a procedure similar to the one employed in our investigation should be alerted to the possible costs. Be prepared for a lengthy time commitment (nearly two years for this study), practical difficulties (e.g., obtaining published materials or securing the cooperation of those in the journal hierarchy), and objections to your research approach.

In what is perhaps a sign of growing concern about the adequacy of our present journal review system, a number of different sources (editors and contributors) have recently offered suggestions for improvement (e.g., Crandall 1978a; Hall 1979; Harnad 1979; Hendrick 1977; McCartney 1973; Scott 1974; Stumpf 1980; Wolff 1973). Several writers have stressed the need to adopt a standard rating form in which an explicit set of evaluation criteria is listed. Some have called for the training of referees to increase the quality and reliability of their reviews. This might involve giving potential referees samples of actual editorial evaluations and subsequent publication decisions. Editors would have the responsibility of explaining the relationship between specified attributes of submitted manuscripts and their publishability. A collection of reviews and articles on which experts could agree (e.g., high quality, very publishable; low quality, reject) might be quite useful as a training or screening device for referees.

As mentioned above, another recommendation has been for a more accountable system of peer review, the basic idea being to establish some form of referee review, with referees formally evaluated by judges, authors, and editors. Systematically monitoring the quality of referee reports should have a corrective effect on journal review practices.

A further extension of author involvement and a move toward more openness and accountability in the review process can be seen in the recent suggestions for (and implementation of) an "open peer commentary" or "open review" system (Harnad 1979) in this journal (and its model, *Current Anthropology*). The basic idea is to complement the conventional closed peer-review system by giving the authors of *accepted* (refereed) articles the opportunity to respond openly to criticism. With the article, commentaries (from first-round referees and others), and the author's formal response published together in their entirety, readers of the journal can have a chance to examine and appraise this process of "creative disagreement" and form their own opinions as to the merit of an individual's work.

If institutional affiliation or professional status can in fact bias peer review – and this bias proves to have no validity, or negative validity – then one possible solution to this problem (as several critics have recommended) would be to establish blind reviews as standard journal policy. We realize that this might not be totally effective in eliminating the problem (authors might, for example, be identified on the basis of several personal citations in the manuscript) and that blind reviewing may present additional problems (see American Psychological Association 1972; Rosenblatt & Kirk 1980), but it would certainly help minimize the influence of such biasing variables if done conscientiously. It is encouraging to note that there has been an increase in the number of psychology journals using blind reviews over the past decade (although the motive behind this move may be largely public relations, as was suggested by an APA editor involved in our study).

Given the professional importance placed on pub-

lished research, and considering that science policy-makers and funding agencies make practical decisions affecting the nature and future direction of scientific research on the basis of what gets reported in our journals, it is essential that we acquire a better understanding than we presently have of our journal review system. For years scientists have assumed that the review process is basically objective and reliable. Is it? Unfortunately, the peer-review process has not received the experimental attention given other research topics - most of which have considerably less significance for and impact on scientists. We are sure that there are those who would not wish to see the somewhat delicate machinery of the review system tampered with by a wave of research projects. However, unless we subject the review process, and suggestions for improving it, to experimental analysis to learn more about the variables that do influence peer review, we are left with little to defend it other than faith.

ACKNOWLEDGMENTS

The authors wish to thank Stevan Harnad, Jason Millman, David Palermo, Paul Ross, William Wilson, and several *anonymous* reviewers for helpful comments on earlier drafts of this paper.

NOTES

*Reprint requests should be sent to Douglas Peters, Program in Social Ecology, University of California, Irvine, Calif. 92717.

1. By "response bias," we mean no more than the value-free signal-detection theoretic parameter usually called "*c*" or "η." This represents a referee's criterion or cutoff point in the trade-off in his "pay-off matrix" between "false positives" (overacceptance - analogous to "Type I errors" in decision theory) and "misses" (overrejection - "Type II errors"). The response bias may be based on various predictors (such as author's identity, reputation, institution) whose validity then also becomes an empirical question. The "signal" itself (acceptability) is usually assumed to have a certain independent "detectability," expressed as its distance (d') from noise.

2. The institutional affiliations of the original authors were: Harvard University, Stanford University, University of California at Berkeley, University of California at Los Angeles, University of Illinois at Urbana-Champaign, University of Minnesota at Minneapolis, University of Texas at Austin, University of Wisconsin at Madison, and Yale University.

3. Other fictitious names used were Tri-Valley Institute of Growth & Understanding, Tri-Valley Institute of Human Learning, Northern Plains Center for Human Potential, and Northern Plains Research Station.

4. The following are the eight rejection statements made by the editors. Journal A: "Two consulting editors have examined your paper and I enclose copies of their reviews. I regret to conclude that the paper is not acceptable for publication, and judging from the reviews, I doubt that any revision would alter this decision." Journal B: "The considerable number of problems noted in the three reviews lead to a decision not to publish the paper." Journal C: "Your paper does not fall within priority areas having relevance to ——. Perhaps you might submit this work to some other journal in ——." Journal D: "The two reviews point to a number of serious methodological difficulties that would have to be overcome before a paper of this type could be considered publishable in ——. It is with regret that I must decline your paper." Journal E: "As you can see, the consultants are not enthusiastic about publishing your paper. Unfortunately, my own reading of your paper leads me to share this opinion." Journal F: "Unfortunately they all question its relevance for the —— readers and suggest that it may be more appropriate for a more applied journal like ——. Therefore I have decided to reject the article." Journal G: "Both consultants, who are experts in this area, point out a number of methodological, conceptual, and stylistic problems with your paper. I agree with their judgments that these problems preclude publication of the paper in ——." Journal H: "Two individuals who provided reviews have raised a number of methodological and conceptual issues that militate against accepting the manuscript. Although we will be unable to publish your study in ——, it does seem that you are onto something good here, and I would encourage you to consider conducting another controlled study which addresses some of the reviewers' concerns."

5. In an unpublished manuscript, P. F. Ross (1981) has made a somewhat similar argument. On the basis of his analysis of interrater reliability (used to set validity intervals), citation index (used to determine quality of published articles), and acceptance rates of journals, he determined that the number of Type II errors (i.e., rejection of quality papers) is more than double the number of quality papers actually accepted for publication (and is also more than twice as large as the number of Type I errors). If his analysis is correct (cf. Garvey & Griffith 1971), it points to a large degree of error in peer reviewing.

Few would expect that the select 20% of submitted manuscripts actually accepted by our journals are completely flawless, and that they would without exception be reaccepted on a second review. However, in order to explain our own findings totally in terms of regression, one would have to accept the fact that the system is so flawed that it rejects quality papers far more often than it accepts them. While we do not deny that quality papers do get rejected, we feel that if there were no additional factors to explain our data other than regression, the implied level of Type II error would be unreasonably high.

(Actually, the level of Type II error is likely to be even greater than what we observed in our study, since the original manuscripts, when first accepted for publication, were probably not as polished as the printed versions serving as our resubmitted manuscripts. Thus, if anything, one would expect the resubmissions to have a *higher* acceptance rate than a typical manuscript of publishable quality being submitted for the first time.)

6. It is also possible that institutional prestige produced a criterion bias in the original reviewers toward being overgenerous, elevating marginal papers to the level of "publish if space permits." If this is taken to be the modal threshold for quality as perceived by the initial reviewers, then the later reviewers can be viewed as (validly) correcting this shift. It may be that the least variability among reviewers occurs in the region of outright rejection. This category is indeed reported to be the most reliable one among raters (Cicchetti 1980), and this could explain why our reviewers were in such high agreement - they may have been evaluating manuscripts of low quality despite their respectable citation count. (Now the question becomes: What if this was in fact a representative sample . . . ?)

Open Peer Commentary

Editorial Note *Peters & Ceci's (P & C's) target article gives rise to several methodological and ethical questions, not only about their object of investigation - peer review - but about their study itself. As the editor of this journal I implicitly made a judgment about these questions in accepting P & C's manuscript for publication. Since the questions are repeatedly raised in Commentary, I feel I should make explicit some of the considerations underlying my editorial decision:*

1. Contrary to normal BBS policy, P & C's paper was accepted despite its obvious methodological weaknesses (e.g., very small sample size, highly unbalanced design) because I still judged it to be a provocative basis for an open analysis of peer review and because I recognized that the authors were unlikely to get a second chance to perform or improve upon such a study after informal disclosure (to the editor-subjects involved) and formal disclosure (e.g., Peters & Ceci 1980). The open peer commentary, I assumed (and the reader can now judge for himself), could be counted on, in turn, not to leave any of the P & C study's limitations undisclosed - this is in the self-corrective nature of the BBS Commentary system.

2. It was foreseeable that some would object to P & C's study on the grounds of the deception involved. I happen to find the deception rather innocuous in this case. Perhaps some editors regard themselves as too important, or their time as too valuable, to contribute to empirical efforts to study and improve the peer-review system under natural conditions. It should be reassuring to them, then, to reflect that, just like thefts of great works of art or of great scientists' brains, enterprises such as P & C's are of necessity self-limiting, by virtue of the very act of openly exposing their findings.

It is hoped that this self-reflective open peer commentary on the peer-review system will contribute to strengthening the scientific communication process to which the BBS project is dedicated.

Please note that in the following text all passages that appear between commentaries in this typeface are editorial notes. Ed.

A physics editor comments on Peters and Ceci's peer-review study

Robert K. Adair

Department of Physics, Yale University, New Haven, Conn. 06511

As a physicist, I read the paper of Peters & Ceci (P & C) with a certain wry (and malicious) amusement. The problems of acceptance procedures in psychology journals reported in the paper scarcely exist in physics. After two martinis I would ascribe the very real difference to the obviously superior intelligence, integrity, and the like of physicists vis-à-vis psychologists - which does not contribute to domestic harmony, considering that my wife, Eleanor R. Adair, is an experimental psychologist. However, in reluctant sobriety, I believe that the difference follows not from what innate differences there may be between physicists and psychologists but from the relative simplicity and objectivity of physics and the complexity and subjectivity of many areas of psychology.

I was disappointed that P & C found it impractical to attempt to describe more fully the different areas of psychology they had considered, or to differentiate between the different areas. Psychology is a rubric that covers an astonishing breadth of scholarship and scholars, ranging from very hard (simple and objective) areas, such as the fields of physiological psychology, sensory psychology, and psychophysics, to soft (complex and subjective) areas, such as clinical psychology and social psychology; and I am dubious about lumping such disparate subjects together. Hard and soft (or simple and complex) fields have very different research practices and sociologies. I would expect that there would be appreciable differences between the degree of referee nonrecognition of a resubmitted paper, for example, in physiological psychology compared to social psychology. The referee chosen to review a paper reporting work in a specific, narrow area of physiological psychology should, and usually will, know all of the work in that field very well. He has to. I understand that this may not be the case in social psychology inasmuch as narrow areas cannot be so easily isolated - the effective referee must be broad, and hence, perhaps, necessarily a little shallower in his mastery of the literature.

I will comment on physics journals as a kind of counterpoint to the journals and problems reported by P & C. I believe that there is a consensus to the effect that the *Physical Review* is the most important journal in physics. The *acceptance* rate for the whole journal (consisting of four sections, publishing about 30,000 pages a year) is about 80 percent. The editors consider that their charge is to publish all properly prepared reports of substantial, competently conducted, researches. Although great skill and ingenuity may be required in the conduct of the researches, both the procedures and the results can usually be described in a lucid and objective manner (somewhat as a mathematical proof may be very ingenious but can usually be described very simply). As a consequence of this simplicity, objectivity, and lucidity of physics, there is a high degree of concurrence among physicists - authors and referees - as to what constitutes a paper that is acceptable to a physics journal. And when controversy does arise - as occasionally happens - the editors, with support from the community, consider that the argument should be settled in the intellectual agora by the whole community rather than by a few referees and an editor working in camera.

There is one partial, but important, exception to the rather sanguine picture I have painted of physics journals. One American journal, *Physical Review Letters*, technically a part of the *Physical Review* but separate in editorial personnel and policies, has become something of a prestige journal; some consider it the most prestigious journal in physics. With a nominal acceptance criterion directed toward publishing short papers that are (exceptionally) "novel and newsworthy," the criterion has been transformed operationally to "important" by authors and referees. Only 45 percent of the submitted papers are accepted for publication. The two or more referees chosen to consider the paper generally agree on the comparatively objective question of scientific adequacy, but often they do not agree on the more subjective criterion of importance. When the editors of the journal reported to the community that chance played a large part in the acceptance of papers for the journal, the community, through the American Physical Society, publisher of the journal, supported changes in the acceptance criteria which we all hope will reduce the degree of random acceptance by the journal (Editorial, 1979).

I add brief comments on some points raised in P & C's article: Blind refereeing will not work in physics. We have made simple (unpublished) studies that suggest to us that about 80 percent of the authors of letters submitted to *Physical Review Letters* can be identified by referees competent in the narrow field of the submitted paper. (We use 4,000 *different* referees a year for *Physical Review Letters!*) Also, I sense an implicit assumption in the paper of P & C that journals and their editors should be selecting *papers per se* - hence the interest in blind refereeing. In physics, we editors are interested in the papers only as an exposition of the *research* that is conducted. As active scientists, we consider a result from a scientist who has never before been wrong much more seriously than a similar report from a scientist who has never before been right! Our rather subjective internal checks suggest to us that our referees do not attach much weight to the reputation of an author (or to the institution from which he comes); but science is not democratic, and it is neither unnatural nor wrong that the work of scientists who have achieved eminence through a long record of important and successful research is accepted with fewer reservations than the work of less eminent scientists.

R. K. Adair is editor of Physical Review Letters. *Ed.*

Barriers to scientific contributions: The author's formula

J. Scott Armstrong
The Wharton School, University of Pennsylvania, Philadelphia, Pa. 19104

Recently I completed a review of the empirical research on scientific journals (Armstrong 1982). This review provided evidence for an "author's formula," a set of rules that authors can use to increase the likelihood and speed of acceptance of their manuscripts. Authors should: (1) *not* pick an important problem, (2) *not* challenge existing beliefs, (3) *not* obtain surprising results, (4) *not* use simple methods, (5) *not* provide full disclosure, and (6) *not* write clearly. Peters & Ceci (P & C) are obviously ignorant of the author's formula. In their extension of the Kosinski study (Ross 1979; 1980), they broke most of the rules.

Why, then, is P & C's paper being published? In my search for an explanation, I learned the following from Peters: (a) After a long delay, the paper was rejected by *Science*, with advice that it would be appropriate for the *American Psychologist*. (b) After a long delay, the paper was rejected by the *American Psychologist*. This history illustrates the predictive power of the author's formula. Submission was meanwhile encouraged by the editor of the *Behavioral and Brain Sciences* - a journal specializing in peer interaction on controversial papers - and, after a final round of major revision, the paper was accepted for publication.

In this commentary, I describe how P & C violated many rules in the author's formula. It may be too late to salvage their careers, but the discussion should be instructive to other authors.

Examined an important problem. P & C examined whether the decision of prominent scientists to recommend a paper for publication constitutes evidence of that paper's scientific contribution. This strikes me as an important issue. It passed one test I use for importance: Would I discuss this issue with people outside my field? It has implications for the communication of scientific knowledge. Few researchers have dared to address it. Most of those who have done excellent work on this issue have met difficulties in getting their findings published in high-prestige journals. For example, Goodstein and Brazis (1970), Mahoney and Kimper (1976), and Mahoney (1977) were *not* published in journals with high prestige. Furthermore, Mahoney (1977) was rejected by *Science*.

Challenged existing beliefs. Scientists believe themselves to be competent and fair when they judge scientific contributions. An alternative hypothesis, such as, "The judgment by scientists of scientific contributions is seriously affected by irrelevant factors," is an affront to scientists. P & C reveal themselves to be insensitive to this fact. Their work does not try to make a "positive scientific contribution"; instead, it merely criticizes an existing belief. (P & C have obviously learned little from the well-publicized case of Galileo, namely, that some beliefs should not be examined.) P & C's use of the method of multiple hypotheses to examine existing beliefs was impressive. It was difficult to think of a reasonable alternative to current beliefs that they did not examine. The strategy of multiple hypotheses is unusual in the social sciences. Typically, the route to academic fame involves adopting a single dominant hypothesis and finding supporting evidence (Armstrong 1979; 1980a).

Obtained surprising results. To challenge existing beliefs is folly. To obtain evidence that these beliefs are wrong is sinful. (It is called the Second Sin by Szasz 1973.) P & C are sinners. They found that biases against the author or the author's institution play an important role in judgments of the value of a scientific contribution. These are surprising results.

For the journals used in P & C's study, the probability of an article being accepted was about 20%; the rate of acceptance for the papers resubmitted by P & C, 11%, was not significantly different. This suggests that we, as scientists, are merely engaged in a game of chance. Either P & C are wrong or we are fools!

Another way to interpret the results is to argue that there are identifiable reasons for acceptance or rejection but that these have nothing to do with the scientific contribution made by a paper. This reminds me of Webster's (1964) studies on the employment interview. Some factors did help explain who would be hired, but they had little to do with job performance. Instead, they related to the similarity between interviewer and interviewee. The decisions were usually made in the first five minutes (often in the first 30 seconds), and the interviewers were not aware of their decision-making process, although they claimed to have based it on the potential job performance of the interviewee. Perhaps the employment situation is analogous to the P & C study; that is, perceived similarity (e.g., in institutional background) is a measure of the author's competence. In fact, none of the reviewers had a background similar to the author's, since the rejected papers were from fictitious institutions.

I agree with P & C that bias against unknown authors and institutions provides the best explanation of their results. In fact, the evidence may be even stronger than they suggest. Noting that one of the manuscripts (manuscript "I" - see P & C's tables) was unanimously accepted (two referees plus two editors) while eight papers were unanimously rejected (16 referees plus 10 editors), I hypothesized that the author and institution for this sole accepted article may have *sounded* more authentic. (Hawkins, Ritter & Walter 1973 demonstrated that fictitious journals had high status among academics if they sounded theoretical.) Peters & Ceci (personal communication) provided some confirming evidence. The fictitious author of manuscript "I" had a common male name, Wade Johnston. This paper was originally submitted as having been from the University of North Dakota. (In a convenience sample I conducted at the University of Pennsylvania's Wharton School, eight out of nine faculty members selected the University of North Dakota as the most prestigious from the list of six institutions used in the P & C study.) Howerer, one other undetected paper had likewise been initially submitted with the institution identified as the University of North Dakota, and that paper was rejected.

You might not agree that P & C's results are surprising. This would be understandable. The experiment by Slovic and Fischhoff (1977) indicated that few scientists are surprised by the results of scientific studies, no matter what the results.

To determine the degree of "surprise" associated with P & C's results, I conducted a survey of five full professors at Drexel University and seventeen at the University of Pennsylvania. Most were from the social sciences, primarily management, but five were professors of physical sciences. (Four research assistants and I attempted to deliver personally a self-administered questionnaire to professors in these schools over a five-day period. Six questionnaires were administered by phone. Although there were few refusals, few professors were available; either they were out or occupied.) Fourteen of the 22 respondents (64%) reported that they were either editors or associate editors for one of the most prestigious scientific journals in their field.

The questionnaire contained a brief description of the P & C study. It was presented as a "proposed study" and the respondents were asked to predict what results would be obtained if the study were conducted with psychology journals. None of the respondents had previously heard about the P & C study.

The results, summarized in Table 1, show sizeable differences between the predicted and actual outcomes. The respondents expected the journals to detect far more papers as having been previously published. Furthermore, they greatly underestimated the percentage of reviewed papers that would be rejected and greatly overestimated the percentage of the rejected papers for which the grounds for rejection would be

Table 1 (Armstrong). *Predicted versus actual outcomes of P & C study*

	Average predicted % from survey (n = 22)	Actual % (P & C)
Papers detected	66	25
Papers rejected (out of total number reviewed)	42	89
Papers rejected as adding "nothing new" (out of total number rejected)	46	0

that they added nothing new. If the respondents had assumed that they knew *nothing*, they could have minimized their maximum possible error for each question by predicting 50%. This prediction would have produced a smaller average error (38% rather than the 45%). Applying the survey outcome to the P & C study, one finds that only five papers would have been expected to go undetected through the review process and two to have been rejected. Only *one* paper would have been rejected for reasons other than that it "added nothing new" (vs. the eight such papers in the P & C study). In short, P & C's results were very surprising. The surprise was much higher among the social scientists than among the physical scientists. (Additional details on the survey are available from this commentator.)

Used simple methods. P & C examined alternative hypotheses by using simple methods. These methods seemed appropriate, and it was difficult to see a need for greater complexity. In short, P & C lacked the methodological rigor desired for publication in a prestigious journal. They would benefit from studying Siegfried (1970), who showed how even the simplest ideas can be made complex: He provided a rigorous proof that 1 + 1 = 2.

Failed to provide full disclosure. To obtain information from some journals, P & C had to promise confidentiality (personal communication with Peters). It is interesting that scientific journals find it necessary to suppress relevant information. Does this protect science - or does it merely protect scientists?

P & C were unable to provide full disclosure. This omission represents the only time that their study clearly followed the author's formula. This shortcoming is unfortunate. How can we be sure that P & C did the study? Hoaxes are not unknown in science. Mahoney (1979, notes 91-95) provides a short bibliography of scientific hoaxes and deceptions, and recent deceptions are described in Trafford (1981). Or perhaps P & C made errors in their analyses: Wolin's (1962) study showed errors to be common for published studies in psychology.

Access to the original data would be helpful in evaluating P & C. What papers were used in the study? Who were the referees? What did the referees say in their reports?

Wrote clearly. Many items in the author's formula can be overlooked if only authors would write obtusely. Obtuse writing also seems to yield higher prestige for the author (Armstrong 1980b). P & C almost followed the rule. Their Gunning fog index was about 18.[1] This represents material appropriate for a second-year graduate student and is typical for scientific papers. (For example, I calculated an average Gunning fog index of 18.7 for papers published in 58 sociology journals.) P & C's note 1 was certainly an excellent example of fog.

Recommendations. On the basis of prior research and their own study, P & C offer suggestions for improving the review system in journals. Three of their suggestions are of particular importance: structured guides for referees, open peer review, and blind refereeing.

Structured guides for referees. A structured guide can clarify the aims of the journal, which should reduce the likelihood of bias due to irrelevant factors such as similarity in beliefs or backgrounds. I suggest that articles be reviewed according to a structured guide designed to refute the author's formula. That is, positive ratings should be sought for importance, challenges to current beliefs, surprising results, simple methods, full disclosure, and clear writing.[2]

Open peer review. Some observers claim that referees will be more open in their criticism if their identity is kept secret from the authors and readers. Anonymity is certainly the traditional practice. In my survey of faculty members, 12 of the 22 respondents (55%) said they, as referees, would object to having their names revealed to the authors. Also, 12 others objected to having their names published along with the papers they reviewed.

Nevertheless, I agree with P & C that an open reviewing system would be preferable. It would be more equitable and more efficient. Knowing that they would have to defend their views before their peers should provide referees with the motivation to do a good job. Also, as a side benefit, referees would be recognized for the work they had done (at least for those papers that were published).

Open peer review would also improve communication. Referees and authors could discuss difficult issues to find ways to improve a paper, rather than dismissing it. Frequently, the review itself provides useful information. Should not these contributions be shared? Interested readers should have access to the reviews of the published papers.

For important issues, referees could publish their review along with the paper. The format would be similar to that currently used by the *Behavioral and Brain Sciences*. Such a procedure would provide a favorable bias for the acceptance of papers dealing with important issues. [*See editorial note following this commentary. Ed.*]

Blind refereeing. My conclusion, based on the prior research cited by P & C, is that blind refereeing should be used by journals. Some studies have found it to be helpful (Schaeffer 1970). Although other studies have indicated no need for blind refereeing (P & C might also have included the experiment by Mahoney, Kazdin & Kenigsberg 1978, and the survey by Kerr, Tolliver & Petree 1977), none has shown blind refereeing to be harmful. Furthermore, the cost of blind refereeing is negligible.

In the survey described above, my respondents thought that the reviewers would be able to guess the author and institution for about one-third of the papers. Furthermore, in P & C's study only three of the 38 reviewers (8%) were able to detect papers that had already been published in leading journals by individuals from prestigious institutions. In short, blind refereeing would be expected to be relevant for most papers.

Most journals do not use blind refereeing (e.g., Coe & Weinstock 1967 found that only 26% of the major economics journals used blind reviewing). Judged in light of the earlier pattern of research results, P & C provide compelling evidence in favor of blind refereeing. Finally, 14 of the 18 professors expressing an opinion in my survey (78%) thought that journals should use blind refereeing, so this change should be an easy one to make.

Fate of the author's formula. Academicians do not believe that the author's formula exists. In my survey, the 22 professors were asked the extent to which they thought each factor increased (+2) or decreased (−2) the likelihood of publication in "the highest prestige journals in your field." Table 2 summarizes the results. Respondents felt that authors should pick important problems, obtain surprising results, and write clearly. The only agreement, and it was modest, was that simple methods should be avoided.

Table 2 (Armstrong). *Academics' opinions of the author's formula (− 2 = greatly decreases to + 2 = greatly increases)*

	Effect on acceptance
If the author does	
not pick an important problem	−1.6
not challenge existing beliefs among scientists	−0.2
not provide surprising results	−1.1
not use simple methods	+0.4
not provide full disclosure	−0.7
not write clearly	−1.1

In contrast to these academics, I believe that the empirical evidence supports the existence of the author's formula (Armstrong 1982). The P & C case represents only a portion of the surprising research on scientific journals. We should examine this evidence and then take steps to penalize, rather than reward, use of the author's formula.

Further research. I wish that I had done the P & C study. One hopes that their study will be replicated and extended by others in fields other than psychology. The possibility of replications of this study should improve the vigilance of editors and referees: perhaps the next paper they receive has already been published.

Particularly important would be a replication with journals that provide blind reviews. This will help to determine the relative importance of two of the most likely hypotheses in P & C: Is it a game of chance or is it bias against unknown authors and institutions?

NOTES

1. The Gunning fog index was calculated in the following way: G = 0.4(S + W), where S is the average sentence length and W is the percentage of words with three or more syllables (not including prefixes or suffixes).

2. We have attempted to do this in the editorial manual for our *Journal of Forecasting*. Copies are available from the author.

J. S. Armstrong is an editor of the Journal of Forecasting.

BBS does not have a policy of open peer review (see C. Belshaw's commentary, this issue). Submitted manuscripts receive multiple, anonymous review (although referee anonymity is optional). Open peer commentary is accorded only to accepted articles (and then, of course, the referees are among those who are invited to submit a commentary). The Journal of Experimental Psychology: General *has a policy of occasionally copublishing referee reports with accepted articles, and* Speculations in Science and Technology *sometimes publishes dissenting reviewers' exchanges of correspondence with authors (see W. M. Honig's commentary, this issue). R. A. Gordon's commentary discusses a similar proposal. Ed.*

The fate of published articles, submitted again

John J. Bartko
Division of Biometry and Epidemiology, National Institute of Mental Health, Bethesda, Md. 20205

Is part of the exercise here to recognize that much of the material in this accepted-for-publication article has appeared earlier in "A Manuscript Masquerade. How Well Does the Review Process Work?," *The Sciences*, 20, September 1980, pp. 16–19, by Douglas P. Peters and Stephen J. Ceci?

A prior account of P & C's work also appeared in New Scientist *in March 1980 (see Cherfas 1980). Ed.*

On the failure to detect previously published research

Donald deB. Beaver
History of Science Department, Williams College, Williamstown, Mass. 01267

Although I agree in general with Peters & Ceci's (P & C's) careful and thoughtful analysis and discussion, I should like to comment on two points connected with the failure to recognize previously published research. Before I do so, however, it ought to be noted that what happened to their fictitious authors has given us a rare chance to see the second half of the Matthew Effect (Merton 1968a) in operation, "from him that hath not shall be taken away even that which he hath."

Given recent reports in *Science* (Broad 1980a; 1980b; 1981b; 1981c) concerning papers pirated and republished, the growing practice of having virtually the same articles published in two or more journals, and what is already known about the readership of journal articles, it isn't at all surprising that so few of P & C's sample papers were detected as having been previously published. What is surprising is that three were indeed detected. Any particular article in a journal issue is actually read by very few people - around 1% or less of the readership, or on the order of 1 in 100 (Garvey & Griffith 1971). [That only nine of the 12 articles escaped detection indicates that for the journals in P & C's study the editors and reviewers know the literature an order of magnitude better - on the order of 1 in 10 articles in a journal.] If a given article has a 9 in 10 chance of not having been seen before, then the joint probability that two reviewers and one editor will fail to recognize it is $(.9)^3$, or nearly the 3/4 of this study. Furthermore, there is another reason to suspect that these estimates are representative of science as a whole: They imply that if the average scientist sampling the journal literature regularly finds five articles to read each month, then the average reviewer reads 50 articles each month. To expect much more is impractical; the reality is probably less.

A significant question this research raises is whether or not its findings are generalizable to research fields other than psychology. Connected with the desirability of answering this question is one of the more worrisome features of P & C's article: its potential misinterpretation. Most scientists will not have read it, but will know something of its outcome. They will know that reviewers in psychology are biased, or that they not only don't know their literature, but don't even recognize quality. That is, some may be tempted to see the research as yet another exemplar of the soft, almost nonscientific, character of the social sciences.

In view of this possible interpretation, and because the natural sciences have provided much of the evidence for the (now somewhat battered) impression that peer review is relatively objective and meritocratic, like the social structure of science, it would be even more desirable to extend P & C's work. There is a potential difficulty in interpreting reviewer comments, however. It arises here in connection with the discussion and dismissal of the possibility that the reviewers were led to reject the submitted research because they recognized it as old. I suspect that this is also connected with the "hard" or "soft" nature of scientific fields.

Keeping in mind that editors and reviewers do not know the published literature thoroughly, consider the following elaboration of the discussion. Suppose that the research fronts of psychology are sharp, well defined, and regularly covered in review articles. Suppose further that the invisible colleges of each research specialty communicate reasonably efficiently so that their membership (which ought to include the majority of referees for journals as well) already knows the results of significant current works in advance of submission for publication (through preprints, conferences, personal communications, and the like). That means that given a mean lag of one year between submission and publication, the effective age of

research 18 to 32 months after publication is close to 2½ to 3½ years. For a hard science, that is a long time, not relatively short, as P & C claim. It is all the longer, if, following the earlier figures, one estimates that in that time a potential reviewer will have read more than 1,000 articles. As time passes, the important residue of the original information is not the details of those investigations but the general understanding that certain questions have been answered. Thus, as P & C envision (see "Discussion," paragraph 3), when an article is resubmitted 18 to 32 months after its publication, reviewers may know that it is outdated and superfluous without having read the original. It is important to realize that at that point the decision to reject has already been taken; for the reviewer, it is now a matter of justification.

Here it is that P & C remark on the absence of any reviewer comments to the effect that the (re)submitted articles represented already established findings, implying that the rejections were therefore not made on that basis. I should like to suggest a possible feature their analysis overlooks.

Neither conviction nor intuition that the research addresses a question already settled is sufficient reason for rejection. Evidence is necessary. The mere statement that the work is old, though effective, leaves itself open to challenge by the authors. Much time can be saved, and the whole problem can be much more conveniently resolved, by attributing the unworthiness of the paper to the universal critical deficiency of choice: "methodological flaws." Note that the flaws mentioned by the reviewers are not specific, but so general that they will "probably be considered flaws ten years from now." That is, they are precisely the ones to cite in order to dismiss a paper cleanly without any chance of rebuttal. Consequently, failure to comment on the redundancy of the research may not necessarily imply failure to recognize it as such.

This scenario suggests that there is a potential ambiguity in evaluating reviewer comments, one that research on psychology journals with blind reviewing might resolve. It further suggests that types of comments may depend on the structure of the field, a factor that ought to be considered in any attempt to extend the research.

Finally, let us hope that publication of this work won't inhibit its replication by having alerted editors and reviewers. Of course, some may hope otherwise - that this work will get the widest possible publicity, in order to forestall further investigation along these lines. Now, I wonder, did P & C consider this before they decided to publish. . . ?

Peer review and the *Current Anthropology* experience

Cyril Belshaw
Department of Anthropology, University of British Columbia, Vancouver, B.C., Canada V6T 2B2

The editorial experience of *Current Anthropology* (CA) bears on the questions raised and the problems revealed in Peters & Ceci's (P & C's) ingenious experiment. P & C's remarks about CA refer, of course, to commentary *after* acceptance for publication, rather than review *before* acceptance. Nevertheless, the CA system has special features that provide additional data. Unfortunately, I have not had time to assemble the exact figures that would quantify my impressions.

When an article is received we specifically select 15 reviewers (the figure used to be 20, which is probably more satisfactory). Since the journal has special international features, reviews are solicited from all continents, and almost always include one from the U.S.S.R. In almost all cases, however, most are from the United States, since this is where most anthropologists live. Most reviewers are professional Associates in CA; they receive the journal at a reduced rate, participate in policy formation, and have a moral obligation to enter into peer review and commentary. Selection is influenced by reviewers' past records (e.g., failure to respond, ad hominem views, etc.) as well as by their familiarity with the topic; an attempt is also made to avoid creating a burden by "overusing" reviewers. Reviewing is not limited to Associates, however, since the primary concern must be expertise in relation to the article's subject matter; further, most articles have an interdisciplinary component, so it is often necessary to go outside the ranks of anthropologists.

Reviewer responses are voluntary, so we cannot depend on a 100% return rate. We do not ask reviewers in advance whether they are willing to participate in each instance, especially if they are Associates, so there are often valid reasons (such as time constraints or insufficient expertise) for nonparticipation, which cannot be counterbalanced by our appeal to professional responsibility. The normal return rate is between 7 and 12 reviews out of 15. Reviewers may use a checklist response form, but they are encouraged to expand on reasons, which are crucial to the editor where there is disagreement or bias, and invaluable when criticisms are conveyed to the author. No attempt is made to disguise the author's identity or affiliation, since in most cases these are discernible from internal evidence. Reviewers are asked whether their names may be used (practices vary widely in different academic traditions), but the author is not given the reviewer's name unless he specifically asks for it or there are compelling reasons to put the two in touch.

When a paper is accepted, all reviewers who responded are invited to provide published Comments, and the panel of Commentators is now extended to 30 (originally 50). Any Associate may submit a later Comment for consideration (we sometimes get complaints from an anthropologist who believes he should have been included as either reviewer or Commentator).

I have gone into this in some detail in order to make the following points. The use of a group of 15 reviewers, instead of the usual one to three, provides an instrument for counteracting bias and revealing differences of opinion. *Almost no article receives a unanimous vote.* It is usually necessary, unless there is some countervailing principle of policy, for an article to cross the hurdle of at least 80% approval. But even there, one cogent criticism, revealing serious flaws in data, omissions in the literature used, fundamental misconceptions in terminology or methodology, and so forth, can occasionally counterbalance 10 bland approvals. Sometimes an article is refused, not because of inherent defects, but because it is more suited to a different type of journal.

Our experience reinforces the main conclusion of P & C's experiment. Dependence on two or three referees is not just insufficient but downright dangerous. However, that danger is reduced by the existence of alternative journals, with the possibility of publication elsewhere. This applies even to CA. Some of our refused articles (I try to avoid the term "rejected") turn up in other reputable journals, to critical acclaim. We sometimes publish material refused by other journals, particularly where an author has come up against an establishment in his own country interested in preventing the airing of an unorthodox position (a circumstance *not* limited to totalitarian political situations). My only criticism of BBS is that it underestimates international influences in its subject matter and networks. [*See editorial note following this commentary. Ed.*]

The linkage of refereeing to potential subsequent public commenting modifies the position of referees and provides some pressure of accountability. That is still far from perfect,

however. Sometimes I am simply forced to censor critical comments when they are so scathing that they would be destructive if communicated to the author. And the most difficult question to handle editorially is the matter of ad hominem attacks seeking publication, and the even more ad hominem (verging on libelous) replies of those who feel they have been attacked. In such situations I am continually accused of attempting to stifle honest debate. If one thing emerges clearly from the editorial experience, it is that our colleagues are emotional, easily hurt, and identify very strongly indeed with what passes for objective, impersonal science. Nevertheless, I am continually amazed at the willingness of most to offer their ideas to public debate. The main exceptions are "big names," some of whom seem extremely sensitive when their authority is questioned, and who dislike the immediacy of the commentary criticism.

Two additional points are relevant. I agree with P & C that it is difficult to detect previous publication, and I am not surprised by the experimental result. It could certainly happen to me as an individual editor or referee. Only the multiple review system has saved us from some extremely embarrassing incidents involving potential breach of copyright because of prior or simultaneous publication elsewhere. Here again, some very senior colleagues have misled us. On a couple of recent occasions, multiple publication has been caught only at the commentary stage, after acceptance for publication. Such carelessness, to put it mildly, is all the more incomprehensible since the copyright release form is clear on the matter (and CA is not necessarily opposed to republication under certain special, previously agreed-upon circumstances). Yet the majority of the review panel may miss the problem (admittedly more often in cases involving material that has been accepted elsewhere but has not yet appeared).

It may be wishful thinking, but I have not detected institutional bias. Perhaps as a discipline anthropology is a little more equitably distributed than some others. In addition, the commentary system draws in, of necessity, a wide range of participants, while the multiple review system is encouraging to junior authors since, whether the article is accepted or not, they stand a good chance of getting some useful advice. If the submission base is broad, and there is a genuine sense of participation, I think the bias in favor of prestigious institutions is partially modified. As is usual, it is an almost impossible task for unmodified M.A. or Ph.D. theses to get over the acceptance hurdle; original papers by graduate students, however, seem to be treated on their competitive merits, and they are sometimes successful in being accepted for CA treatment.

C. Belshaw is editor of Current Anthropology, *the experiment in scientific communication on which the BBS project, with its Open Peer Commentary service, is modeled.*

Professor Belshaw's commentary prompts me to consider five important points of comparison between the two projects:

1. Internationality can be said to be an intrinsic *factor in anthropology, inherent in the subject matter of the discipline, whereas in most areas of the biobehavioral sciences it is merely an* extrinsic *factor, as it is in most other sciences. There are certainly exceptions (e.g., human ethology, sociobiology, cross-cultural psychology), in which case reviewers and commentators are selected accordingly; but in most cases it is area of expertise, rather than geographic area, that dictates optimal selection strategy for BBS. Of course, BBS's Associateship, authorship, commentatorship and subscribership are all fully international (32%, 32%, 32% and 31% non-U.S. respectively) as is our readership-at-large (as indicated by reprint request data, reviews, citations, correspondence, etc.); and a special effort is in fact made to include European, Eastern European, and Far Eastern contributions in all BBS treatments.*

2. BBS usually only uses 5-8 referees, compared to CA's 15. However, these are all precontacted by telephone to ascertain willingness to referee and ability to meet our deadline; hence responses are close to 100%, yielding a return rate more comparable to CA's 7-12. Furthermore, it is our experience that whereas 5-8 is a suitable improvement over the conventional journal's usual sample of 1-3 referees (recall that a third coin toss will always ensure a tiebreaker if the first two come out heads and tails), still larger samples just increase the noise, reduce the likelihood that a manuscript can be successfully revised to everyone's satisfaction, and take on some of the undesirable features of committee writing. On the other hand, we do make an effort to ensure that our sample of referees is a fair and representative one, in part by using computer-assisted referee selection techniques (see H. R. Bernard's commentary, this issue) to search the current literature and the BBS Associateship for qualified specialists in the various areas on which a particular submission impinges. It is true, however, that time pressures and telephone costs have largely constrained our precontacting to U.S. and Canadian referees. Nevertheless, in cases in which particular foreign referees would clearly be optimal, they too are sent a copy of the manuscript for review, over and above the full domestic quota, but without precontacting. And, of course, commentaries (50+) are solicited worldwide. Subsequent Continuing Commentary sections permit further rounds of unsolicited commentary.

3. BBS too has a nonblind review procedure, with optional referee anonymity – the former because author nonanonymity has not yet proven to be a problem, and the latter to ensure the referee's freedom of judgment (as anonymous voting does the voter's in a free election). The referee is answerable, however, in that (a) the editor evaluates his report (checklist plus written text – see Cicchetti commentary, this issue) relative to the other reports and the manuscript itself, and (b) the author is asked to rate each referee report in terms of how favorable and how useful he finds it. Referee (and commentator) performance is thus cumulatively and systematically monitored; this information (along with data on willingness and timeliness) is further supplemented by ongoing internal statistical studies of the relations among referee anonymity, referee ratings of authors' papers, authors' ratings of referee reports, authors' ratings of commentaries (sometimes permitting a matched-guise comparison between an anonymous referee report and an open commentary by the same individual), editorial ratings, and the like. (Reports of these findings are in preparation.)

4. BBS experience does not indicate that the most distinguished investigators are less inclined to avail themselves of the Commentary service (authors in our first half decade have included Noam Chomsky, Iränaeus Eibl-Eibesfeldt, H. J. Eysenck, John Searle, E. O. Wilson, and many other senior colleagues); nor are the most prestigious names in the biobehavioral sciences missing from among the ranks of our active commentatorship and Associateship. On the contrary, it is remarkable how the entire BBS spectrum – from molecular neurobiology to artificial intelligence and philosophy of mind – has taken to the Commentary process, at all levels of distinction and seniority. We take this as a sign that CA's experiment in scientific communication can be successfully and productively generalized (with suitable adaptations) to other scientific and scholarly domains.

5. Interestingly enough, BBS too has on occasion been misled by the authors of manuscripts that had appeared or were going to appear elsewhere. One case to which the referees, editor, and *author had devoted considerable time and effort was just about to be circulated for Commentary when the editor chanced upon a suspiciously similar piece by the same author in the latest issue of another periodical. (And, yes, here as in the other similar cases, it was indeed a very senior author who had attempted the supernumerary submission.) Ed.*

Computer-assisted referee selection as a means of reducing potential editorial bias

H. Russell Bernard
Department of Anthropology, University of Florida, Gainesville, Fla. 32611

In this commentary I will discuss a mechanism for eliminating lengthy delays in the review process. It turns out that this mechanism (selection of referees from a large, computer-based file) also cuts down on some opportunities for bias (such as those identified by Peters & Ceci [P & C]) in the review process. I will not say much about the P & C article. It reports clever research, demonstrates competent analysis, and presents plausible explanations for some interesting findings. P & C's study focuses our attention on the scientific review process, and demands that we take a good hard look at it.

Six years ago, as a new editor of a major journal, I was faced with the problem of locating referees for a large number of manuscripts. I soon exhausted my own network and was forced to comb the journals in order to locate appropriate reviewers. Many authors were not to be found at the addresses listed in their articles. Furthermore, the search process was tedious, and the whole thing became a little intimidating.

I was very much heartened to learn that my discomfort was widely shared by editorial colleagues. Long delays in the review process, it turned out, could be attributed to the simple fact that editors can't find enough reviewers who will respond quickly. The solution to this problem was the creation of a very large, computer-based file of potential referees. Names were culled from the guides to departments of anthropology and sociology. Reviewers received a letter explaining how I had found them and requesting names of colleagues who might not be listed in the usual academic guidebooks. The file quickly grew to over 3,000 names.

A complete description of the procedure that is used to manipulate the file is given in Bernard 1980. Briefly, names and specialties are entered in free format into a logical text editor. The text editor is used as a data-base management device to find potential reviewers with any number of simultaneous specialties. For example, all persons who have listed "Africa" as a specialty can be located; or the much smaller lot of Africanists who also list "urban" as a specialty can be found. Or the roster of those who list "agriculture," "women," and "Latin America" can be found. And so on.

Use of this technique cut the review time down to less than two months, on average. If a reviewer fails to respond within three to four weeks, I just send out copies of a manuscript to alternative referees. The computer is utterly unintimidated by my requests for names of more potential referees.

Moreover, the use of this technique makes it easier for editors to overcome lurking biases they may have toward an article. P & C have shown that the authors' names and affiliations may predispose an editor favorably or unfavorably toward a manuscript. In addition, the subject matter, the conclusions, the epistemological approach, or the political overtones may rub an editor the wrong way. The challenge, it seems to me, is to exercise our craft, and to make good editorial judgments *in spite of* these biases, because the biases don't go away. The computer-based referee file can help editors meet this challenge. Here's why.

It is really quite simple for me as an editor to guarantee that an article will be killed by referees. All I need to do is to select referees I know can be trusted to clobber a particular manuscript. (Of course, it is easier to do this if I can offer the referees anonymity, a practice I fully endorse.) The availability of a very large, tireless robot who finds potential reviewers for me does not guarantee that I will never be punitive, vindictive, petty, or ruthless with authors whose work I don't like personally. But it sure helps me to avoid those sins. If it is *my* prejudice that makes someone's good work look bad to me, then a sample of referees (whom I don't know personally) chosen from a big list is likely to show me just that. The referees will probably like the work a lot. I'll be faced with some cognitive dissonance, and will have to reconsider my opinions. It's a bit masochistic, but very illuminating. And, from my own experience (an n of 1, unfortunately), it works.

P & C's article is very important. It is an excellent case study of the dilemma of management. In complex organizations, the decision-making environment is just too complex for people to comprehend. (I am indebted to Bruce Mayhew for his thinking on this problem.) The result is that managers guess a lot, and they often base their guesses on input from "consultants." One technique, well-known in government and industry, is to select consultants who will tell you what you've already decided is true. In scientific editing, I believe, we cannot afford the luxury of this technique. It is not conducive to the free flow of ideas.

On the other hand, space in good journals *is* limited; and pressure on that space is mounting. And decisions *do* have to be made, no matter how we tackle the problem of selecting among competing articles (either with traditional review or the *Current Anthropology* approach). [*See commentaries by S. Tax & R. A. Rubinstein and C. Belscaw, this issue. Ed.*]

It is important to point out that my robot file does not cause *me* to become a robot. I still decide whom to send manuscripts to; and there is still room for prejudice to operate in favor of or against a particular manuscript. But my experience has been that the computer-based referee search procedure has helped me to be a more responsive and a more responsible editor than I could otherwise be. I am convinced that the craft of scientific editing and decison making can be improved through the use of computer-based referee selection files. And we need to move decisively toward the development of a more valid computer-based referee file than the one I have been using.

What is needed is a central service that produces and updates a master file of reviewers for major scientific areas. "Social and behavioral sciences" seems like a plausible area. The file should consist of names of persons who have actually published in, say, the last five or six years in a given list of journals. Books and theses should be included as well. This would be a behaviorally based reviewer file rather than one based on self-reports of expertise (which is the major flaw in the system I use now). Each discipline or professional association might develop a system; or a group of associations might seek cooperative findings to initiate a more cross-disciplinary service effort. Journals could subscribe to such a service, paying a subscription or fee for subfiles in particular areas. When I began experimenting with the technique I've described, only mainframe computers were up to the job. Today, most journals could use a friendly microcomputer for handling a few thousand potential reviewers.

The benefits of such a system are many, and the cost is small. We should get on with it, and with other experiments to improve the flow of scientific information.

H. R. Bernard is former editor of Human Organization *and current editor of* American Anthropologist.

BBS too has developed a computer-assisted referee- and commentator-selection system, which includes searching the current file of BBS Associates on the basis of their coded areas of expertise as well as searching the current biobehavioral literature on the basis of keywords and citations. Ed.

Explaining an unsurprising demonstration: High rejection rates and scarcity of space

Janice M. Beyer
School of Management, State University of New York at Buffalo, Buffalo, N. Y. 14214

Peters & Ceci's (P & C's) study belongs in that class of scientific experiments in which scientists demonstrate rather

than discover relationships. Because the authors knew in advance what results they hoped to demonstrate, they consciously or unconsciously designed their study to ensure those results:

1. P & C's choice of psychology journals assured them of high rejection rates, and thus increased the probability that the previously accepted articles would be rejected. If they had chosen to study journals from fields with lower rejection rates - the physical sciences, for example - they would have been more likely to find that previously accepted articles were accepted again.

2. P & C's alterations of the previously published articles were designed, by their own admission, to make recognition of the articles more difficult. Nevertheless, in one fourth of the cases, the article was recognized. The authors obscure this result by reporting their results for individuals, not articles. This is not entirely appropriate, since the individual observations are not independent.

P & C's results regarding the rejection of the previously accepted articles are presented under the heading "Reviewer reliability." This implies that the later rejection of previously published articles is evidence of low reliability among reviewers and editors. But we do not know that exactly the same articles were accepted and later rejected. It is quite common to ask authors to shorten their articles between acceptance and publication, and often methodological detail is sacrificed to make the cuts. In my experience, this cutting, plus the additions needed to meet reviewer comments, can also lead to uneven writing and unclear passages.

P & C make other unwarranted assumptions in interpreting these data. First, they assume that the original acceptance of the articles they resubmitted means that there was "presumably near perfect agreement among the original reviewers in favor of publishing." They have no data to support this assumption. My own research (Beyer 1978:76-77) suggests that it is very unlikely that all of the articles received unanimously favorable reviews at the time of their first submission. P & C's assumption may thus exaggerate the differences between the two sets of reviewers. Finally, it should be pointed out that the results presented in Table 1 show very high interrater reliability.

Because they changed the authors' affiliations on the articles and blind reviewing was not used by these journals, P & C interpret the change from acceptance to rejection as suggestive of bias based on authors' affiliations. This interpretation is supported by past research. Evidence of partiality has also been found in studies of journals from other fields (Beyer 1978; Pfeffer, Leong & Strehl 1977; Yoels 1974), including fields in which blind reviewing is the norm. Unfortunately, P & C interpret the data as if the bogus affiliations they supplied had changed the authors' affiliations only in terms of prestige. In fact, they changed the affiliations from known universities with high rankings to unknown nonacademic institutions with no known ranking. Not only is this a change in relative prestige, it is also a change from academic to nonacademic. Any bias their study uncovered was just as likely to result from the nonacademic origin of the articles as from the status of their authors and their affiliations. In fact, it could be argued that the academic-nonacademic distinction may have been more noticeable and important to referees (causing them, for example, to wonder how people working in such places could produce the types of research reported in the articles).

Also, by relegating it to note 6, P & C play down the most obvious and likely explanation for their findings. They acknowledge that referee judgments may show "least variability... in the region of outright rejection." When 80 to 90% of articles submitted to behavioral and social science journals are rejected, the most likely fate of any submitted article is to be unanimously rejected. Of the nine articles they submitted, eight were rejected and one was accepted. This proportion is probably pretty close to the average acceptance rate of the journals included in their study. The higher rejection rates of journals in some fields compared to others is probably more a reflection of the availability of space (Beyer 1978:79-81) than of concern with the quality of articles. Ratios of circulation to numbers of articles published show much greater scarcity of space in fields that have high rejection rates (Beyer 1978:79; Hargens, 1975:20-21). Because of their shortage of space, "social science referees know they must reject most articles or they will be deviant referees. They must find and document reasons for that rejection" (Beyer 1978:81). Acceptance rates that exceed available space create backlogs of accepted articles waiting to be published. Any persistent mismatching of rejection rates and available space leads to intolerable situations: either a lack of material to print or ever-increasing backlogs.

Finally, I feel I have to comment upon the use of deception in this study and to question its justification. If such research can be justified at all, presumably it must be justified in terms of beneficial results that outweigh costs and could not otherwise be obtained. In the case of the P & C study, it is hard to see what benefits accrue to society, or even the scientific community, that outweigh the distress brought upon the editors and reviewers of the journals included in their study. P & C's results do not stand up well under close scrutiny, and lack generalizability in any case.

The problems associated with the publication of scientific results are difficult and important ones. They deserve studies that are more than demonstrations that problems exist. What is needed is research that can discover new relationships and unravel the causes of the problems.

Peer review and the structure of knowledge

Marlan Blissett
LBJ School of Public Affairs, University of Texas, Austin, Tex. 78712

Peer review may be hard to define, but whatever it is, all disciplines use it, scientific reputations are made by it, and research grants are rarely given without it. There are different versions of peer review - almost as many as there are different kinds of science. In *Politics in Science* (Blissett 1972) I suggested it might be possible to develop a typology of peer "regimes" based on the degree of theoretical consensus in a discipline and the number of individuals permitted to influence decisions. While this exercise might be useful in explaining *how* scientific disagreements are resolved, it does not address the more important questions of reliability and fairness. In the field of psychology Peters & Ceci (P & C) have made a significant step in this direction. What I offer here are some marginal comments on their findings.

Like most of the behavioral sciences, psychology has become a highly specialized discipline, but one without a unifying structure. This fact alone greatly complicates the reliability of peer review. The immediate effect is that fewer and fewer psychologists can agree on research priorities, and fewer still are capable of integrating different specialties. Perhaps this condition accounts for the failure of an overwhelming majority of editors and reviewers (92 percent of editors and reviewers combined and 87 percent of reviewers alone) to detect published articles that had been resubmitted for review. It may also explain why the most serious criticisms from reviewers were methodological and not substantive in character. When agreements on content are few and far between, one might think the only things left to talk about are research design and the use of numbers.

But is this the case? If reviewers and editors can reject 89 percent of previously published articles - largely for methodological reasons - can we assume even a consensus on research methods? There is certainly room for doubt. And if a consensus does not exist on the proper way to do research, peer reliability is further weakened. Questions begin to pop up from everywhere. What has happened? Are researchers receiving poor training in graduate school? Is there a rapid obsolescence of research skills among practitioners? Do the problems selected for research defy quantitative analysis or does quantitative analysis defy the problems under study? The answers eventually lead back to the structure of the discipline and its lack of unity.

The reliability of peer review also affects the distribution of research grants from public agencies. Although applied less rigorously than in an academic discipline, peer evaluation in the form of panels, study groups, site visits, and the like can be an important stage through which an administrator goes in making allocation decisions. From this standpoint, as William Carey (1975) has pointed out, "peer review is a proxy for assaying the standards of the scientific community." If those standards can be questioned, then scientific merit will be given a lower priority in the decision process.

Maintaining the reliability of peer review is essential to the growth of any discipline. But satisfaction with the review process can also be influenced by questions of fairness. Do nonestablishment scientists get short shrift? The findings of P & C strongly suggest they do. The Tri-Valley Center for Human Potential just doesn't stack up to "prestigious and highly productive" psychology departments. If the research were of uniformly high quality, one could at least justify this discrimination by appealing to a higher principle - (say) the advancement of knowledge. But the peer practices uncovered here undermine such confidence. The bulk of published research cannot be rejected the second time around without creating the impression that what came out in the first place was of marginal quality.

A number of mechanisms have been proposed to rectify individual mistakes and improve collective judgments of quality. These include: appeal panels, higher tribunals, open peer commentary, improved selection procedures, standard rating forms, and referee reviews. In some cases these devices may prove helpful. But the problems facing psychology and the behavioral science community cannot be solved by process solutions. The problems are content problems and go to the nature and structure of the disciplines involved. P & C should be congratulated for pointing to the tip of the iceberg. But massive problems lie hidden beneath the surface.

Reforming peer review: From recycling to reflexivity

Daryl E. Chubin
Technology and Science Policy Program, School of Social Sciences, Georgia Institute of Technology, Atlanta, Ga. 30332

The Peters & Ceci (P & C) paper invites two kinds of criticisms: One focuses on what was done and how it was reported, the other on what was overlooked and the significance of those missing elements. With all the hubris of an outsider - a nonpsychologist and confirmed critic of peer review *as practiced* (not as a concept or principle) - I prefer to stray from the authors' text, pose some different questions, and cite literature relevant to those questions that eluded the authors.

My chief objection to P & C's approach is its experimental guise. The recycling of published papers through the journal review process may be the only way to generate primary data since editors and referees are typically reluctant to grant researchers access to either their files or their recollections. Nevertheless, such recycling strikes me as ethically suspect - but I'll let the editors in question wrestle with that issue. What I find even more questionable is P & C's effort to legitimate their research as *good* experimentation. What is presented as procedurally sound, however, lends an aura of rationality to a process that appears to be largely irrational. Referees will muster whatever methodological rhetoric is needed to justify rejection. But does P & C's paper help us to reconstruct *this* process? Not a bit. Indeed, all that is established is the myopia of referees in psychology. That doesn't surprise me (or other sociologists, e.g., Freese 1979); rather, it underscores my reservations about a system that has been perverted by those who participate in it.

Therefore I ask, How might the system be reformed? Why not afford authors an opportunity to reply to reviews *prior* to the editorial decision (Glenn 1976)? Such a dialectical review-and-recourse process would help recognize referees for performing valuable but thankless work, and restore "accountability" to the performance that conscientious reviewing demands. Subsequently, we could remove the anonymity of reviewing and append to the published article the names of referees who recommended publication. Wouldn't that represent the communal spirit?

Other questions that P & C fail to consider include: How are referees selected to review a manuscript? Are associate editors pivotal in this process? And do authors' statuses and institutional affiliations define who are their "peers"? These questions, in the context of journal publication, have been addressed by Lindsey (1978) and the commentators (Symposium 1979) on his *The Scientific Publication System in Social Science*. At issue in this work and commentary is the politics of reviewing and the underlying epistemologies that clash in the review process. As Mahoney (1979) has suggested in another context, the psychology of scientists predisposes them to intellectual postures that can undermine as well as champion the claims to knowledge made in submitted manuscripts and research proposals.

It is in the grant proposal domain that peer review becomes so odious (Greenberg 1980; Mitroff & Chubin 1979; Roy 1981). P & C barely acknowledge this crucial link between scientists' opinions of one another under the conditions of competition for scarce resources. For resources afford differential advantages that tend to color merit, if not to obscure it beneath a researcher's reputation. These advantages accumulate; just how they warrant funding decisions and advance careers to the detriment of science is unknown. What collusive role do scientists play in sanctioning the priorities and criteria of funding agencies, programs, and editors? Social scientists have differed regarding the propriety of voicing such concerns about peer review (Chubin 1980; Cole, Rubin & Cole 1978) so the oversight of P & C is not theirs alone.

Nevertheless, such "oversights" must be challenged if peer review as a process and a system of self-governance is ever to improve. For peer review is not merely a problem in psychology; it is a problem for all of science. Surely reform is essential if the research - promised or produced - that is certified as original scientific knowledge is neither original nor good science. If it is merely recycled and reclaimed but undetected science, then our specialization has betrayed us. If we smugly reject studies of peer review, we foreclose the prospect of reforming the system and process so dear to us all. Similarly, if we embrace uncritically the approach and results of P & C, we endorse both the practice *and* the principle of peer review. That would be irresponsible and shamefully unreflective for students of human behavior - especially that behavior in which we ourselves engage, benefit from, and ratify as peers.

On peer review: "We have met the enemy and he is us"

Domenic V. Cicchetti
VA Medical Center, West Haven, Conn. 06516

The rather provocative article by Peters & Ceci (P & C) sets off in bold relief the fact that it is not at all unusual for one referee to impugn the quality of the same research that another independent evaluator extolls as a worthwhile contribution to science (see also Patterson 1969). What does this sorry state of affairs portend for the future of peer review in behavioral and medical research? Let me first evaluate the remedies discussed by P & C and then discuss some of the more subtle, rather ephemeral aspects of the review process that militate against an easy solution to this critical problem.

With the possible exception of two proposed antidotes, namely, ceasing to call upon reviewers who receive repeated legitimate complaints, and increasing the role of "creative disagreement" (Harnad 1979), the remaining suggested cures seem almost certainly doomed to failure. Let me explain why. Consider the recommendation for reviewer disclosure. A bright, prolific, promising young research investigator is selected to evaluate some badly flawed work of a notable research titan. Fearing reprisal, the young author might easily decline the invitation to serve as referee under the stipulation of reviewer disclosure. After all, well-established scientists serve on review and editorial boards with much greater frequency than do young Turks. Reviewer disclosure policies might then produce an excess of well-established referees, in relatively secure academic positions, who may have a stake in perpetuating the status quo in research. (Such a situation is implied in the delightful Swindel & Perry 1975 spoof on the scientific review process.)

With regard to (a) the continued development and refinement of more sophisticated manuscript attribute rating forms and (b) the training of referees to perceive the relationship between high scores on such lists and the publishability of a given manuscript, past research indicates that such endeavors will probably contribute little toward improving the quality of the review process. We found recently that reviewer agreement levels were in fact lower for about 450 manuscripts for which a 7-item rating form (see Table 1) was applied by referees (R intraclass (I) = .15) than for approximately 600 manuscripts that were evaluated without the presumed benefit of our rating form (R (I) = .21) (Cicchetti & Eron 1979). Consistent with this finding was a study of peer-review practices for the *American Psychologist*, which reported the highest interreferee agreement levels to date (R (I) = .55), despite the fact that a rating form was not available to reviewers (Cicchetti 1980).

Our research (Cicchetti & Eron 1979) also implies that training reviewers to understand the relationship between rating-form items and the publishability of a given manuscript may also prove futile. Specifically, we correlated our rating-form scores (rank ordered on a 4-category scale from "excellent" to "devoid of scientific merit") with reviewer recommendation levels (rank ordered on a 4-category scale from "accept as is" to "reject"). These correlations were computed separately for each of two independent reviews of about 400 manuscripts. Interestingly, both sets of reviewers were in high agreement on the relationship between our checklist variables and the publication potential of individual manuscripts: (1) Both sets of reviewers agreed that the "importance" of the contribution to the field is the most relevant criterion on which to base this final recommendation (R = .71 and .73, respectively); (2) there was also high agreement that the next most relevant attribute is the quality of the "research design" (R = .63 and .62); (3) "succinctness" was regarded as the least relevant criterion (R = .32 and .37); and (4) the remaining checklist criteria were rated "in between" by the two independent sets of referees (correlations ranging between .38 and .46). What does this mean? Simply that reviewers are in considerable agreement about the relative weighting of scientific attributes. They just cannot seem to agree on which high and low scores should be paired with which manuscripts.

Table 1 (Cicchetti). *Manuscript attribute rater form*, Journal of Abnormal Psychology (JAP)

1. Probable reader interest in the problem[a]
2. Importance of present contribution[b]
3. Attention to relevant literature[b]
4. Design of research[b, c]
5. Analysis of data[b, c]
6. Style and organization of the material presented[b]
7. Succinctness[b]

[a] 4-point scale.
[b] 10-point scale relative to average JAP article.
[c] To be disregarded with case reports and theoretical articles.
Source: Based on Scott 1974.

Let us now consider some of the more ephemeral aspects of the peer-review process. Ponder the situation in which two referees agree (for basically similar reasons) that a particular research study should receive high priority for publication. The editor accepts the manuscript. However, once the article appears, a knowledgeable specialist in the field detects a fatal flaw in the research design that invalidates the authors' conclusions. Thus, a highly reliable decision becomes devoid of validity. The same phenomenon occurs on peer-review panels (simply substitute primary and secondary reviewers for the two referees and an ad hoc expert in the field for the person who detects the flaw in the published research article). Another even more subtle problem that plagues the review process is one noted by Smigel and Ross (1970), in which two referees cite essentially the same criticisms of a manuscript. One reviewer recommends rejection because he considers his criticisms serious ones. The second reviewer regards essentially the same criticisms as minor and so opts, instead, for publication of the article. A third class of elusive problems occurs when both referees agree to recommend acceptance, revision, or rejection, but for entirely different and sometimes conflicting reasons.

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. Moreover, as long as journals continue to support and encourage the rejection of about four out of every five manuscripts sent out for peer review, the evaluation process may never improve all that much. For example, our own research shows that two out of every three manuscripts receiving a split decision (accept vs. resubmit or accept vs. reject) will ultimately be rejected (Cicchetti & Eron 1979, p. 600). And so, as we peer once again into the "peer view" mirror, we may be forced to agree with Pogo: "We have met the enemy and he is us" (Kelly 1972).

Manuscript evaluation by journal referees and editors: Randomness or bias?

Andrew M. Colman
Department of Psychology, University of Leicester, Leicester LE1 7RH, England

One of the most influential discoveries in modern science, the first law of thermodynamics (sometimes called the law of the

conservation of energy) was first reported by the German physician J. R. Mayer in 1842. But Mayer's revolutionary paper was rejected by the leading physics journal *Annalen der Physik* and was eventually published in a relatively obscure and much less appropriate chemical journal. It was therefore almost entirely ignored by physicists, and, possibly as a result of this, Mayer suffered a mental breakdown from which he never recovered (Ziman 1976, pp. 103–4). This is just one dramatic example of the fallible judgments to which journal editors and referees are sometimes prone; I have cited some other equally shocking examples elsewhere (Colman 1979).

Peters & Ceci (P & C) have reported some interesting empirical evidence from a controlled investigation of the peer-review system. Their data demonstrate that the system is vulnerable either to random error or to systematic bias, or possibly to both. The authors acknowledge that in order to test the bias hypothesis properly it would be necessary to compare the fate of resubmitted articles purporting to come from high-status institutions with the fate of others purportedly from low-status institutions. Since this manipulation was not performed, P & C's interpretation of their results as supporting the bias hypothesis lacks force, and I believe that their statistical arguments against the random error hypothesis are unsound.

Let us assume that the ultimate fate of a submitted manuscript or an undetected resubmission is a purely random event, unrelated to its quality or to the authors' reputations or their institutional affiliation. Suppose that there is a fixed probability p that the manuscript will be accepted and a probability of $q = 1 - p$ that it will be rejected. (In reality, of course, there are other possible outcomes apart from outright acceptance and outright rejection, but I shall ignore this complication as P & C have done.) If the number of submitted - or resubmitted and undetected - manuscripts is N, then the exact probability $P(x)$ that x of them will be accepted is given by the binomial probability function:

$$P(x) = \frac{N!}{x!(N - x)!} p^x q^{N-x}, x = 0, \ldots, N.$$

The logical derivation of this formula is explained from first principles in Colman (1981, chapter 4). P & C correctly state that if $N = 9$, $p = .43$, and $q = .57$, then the probability of less than two acceptances - that is, one or zero acceptances - is $P(1) + P(0) = .046$.

This does not, however, answer the question, How improbable is the observed outcome of less than two acceptances out of nine on the basis of chance alone given the actual acceptance rate (20 percent) of the journals studied? The required probability is obtained by setting $N = 9$, $p = .20$, and $q = .80$. Then, according to my electronic abacus, the probability of less than two acceptances is $P(1) + P(0) = .44$. This means that, *on the assumption of purely random selection*, the probability of an outcome as extreme as that observed by Peters and Ceci is .44, which is certainly not low by any standards. Furthermore, the *expected* number of acceptances, given the above parameters, is $Np = 1.80$, which is fairly close to the observed outcome of one acceptance (the standard deviation is 1.20). In other words, if the experiment were repeated many times, then between one and two manuscripts, on average, would be accepted per experiment. It seems imprudent, therefore, to reject the hypothesis that the fate of the manuscripts resubmitted by P & C was determined purely randomly. This does not, of course, prove that bias was absent, but Occam's razor bids us to reject the bias hypothesis in favor of the null hypothesis of random selection.

P & C attempted to show that the relative frequency of favorable reviews by referees and editors was significantly less for the resubmissions than for the original submissions. Using a conservative estimate of the latter, they rejected the null hypothesis of no significant difference on the basis of a chi square test. Unfortunately, this conclusion is invalid because editors' reviews are clearly influenced by those of their referees; hence the crucial assumption of stochastic independence of observations underlying the chi square statistical model was violated in P & C's calculation. It cannot, therefore, be inferred that the resubmissions received significantly fewer favorable reviews than the original submissions, or that the observed outcome was "quite improbable," as P & C claim.

Whether referees and editors are systematically biased or operate in a quasi-random fashion, the peer-review system evidently lacks validity. When referees claim to have found serious flaws in a manuscript, therefore, there is no a priori reason to assume that they are right and the authors are wrong. If editors lack sufficient specialized knowledge to evaluate the criticisms, how ought they to respond to unfavorable referees' reports? The following procedure, which has been successfully used by the new journals *Current Psychological Research* and *Current Psychological Reviews*, seems to me to be most fair. The authors should be sent the referees' criticisms and be invited to rebut them if they consider them invalid. The original manuscript, together with the referees' criticisms and the authors' rebuttals, should then be submitted to an independent arbiter for a final verdict. This procedure would perhaps eliminate some of the more blatant injustices of the peer-review system and act as a corrective to referee error.

A. M. Colman is executive editor of Current Psychological Research *and* Current Psychological Reviews. *Ed.*

Criterion problems in journal review practices

John D. Cone
Department of Psychology, West Virginia University, Morgantown, W. Va. 26506

The provocative paper by Peters & Ceci (P & C) further documents persistent problems of unreliability and possible bias in the peer-review practices common to professional journals in the behavioral and physical sciences. As an associate editor of a psychology journal (*Behavioral Assessment*) I was surprised at P & C's finding that the same article was not recognized by the editors who had handled it just 18 to 32 months earlier. This is especially surprising in view of the manuscript's acceptance the first time around, since accepted papers are usually handled several times as they wind their way through revision, copy-editing, and final processing for publication. The forgetting of rejected manuscripts would be less surprising.

Nonetheless, the P & C results are compelling, and the editors apparently did forget. It is not the editors' faulty memory that is the primary focus of this study, nor of these comments, however. Editors of APA journals typically handle hundreds of manuscripts each year, and it is not expected, or even desirable, that they remember each one. Perhaps associate editors should be expected to do better, but even for them the implications of the P & C findings are that editorial recollection is but one element in a complex, often hastily enacted process that requires serious study and our commitment to overhauling.

Such study would begin with an analysis of the variables controlling the reviewing process. Disagreement among referees is not surprising when it is realized that reviewers, typically prominent and overextended researchers themselves, work independently, under tight time constraints, with minimal criteria to guide them, with no opportunity to question the author for clarification, with minimal feedback as to the adequacy of their reviews, and with few rewards for the long hours devoted to the process. Judgments by independent experts concerning simpler sets of stimuli than those repre-

sented by journal submissions have long been known to vary considerably. Efforts to enhance agreement between judges have frequently included greater explicitness in defining the attributes to be judged, better training of judges, and, occasionally, financial or other incentives for doing a good job. In recent years an extensive body of literature has emerged dealing with judgment problems in the direct observation of human performance (for reviews see Cone & Foster, in press; Hartmann & Wood 1981; Weick, in press). Lessons learned from that literature could be of some value in the study of the journal review process as well.

It should be noted that the very practice of securing multiple reviews of a submission suggests a less than perfectly objective enterprise. The implication is clearly that we cannot know a paper's worth in any absolute sense and therefore its definition by consensus is required. This further implies the futility of endeavoring to discover "an explicit set of evaluation criteria" against which "a standard rating form" and "training of referees" could be developed, as P & C recommend. The connection between the practice of securing multiple reviews and the assumption of an "unknowable criterion" has apparently been lost on the purveyors of suggestions such as these. And well it should have been. For, while there is certainly no standard in any absolute sense, there is very likely to be a set of ingredients that acceptable papers should include and the inclusion of which could be agreed to by any number of competently trained reviewers or journal editors.

The problem is not a new one, nor is it particularly mysterious. It is merely an issue that has been underresearched in psychology, specifically, and in the realm of scientific publication, generally. Doubtless many reasons for this relative neglect could be offered, and many would reflect the basic reward system underlying the review process itself. The many hours necessary to accomplish a thorough, well considered review and subsequent, constructively articulated report thereof fall into the relatively invisible realm of service to the profession. Promotion and tenure committees do not weigh reviews heavily, and the reviews' very anonymity further underscores the minimal recognition accruing to their authors. Reviewers are willing to go just so far in return for having their names listed inside front covers and for the opportunities of seeing research reports at their earliest stages and keeping their analytic skills finely honed. To ask them to submit to a systematic program of reviewer training, to use a standard rating form, or to have their performance evaluated in regular and objective ways may be asking too much in view of the present reward system.

Moreover, research on the reviewing process itself is something only slightly more professionally enhancing than performing the reviews. The methodology of science is, after all, a rather pedestrian affair when compared with the discovery of basic variables and the laws governing their interaction. It is generally appreciated that applied research is less prestigious than its basic counterpart, and research in the methodology of science is no exception.

But, having said all this, what are my recommendations?

1. I support P & C's call for more research on the peer-review process.

2. I urge professional societies to lobby for increased federal support for research in this area.

3. I urge professional societies to support such research themselves.

4. Research needs to address the problem of specifying criteria for acceptable papers.

5. Research needs to consider the training needed for consistent and accurate application of these criteria by independent reviewers. In such research, the accuracy of the individual reviewer rather than agreement among several should be the principal focus. In the P & C paper, for example, we do not know which set of reviewers performed more creditably. The implication is that the first were positively biased by the name and institutional affiliation of the authors. Perhaps the second were guilty of negative bias, however. The existence of accuracy criteria would enable answers to be provided for such questions.

6. Procedures for monitoring the continued accuracy of trained reviewers must be developed.

7. A reward system capable of maintaining accuracy should be established. While this might eventually involve paying reviewers for their work, it is conceivable that performance-dependent selection and retention on editorial boards might be sufficient to encourage consistently high-level reviewing.

Postscript. Having said these things in the initial draft of this commentary I turned, appropriately enough, to the completion of an overdue review. It provided a sobering stimulus for some reality testing with respect to the suggestions I had just made since my review turned negative after I discovered a crucial flaw in the design of the study. I wondered whether crucial-flaw-discovering was really an art (or the product of genius!) that could never be rendered objective. Understandably, I initially decided that it was, and that my suggestions could never be fully implemented. However, more reflection showed the folly of this reasoning. Of course, crucial-flaw-discovering is a skill that can be objectified and taught along with other components of the "artful" reviewer's repertoire. It merely requires a more refined technology.

Having survived this "test" I can conclude that the above suggestions are sound and can be offered for the serious consideration of the scientific community.

J. D. Cone is associate editor of Behavioral Assessment. *Ed.*

Editorial responsibilities in manuscript review

Rick Crandall
Department of Psychology, Dominican College, San Rafael, Calif. 94901

Peters & Ceci's (P & C's) article on peer review has some limitations that could be criticized. Because they didn't do a full experimental design, submitting good and bad articles from high and low prestige institutions and authors, their results could have been caused by some factor of which we are unaware. Thus their study is only an indirect test of the notion that there is a bias in peer reviews based on status. However, for me, the most important thing is not to criticize the methodology but to identify why the article really scared me. Perhaps the simplest reason was that I had always assumed that blind review would be a good idea, just in case there was any minor bias in the editorial process. Now I'm faced with the possibility that the difference between a good article that's accepted and a good article that's rejected may be a minor factor like the status of the author and institution.

In a loaded area like peer review, it is important to make clear our own assumptions and experience. A few years ago I drafted a paper outlining why I thought editors should have more obligations to ensure fair review practices. I never finished the paper, but I did go on record as arguing that agreement between reviewers was better than it looked because the wrong statistics had been used (Crandall 1978a); I also discussed other publication-related issues (Crandall 1977; 1978b; Crandall & Diener 1978). There is some ironic comfort for me in the fact that P & C showed 100% reviewer reliability for their papers *on the one occasion.* I believe that most editors and reviewers are responsible. As an author I've gone through a formal review process about 30 times. My personal anger at some rejections was objectified by unpublished work by N. J. Spencer demonstrating through linguistic analysis that reviews include a considerable portion of nonconcrete, unanchored generalities, with at least a dash of simple bias or gratuitous

insult. I've reviewed over 50 papers as reviewer, consulting editor, and deciding associate editor across several journals. As I'll explain shortly, despite being on both sides of the fence, my sympathies are definitely with authors as the lower-power part of the editorial exchange.

Before I offer some value judgments about editorial responsibilities, I would like to offer two possibilities showing how status may not be as important a factor as P & C suggest. It is possible that when the articles in question were first accepted and published, they constituted a reasonably important breakthrough or presented some new data. Even though none of the reviewers criticized the papers on the basis of their results not being new, they may still have had in the back of their mind a sense that these general results had already been found; thus they could have switched into a more critical mode, requiring cleaner methodology and thus subjecting the article to much more criticism. This is an area where it would have been particularly nice to have some other experimental groups and more cooperation from editors to answer this question.

My second guess is that very low status has more negative effects than high status has positive effects on editorial decisions. There may be hundreds of acceptable institutions that are not discriminated against by reviewers. I found the authors' fake institutions almost negatively prestigious. We should know just how low they are. As a start, I quickly asked four psychologists to rate four of the authors' institutions plus two actual lower-status places on a 1-7 point prestige scale, with 7 being the most prestigious. The results follow:

High prestige: Yale = 6.5, University of Wisconsin, Madison = 5

Modest prestige (real): George Mason University (unknown to all) = 3.25, University of Texas, El Paso = 4

Low prestige (fictitious): Tri-Valley Institute of Growth and Understanding = 2.5, Northern Plains Center for Human Potential = 2

If these results are representative, the authors' low-prestige places were indeed very low. There may be a status step function. For instance, below 3 bias may increase greatly.

Even if more complete studies moderate the P & C effect, several conclusions seem appropriate. The editorial process has tended to be run as an informal, old-boy network which has excluded minorities, women, younger researchers, and those from lower-prestige institutions. It is time we used our methodological skills to ensure the validity of the editorial process. It should be the responsibility of each journal to ensure the timeliness and quality of the review process. Editorial positions "pay" enough in prestige and influence on the field so that editors must be willing to invest more time helping authors produce better work.

Without space to elaborate here, I conclude that the review system can be that worst of all worlds, where both sides feel abused by the other. Editors and reviewers can feel unappreciated and put upon by careless and incompetent authors. Authors can feel that they're dealing with hostile gatekeepers whose goal is to keep out manuscripts on picky grounds rather than let in the best work. A publication can be worth a great deal to an author (e.g., Tuckman & Leaky 1975). An 85% rejection rate puts the editor in a high power position and inevitably produces pressure to find reasons for rejecting an article rather than spending time helping the best get published. Where previously I had thought blind reviews were a good idea in principle, but not very important in practice, I think what we all have to assume from P & C's provocative paper is that blind reviews should be used unless evidence to the contrary is forthcoming.

Since I have been involved in talking about the ethics of research (Diener & Crandall 1978), I suppose a note is in order acknowledging that sending papers out under false pretenses is probably technically unethical. However, I think that this problem is important enough that a larger and more comprehensive study is not only justified but should be sponsored by journals. It is clear from the shocking results of this study that we need to look more carefully at this area, and a little deception toward journals and our review process may be more than in order in response to the careless standards and intentional or unintentional biases that seem to be prevalent in the editorial process. It is clear from previous work by others (e.g., McCartney 1973) and from P & C's own review, that we have the material on which to base great improvements in the editorial process. Prior to this point I had thought that these improvements would be ideal but not of great practical import. Now we must face the fact that the quality distinctions between many articles may be so small that all the "small details" may end up causing editorial decisions.

Much of what needs to be done to improve the review process, such as screening reviewers to make sure they are competent before sending them reviews, can be done by editors without any further research. Reviews should also be screened as they're returned, to eliminate gratuitous and irrelevant comments. At least one thorough, expert person should read the paper and the reviews and give that mature, balanced, editorial wisdom about a paper that is the ideal of the review process.

Authorship and manuscript reviewing: The risk of bias

Lois DeBakey
Baylor College of Medicine, Houston, Tex. 77030

Peters & Ceci (P & C) are addressing a recurrent question about manuscript reviewing, and their speculations about bias concur with those of many others. The traditional anonymity of manuscript reviewers has aroused distrust among authors for some time (DeBakey 1976; DeBakey & DeBakey 1976). Thomas Huxley (1900, vol. 1, pp. 97-98) complained that because an established "authority" considered as his own "special preserve" a subject on which Huxley was writing, Huxley would have to "manoeuvre a little to get my poor memoir kept out of his hands" as a reviewer. Wright (1970) referred to reviewers' undue delays and repeated demands for revision as "psycho-political manoeuvres." Some years ago I recommended signed reviews to encourage factual, impartial, and documented evaluations and to eliminate such blanket condemnations as "topic inappropriate," "methodology defective," or "data weak" (DeBakey 1978).

A serious problem in manuscript reviewing is the lack of stringent criteria for the same kind of supporting evidence and documentation in reviews as are required in authors' manuscripts. Reviewers who provide objective evidence might be less reluctant to disclose their identity, since a well-reasoned decision, even if negative, is not likely to cause resentment or prompt questions about the reviewers' dark motives.

As I read P & C's article I found myself asking a number of questions:

1. Did P & C obtain permission from the authors of the published articles used in this study and from the copyright owners of the journals to substitute fictitious authors and affiliations, to make minor changes in the manuscripts, and to resubmit them for evaluation for publication?

2. What is an "ecologically valid study of the journal review system"?

3. Were all the fictitious names common and plausible, or were some unusual and obviously contrived?

4. What kind of "slight" changes were made in the titles, abstracts, and opening paragraphs? These critical parts of the manuscript make an initial favorable or unfavorable impression on editors and reviewers, and "minimal" changes in usage and

structure can sometimes mar substance as well as style. I would like to have seen a sample of "minimal and purely cosmetic [changes] (e.g., changing word or sentence order, substituting synonyms for nontechnical words, and the like)."

5. What effect might the conversion of tables to graphs and vice versa have had on reviewers? When used properly, these synoptic devices serve distinctive purposes and should not be indiscriminately interchanged. A discerning reader would have reacted negatively, for example, if data intended to disclose trends, shapes, or correlations were presented as absolute values in tables.

6. What did P & C discover from the few original reviews that editors were willing to show them? Did the first reviewers label the manuscripts "marginally acceptable" or "outstanding"?

7. What was the reaction of the reviewers who detected the disguise? On what basis was their discovery made? And how did the unsuspecting reviewers react when told the truth? Did any consider themselves victims of entrapment?

One wonders how the results would have been affected:

1. if the selection of journals had been truly random instead of restricted to so-called prestigious periodicals, based on authors' affiliations with high-ranking departments;

2. if any of the original reviewers had received the manuscripts they had reviewed before, but now slightly altered and resubmitted. There were two changes of editors, but how many changes of reviewers? There is no way of knowing, of course, if the second set of reviewers was simply more discriminating than the first set and might have rejected the manuscripts on the initial submission. The skill of the editors in selecting reviewers who were most competent to evaluate particular manuscripts would also affect the results.

Several other points deserve comment. Even when authors' names are removed from manuscripts, experts in the discipline represented by the article often recognize other clues to auctorial identity, including references cited. One wonders why these reviewers detected no such clues. The fictitious academic institutions should also have aroused suspicion; one wonders why the reviewers did not become curious about an institution with which they were unfamiliar and why they did not then seek some information about it. Had they done so, they would have detected the camouflage.

Selecting authors from departments "with the 30 highest productivity rankings" was not intended, I trust, to suggest that their publications were necessarily outstanding, since bibliographic quantity is not necessarily equated with quality. As for using the citation index as a measure of the impact, prestige, validity, or quality of a publication, one must remember (1) that because scientists usually pursue a research subject for years, they are likely to cite their own publications more often than anyone else's, (2) that some authors deliberately omit references to their rivals' work, even when these are undeniably relevant and valid, (3) that every citation is not a positive one or an endorsement, but may be a refutation of a previous publication, and (4) that citations do not necessarily signify the best publications on a subject, but may simply reflect the haphazardness or thoroughness with which an author did his bibliographic research. Finally, it is difficult to evaluate the statistics for a sample that is relatively small and that does not include the resubmission of previously rejected articles by authors from prestigious institutions.

Theoretical implications of failure to detect prepublished submissions

Douglas Lee Eckberg
Faculty of Sociology, University of Tulsa, Tulsa, Okla. 74104

Certainly, Peters & Ceci's (P & C's) is a provocative paper. While the authors deal, primarily, with straightforward issues of bias, its central theme has to do, ultimately, with questions of subjectivity in science. Pushed to an extreme, it might join the literature that treats beliefs in the "progressive nature" of science in the search for "truth" as, at least, open to question. To most scientists, such a stance would amount to apostasy. In psychology and sociology, this questioning has been most closely associated with the work of Thomas Kuhn (1970; 1974), who, however, treats scientific groups as communities of practitioners who share ways of conceptualizing their subject matter, evaluative criteria, and knowledge of one another's work (this is the essense of the "paradigm" concept; see Eckberg & Hill 1979). P & C's findings might challenge even the assumption of community. I wish to comment on theoretical implications of the study with regard to this. Having made the above statements, I must add that P & C's data would be a very weak set to use in making any such judgments. With the possible exception of the anomaly of 100% interrater agreement on the question of acceptance or rejection, their findings are compatible with views of science that stress its community nature *and* that hold it to be generally progressive.

The clearest finding in the paper is of some support for the "Matthew Effect" in science (Merton 1973), a situation in which those who have achieved eminence have further eminence showered upon them in excess of their continuing contributions to their fields. Two manifestations of this are the easier time eminent people have getting work published and the more favorable general reception accorded their work. These are well-documented (most recently by Snizek, Fuhrman & Wood 1981). Here, the important theoretical question has to do with the functional significance of the Matthew Effect for the development of science: That is, does it help or hinder scientific development? In his original article describing the Matthew Effect, Robert Merton argued that it had an overall functional quality, while admitting it could have dysfunctional qualities as well (as in the case of the unknown Gregor Mendel, whose work in biology went unrecognized for years).

P & C's paper raises anew the issue of the functionality of the Matthew Effect, for two reasons. First, if papers that were written by major writers can, on evaluation with the "halo effect" removed, be shown to contain significant errors, then *is* it true that the Matthew Effect helps significant findings achieve publication? Might it not be true that poor work comes to be published, while good work by unknowns is crowded out? (Of course, this assumes a clear criterion of "worth" for research.) Second, if research by significant individuals is utterly unrecognized by reviewers in the writers' fields, then does not the claim that the Matthew Effect helps major work get the reading it deserves (Merton 1973, p. 448) suffer disconfirmation? If, even here, major work is simply filtered out, then how can science clearly be progressive?

The answer to this interrelated set of questions becomes obvious if we take into consideration the paradigm and "invisible college" (Crane 1972) concepts. According to these, a scientific "community" consists, most strongly, of the small number of practitioners in a subspecialty who read and criticize one another's work, who probably read one another's work in manuscript form, and so forth. It is only at the subspecialty level that familiarity with papers is likely to develop. After all, there are a large number of journals (some 179 refereed, English-language journals in psychology alone, and hundreds of others in related fields; see Social Science Citation Index 1977) and thousands of research articles to which one can attend annually.

An implication of this is that only a tiny handful of people in a given discipline will read any given article, and, even within a specialty area, not many will be acquainted with a given work, unless it is one of those that has a truly major impact on a field. Kuhn (1970) indicates that the predominant efforts of most scientists can be described as "mop-up" work - work that

fleshes out a paradigm but is itself not terribly innovative. This explains the failure of editors and referees in P & C's study to recognize the articles submitted. Assuming that only .5 to 1 percent of psychologists - and only a slightly larger percentage of those whose specialty areas are covered by a given journal - can be assumed to have read a given article (see the citations in Merton 1973, p. 448), it is quite unlikely in any given case that any reviewers will be familiar with the work in question. Of course, an implication of this is that one should expect a great deal of redundancy in published research (since the same type of stuff may be published several times), but this does not call into question the generally progressive nature of science.

We can assume the lack of general importance of the articles in question. By "importance," I mean specifically a sense on the part of people in the field that the work has major implications for the development of a subspecialty or specialty area, or that an article is perceived as very controversial. From data provided in P & C's paper, we can assume that the resubmitted articles had been cited once or twice each. While this may be statistically a greater number of citations than an "average" article gets, it certainly does not indicate that a work has set the discipline on fire. Hence, there is little reason to suppose that any given psychologist should be aware of it. I make this point - that these articles are not important - explicit, because it can help us to understand why such a high proportion of the resubmitted pieces were rejected. I assume that there are a large number of articles "out there," most of which are not in any sense major, but most of which would have something to say to people in a specialty area, were space to permit. Researchers, especially at major research schools, cannot wait for their work to be perfect to send it off; pressures to publish quickly or lose their positions ensure that work will be sent out without delay. It may be here that the Matthew Effect operates to the clear advantage of major authors, a point that P & C mention (see their note 6), and which bears following up. In any event, the point is that even most work by "major" departments is not "important," though it may be decent, so that a decision to accept or reject one of many "minor" pieces might easily hinge on halo characteristics. Hence, the Matthew Effect helps those who "have," and hurts those who "have not," but may not affect the general quality of scientific literature.

In summary, the data presented by P & C really are not terribly surprising, given our understanding of the way science operates. Neither do they bear on the issues of the quality of science, or its theoretically progressive nature, though they do bear on the interesting questions of *who* shall partake of the reward system of science, and of the legitimacy of a stratified distribution of rewards.

Deception in the study of the peer-review process

Joseph L. Fleiss
Division of Biostatistics, Columbia University School of Public Health, New York, N.Y. 10032

I shall leave it to the other commentators to address the important and disquieting findings reported in the article by Peters & Ceci (P & C) and shall instead address what I see as serious ethical problems in their study. In my opinion, the authors violated at least six of the American Psychological Association's (1973) 10 ethical principles for research with human subjects. As a result, the study should not have been undertaken as designed. Given that it was, its results should not have been published.

Item: The actual authors of the 12 articles apparently did not give their consent to have the articles used in this study. At least, there is no indication that informed consent was obtained. Given the risks of embarrassment and misunderstanding that the original authors were subjected to (see the final item below), not obtaining consent was inexcusable.

Item: P & C may have violated United States copyright law, inasmuch as permission apparently was not granted them by the publishers of the 12 journals to have the papers reproduced.

Item: The editors, associate editors, and referees were deceived and abused. Participation in the peer-review process is a voluntary professional activity, with participants believing that the time they invest and the critical skills they bring to bear are for the purpose of judging the suitability of a submitted paper for publication. The 38 individuals who served as editors and reviewers of the 12 articles were, instead, unwitting participants in an experiment.

Item: A few of the editors may have acted unethically in providing P & C with copies of reviews from some of the articles' original submissions. In their Discussion section, P & C distinguish between uncooperative, resistant editors and cooperative, "gracious" editors. A more accurate distinction would be between editors who respect the implicit understanding reviewers have that their comments and recommendations will be made available only to the editors and authors, and editors who do not.

Item: Some of the original authors might be identifiable, given the quotations from several of the critical comments made by the reviewers of the resubmissions. For example, the comment that "players had the ability to get different numbers of markers to the goal, [but] the game... was such that each player had only one marker" may be sufficient to identify the original article and its authors. Forums exist for the open criticism of one scientist's work by another, and for rejoinder. Anonymous criticisms of a scientist's work now appear in print; to whom, and how, may that scientist respond?

P & C do not attempt to defend their ethically questionable procedures and methods; indeed, they nowhere point out that they practiced deception on the 38 editors and reviewers. As Weiss (1980) has stated,

> Deception is morally hard to justify, even or especially in the "pursuit of the truth."... if one rationalizes deception for the purposes of research, inevitably it can and will be rationalized for other purposes.... I know of no research involving deception in which the results could not have been obtained without [its] use.

This final point may be made specific to the study of peer review. Journal editors are in the position to superimpose on their routine peer-review practices a controlled study of the effects of some components of the process. Consider, for example, the editor of a journal that relies on nonblind refereeing by two reviewers. He might design as follows a randomized study of the effects of three factors on the fate of papers submitted for publication: the status of the senior author (perhaps dichotomized on the basis of the prestige of his institution), the status of the reviewer (dichotomized similarly), and blind versus nonblind review.

The editor would first stratify the papers by the status of the senior author and would then, within each stratum, randomly assign a paper to receive one of two pairs of reviews. One pair calls for a "high status" referee to conduct a blind review and for a "low status" referee to conduct a nonblind review. The other pair calls for the reverse. If the papers within each stratum are paired, a Latin square design can be used. As shown, for example, by Winer (1971, chapter 9), the data may be analyzed to measure the main effects of each of the three factors on the reviewers' recommendations, and of interactions involving the status of the author.

Researchers studying human subjects have, with ingenuity but honesty, successfully overcome many of the constraints imposed by institutional review boards to ensure trust, confidentiality, and privacy. Similarly, students of the peer-review

process will successfully overcome the constraints I hope will evolve in reaction to the abuses perpetrated in this study.

Review bias: Positive or negative, good or bad?

Russell G. Geen
Department of Psychology, University of Missouri, Columbia, Mo. 65211

Most people who have done much publishing, reviewing, or editing have at some time been convinced that bias plays a part in the peer-review process. Demonstrating such bias is another matter, however, and Peters & Ceci (P & C) correctly point out how little hard evidence we have to support our suspicions. Their own study is a step in the right direction. The methodology seems to be basically sound (given the natural constraints on this sort of research) and the results clear. A few possible problems should be mentioned, however, before we begin to reassess the review process on the basis of these findings. Although I am largely in agreement with the authors' purpose in doing the study, and partly so with the conclusions they draw, I would nevertheless suggest that (a) the data reported in Table 1 may be somewhat inflated due to a confounding factor; (b) granted that response bias was found, we cannot be sure of its direction; and (c) even if the response bias is in the direction proposed by the authors, it may not necessarily be something that we want to eliminate entirely.

The review process. No mention is made of any cases in which the persons who reviewed a resubmitted manuscript were the same ones who reviewed it originally. Judging by the small number of recognitions, I would guess that such cases, if they existed at all, were rare. Thus we may assume that the second reviewers were, in most or all cases, not the people who made the original decision to accept. This change in reviewers is confounded with the change in manuscript authorship which is the main independent variable. Thus, part of the impressive effect shown in Table 1 may be spurious. Lacking any data on reviewers, we cannot estimate the extent of this inflation. What is there in a change in reviewers that might account for at least some of the variance in Table 1? One possibility may be considered. We are told that the papers were resubmitted two to three years after they had been published. Allowing (conservatively) for a lag of six to nine months between original review and publication, the review came between 2.5 and 3.5 years after the original review. In that span of time the personnel in any area of research could include considerable numbers of young investigators not involved in the original reviews. It is my impression that in some fields, such as social and clinical psychology, people who have been out of graduate school two or three years are assuming more and more of the duties of reviewing. Frequently they are willing to review papers for which their senior colleagues do not have time. It is also my distinct feeling that these young people are tougher and more critical reviewers than their elders.

Direction of the bias. Despite the argument stated above, let us still conclude that response bias is shown in Table 1 (as I think it is). The question I would raise is whether it is a "positive" bias, which favors the famous people and prestigious departments, or a negative one against places bearing the names that appeared on the resubmitted papers. The data strongly suggest a negative bias in the second review process as well as a positive one in the first. Otherwise, how could we account for the near unanimity among reviewers in rejecting the papers on resubmission? P & C draw the same conclusion. It would appear that the negative bias could be just as strong as the positive one, and I think that it makes a great deal of difference which of the two we are talking about. Given the number of institutes for "growth" and "potential" in existence these days and the standards that such places usually represent, I would probably be as likely as the next person to form a critical bias against a paper bearing such a designation. I also think that any error in a conservative direction (rejecting a good paper on the basis of such a bias) is less harmful to science than allowing a bad paper to be enshrined in the archives. If we are allowing a lot of bad research to get published just because it comes from Stanford or Wisconsin, we are, I think, more justified in seeking reform in the system than if we are rejecting good papers because they come from places with funny-sounding names. What this study really needed was something analogous to a zero-treatment control in which papers were resubmitted bearing the names, not of exotic-sounding institutes, but of less prestigious, but credible, colleges and universities. Such institutions would probably not raise spurious negative bias to such an extent, and would thus serve as a control for any positive bias attached to papers from the big institutions.

Positive bias: Is it bad? If we assume that the study shows mainly positive bias for prestigious schools and researchers, I would wonder whether such bias is *always* inimical to the interests of good reviewing. Individuals reporting a study from Stanford, for instance, hold their appointments at that school because in all probability they have demonstrable ability and a record of good research. A reviewer may be justified in assuming at the outset that such people know what they are doing. It is widely recognized in most areas of science that when writing a technical report one does not necessarily report every procedure that was undertaken in the laboratory. Colleagues in the field simply know that certain things are done as standard operations even though they are not always reported. They know it, that is, when the work was done by someone they recognize and respect. It is not always possible to make the same assumption in the case of unknown colleagues, and hence the latter are apt to be judged more closely on what they actually describe.

R. G. Geen is editor of Journal of Research in Personality. *Ed.*

The journal article review process as a game of chance

Norval D. Glenn
Department of Sociology, University of Texas, Austin, Tex. 78712

The Peters & Ceci (P & C) findings clearly indicate that, in the case of the journals studied, the review process has not worked as it is supposed to work; but the reasons for the troublesome findings are not very clear. The authors seem to favor the explanation that there was a systematic "status" bias for the papers on the first submission and against them on the second submission, and the evidence is indeed consistent with that explanation. Lack of blind reviewing, then, could account for the deficiencies in the review process of these journals.

I have a strong suspicion, however, that systematic status bias was not the only culprit, or even the most important one. Rather, the findings are quite consistent with my long-standing impression of the capriciousness of the review process in my own discipline, which is quite similar to that in psychology, except that almost all of our journals use blind reviewing to reduce systematic bias. It seems likely to me that most of P & C's manuscripts were rejected (although they had all been accepted previously) because even the best papers submitted to the journals studied had a low probability of acceptance and because, except for papers conspicuously poor according to generally accepted criteria, the outcome of submissions depended largely on luck. Some elaboration of this explanation is in order.

Suppose that (as I suspect is the case) most referees for the

journals studied are predisposed toward making negative evaluations of the papers they read, both because they feel that being able to find flaws is a measure of their competence and because most of them are among the competitors for the scarce space in the journals and thus have a self-interest in making negative recommendations. Referees predisposed toward finding flaws can usually find them because (a) there are few, if any, flawless papers, (b) there is much less than perfect consensus on standards of quality in psychology, and (c) the referees (assuming that they are similar to those for sociology journals) are often undeniably wrong in their criticisms, usually, perhaps, because of haste and carelessness, combined with an eagerness to find fault. Thus, if an editor decides to accept or reject simply by tabulating the recommendations of the referees, as is often the case, the probability of acceptance will always be low. Acceptance will depend on drawing referees who do not have the usual negative predispositions or who, for some reason, are positively biased toward the paper (e.g., because their work is favorably cited in it).

The rejection rates of the journals studied by P & C were around 80 percent. Suppose that a fourth of the rejected papers were so conspicuously deficient that almost any reasonably well trained referee in the specialty would recommend rejection. For all other papers, considered together, the probability of acceptance would be around .25, and if the other papers did not vary substantially in their probability of acceptance, as I suspect was the case, the expected proportion of acceptances among the nine undetected papers resubmitted by P & C would have been .22, or two of the nine. In fact, the proportion was .11, or one of the nine, but the real proportion is not significantly different from the expected proportion at the conventional .05 level. Of course, the .25 probability of acceptance for reasonably good papers is only a guess, but the exercise illustrates how the P & C findings can be explained without resorting to status bias if one works from certain undemonstrated but not implausible assumptions.

Whatever the correct explanation of the P & C findings may be, they and similar evidence clearly indicate a need for improvement in the review process of academic journals. It would help if editors would truly be editors rather than clerks who tabulate referees' recommendations and let the "vote" decide the outcome of submissions. Editors should evaluate the performance of referees, making proper adjustments for the varying strictness of the different ones, and should make an honest effort to achieve consistency in the standards used to judge different papers. It would be immensely useful for editors to send referees' comments to authors for their counterarguments *before* deciding to accept or reject. And if editors are to make truly intelligent decisions, their workloads must not be too great. I doubt that many editors can adequately handle more than 150 to 175 submissions per year, and journals with submission rates well above that level should have more than one editor making decisions on papers.

If substantial improvements are not made, we should stop pretending that the review process is something that it is not. The letters of rejection from which P & C quote are polite in tone but also rather dogmatic and condescending, and the editors do not come close to admitting that the review process is highly arbitrary. Editors, department heads, deans, and other concerned persons should acknowledge the weaknesses in the review process and act accordingly. For editors, that would mean letters of rejection with a somewhat different tone. The following letter would have to be modified to meet the specifics of each case, but it illustrates the kind of letter I think would usually be appropriate:

> Dear Professor _____:
>
> We have completed consideration of your manuscript entitled _____, and I have decided not to accept it for publication. My decision was based partly on the advice of three referees, whose comments are enclosed. Most of their criticisms concern issues on which reasonable people may disagree. However, I agree that _____. On the other hand, I disagree with the referees concerning _____.
>
> In the last analysis, my decision to reject was based on my judgment that among our recent submissions other papers were more deserving of the scarce space in this journal. Other editors might well disagree. I of course do not pretend that we have subjected your paper to a definitive evaluation.
>
> I hope you will give us a chance to look at papers you write in the future, and I wish you success in placing a version of this paper in another journal.

N. D. Glenn is former editor of Contemporary Sociology. *Ed.*

When will the editors start to edit?

Leonard D. Goodstein
University Associates, Inc., San Diego, Calif. 92064

There is little question in my mind that Peters & Ceci (P & C) have addressed an important current problem - the low reliability of peer review by journal referees. Recognizing that they have generalized from a small sample and that they too might be criticized for "serious methodological flaws," I, however, personally have little doubt that their findings can be replicated by other investigators with other, larger samples.

Given the fact that we are asking our peers to make important decisions, using complex and ambiguous criteria, it should come as no surprise that the interjudge agreement of these decisions is low, more or less bordering on chance. If we were conducting an experiment and found our judges behaving in such a fashion, the solution to the problem would be readily apparent - develop behaviorally anchored criteria and train the judges against these criteria. We would allow no judgment to be used in the experiment until the reliability of the judgments reached an acceptable level of confidence.

Since we know the solution to our problem, why does the problem remain with us? For me, there is a rather simplistic set of answers to this question. Neither journal editors nor editorial referees are chosen for their editorial wisdom, their sagacity, or their willingness to work hard at the editorial task. Rather, they tend to be chosen for their research competence, their political connections, or their need for visibility. While these persons may be well intentioned, I know of no journal, in any discipline, that requires, expects, or even encourages the kinds of training that are likely to produce the necessary levels of judgmental reliability.

I would argue that those most loudly damned by the findings reported by P & C are the journal editors themselves, not the reviewers. It is the editors who are responsible for monitoring their reviewers' reliability, not the reviewers. Indeed, it seems inconceivable to me that only three of the editors detected the resubmission. I would like to believe that, after serving six years as editor of the *Journal of Applied Behavioral Science*, I would not have been so unaware of the contents of the journal that bore my name as editor.

But I have at least one personal anecdote that suggests that other editors might have a rather different view of their task. I recently submitted an article to a professional journal and received a routine acknowledgment. When no decision about the article was forthcoming after six months, I wrote the editor. Without apology, he replied that the manuscript had been rejected and enclosed comments by a consulting editor. As the only comment referred to the opening and closing paragraphs, it was clear that the reviewer had indeed not read the body of the paper. I again wrote to the editor, pleading that I had been done an injustice. The second letter from the editor, again without apology, enclosed a second, favorable evaluation from a second consultant which had "arrived after I first wrote you."

In no case did I ever have the feeling that the editor himself had ever read my original manuscript, nor was he prepared to deal with either the substance of the paper or the consultant's criticism. While I may be accused of generalizing from a single instance, I believe that such behavior is much too common among journal editors.

The role of an editor of a major professional journal is an important, perhaps critical, one. Yet I have serious reservations about the manner in which editors are chosen and the nature of their assignment. Since I have already briefly addressed the first of these concerns earlier in this commentary, let me comment on the second. Virtually all of the more prestigious journals have an astronomical number of submissions, and an acceptance rate of 20 percent or less. The role of the typical editor seems to have degenerated into that of a traffic cop, sorting out which manuscripts to send to which reviewers, collating the reviews, and returning them to the author. It is the rare editor who even bothers to read the reviews carefully in order to give the author(s) some hints on how to resolve the conflicting, even contradictory, recommendations about revision.

While I find myself in general agreement with the position of P & C, I would argue that they have simply not gone far enough. Our journals will continue to have serious problems with content decisions until we insist that editors assume the responsibility that is theirs alone. This requires a commitment of time and energy that can only be attained when those responsible for choosing editors take their responsibility, in turn, equally seriously.

L. D. Goodstein was editor of the Journal of Applied Behavioral Science *for six years. Ed.*

Cognitive relativism and peer-review bias

M. D. Gordon
Primary Communications Research Centre, University of Leicester, Leicester LE1 7RH, England

Peters & Ceci's (P & C's) original and somewhat cheeky methodology has certainly produced findings that throw valuable light on the refereeing process. However, this value does not lie so much in exposing systematic bias (which has been done in many previous studies cited by the authors) or in determining its extent (since the study is too small, and it is difficult to disentangle the extent to which referees' evaluations are biased from the degree to which they are random). While thus adding little to our knowledge in these respects, the value of the findings lies in helping to refine our understanding of the nature of bias, and this point, somewhat surprisingly, is not fully recognised by the authors. Indeed, their discussion of the nature and "mechanism" of bias is speculative and does not draw directly on their data. This is an oversight that appears to derive from a concern with arguing what the nature of the bias is, rather than using their data to show what it is not. This can be illustrated by briefly describing two previous studies.

The first (Crane 1967), examined U.S. social science journals and found that as the proportion of referees from particular groups of institutions increased, so did the proportion of successful authors from those groups. Two possible interpretations were offered:

a. As a result of academic training, editorial readers tend to respond to certain aspects of methodology, theoretical orientation, and mode of expression in the writings of those who have received similar training.

b. Doctoral training and academic affiliations influence personal ties between scientists which in turn influence their evaluation of specific scientific work. Since most scientific writing is terse, knowledge of details may influence the reader's response to an article.

Crane found that her data supported the former mode of interpretation rather than the latter; this implied that a notion of cognitive relativism was needed to account for referee bias.

A similar mode of interpretation was also invoked in a study of U.K. physical scientists (M. D. Gordon 1980). When they were split into those affiliated with major universities and those affiliated with minor ones, significantly higher frequencies of favourable evaluation were found when author and referee shared membership of institutional groups ($p < 10^{-6}$). These findings were attributed

> primarily to there being higher levels of consensus on research beliefs within these [institutional] groups than across them. Personal ties and extra-scientific preferences and prejudices might, of course, be playing a part as well. But it appears that even in the absence of these personal factors, the scientific predispositions of referees still bias them toward less critical evaluation of colleagues who come from similar institutional or national groups, and so share to a greater extent sets of beliefs on what constitutes good research. (pp. 273-75)

This mode of interpretation of patterns of bias is implict or explicit in other studies cited by P & C. It is easily accommodated within relativist sociologies of science (see, e.g., Collins 1981) while not being inconsistent with the notion of a scientific community endeavouring to live up to the standards of social conduct described by Merton's norm of "universalism" (Merton 1968b). However, this mode of interpretation cannot be invoked to account for the bias identified in P & C's study. For the bias they detect can be assumed to derive solely from the perception of authors' status and credibility, as judged from their name and institution.

The value of P & C's study therefore lies in showing that bias remains when cognitive relativism cannot be assumed to be having any systematic effect. It remains to be seen whether these findings would be replicated in future studies of other disciplines.

Optional published refereeing

R. A. Gordon
Physics Laboratory I, Technical University of Denmark, DK-2800 Lyngby, Denmark

Peters & Ceci's (P & C's) target article draws sharply focused attention to major inconsistencies or errors that sometimes arise in the refereeing process. Similar inconsistencies or errors have been noted before (e.g. Ruderfer 1980) and have motivated a large number of proposals for changes in the refereeing system. It is not, however, the refereeing system per se that is at fault but rather the undue weight that is sometimes attached to the occasional poor or incorrect referee report, leading in the worst cases to the unjustified rejection (or acceptance) of a submitted manuscript. Unfortunately, few if any of the proposed or already existing variations of the refereeing system are well-suited to the elimination of this single major defect of the refereeing system - either because they would be ineffective in reducing the incidence of poor or incorrect referee reports or because they would seriously compromise the value of the refereeing system by placing too much or too little weight on referee reports. Thus, changes of a more peripheral nature, such as the introduction of double-blind refereeing (Benwell 1979), or the automatic publication of the abstracts of all submitted manuscripts (Carta 1978), would not, in themselves, necessarily reduce the incidence of poor or incorrect referee reports or eliminate the worst effects of such reports. On the other hand, refereeing systems that rigidly attempt to eliminate all manuscripts that might otherwise be unfairly accepted solely or primarily on the basis of referee reports, serve only to place a disproportionate weight on precisely those inconsistencies or errors that must inevita-

bly occur with a less than 100% infallible review system. In addition, as the work of P & C strongly suggests, a high rejection rate does not in itself provide any guarantee that the accepted manuscripts are measurably better than the rejected ones. At the opposite limit, refereeing systems that rigidly attempt to accept all manuscripts that might otherwise be unfairly rejected by adopting measures that in practice effectively ensure a very low rejection rate (e.g. the removal of referee anonymity, Robertsen 1976, the inclusion of a published author rebuttal to all referee comments, Kumar 1979, or the creation of new specialized journals or letters sections, Lazarus 1980), unnecessarily compromise the effectiveness and inherent real value of the refereeing system and lead to a reduction in the general estimation of the value of the work that is accepted for publication.

Journals could, however, readily eliminate the undue weight sometimes attached to poor referee reports without compromising the effectiveness of the refereeing system by adopting a simple system of optional published refereeing (R. A. Gordon 1980). This system would give authors the option of (and responsibility for) publishing a manuscript provided it was accompanied by the anonymous unanswered comments of the referee where relevant. Such a system would not only leave the responsibility of publishing with the authors, to whom it must ultimately belong, but would be more consistent with the only real, and in fact the only truly realizable, goal of the refereeing processs, namely to provide as effective an evaluation of a submitted manuscript as is practical, but without any implication of infallibility on the part of referees, journals, or authors. In practice, such a system of optional published refereeing would have a significant advantage over other proposed changes in the refereeing system, in that it could readily be implemented without any changes in existing refereeing procedures. In cases in which the referees were in substantial agreement with the manuscript, the manuscript would simply be published as under present refereeing procedures. Journals wishing to provide a minor, noninhibiting degree of referee accountability could require that the names of the referees be published at the end of the accepted manuscript, although authors would, of course, still bear the major responsibility for the manuscript.

In cases in which major fundamental points of disagreement remained after the usual manuscript revisions and exchanges between referees and authors presently allowed by most refereeing systems, the journal would simply give authors the option of publishing the manuscript together with the anonymous comments of the referee without any rebuttal by the authors. This would allow the contested points to be brought to the attention of the interested scientific public (which is in fact the only authority that can ever resolve such fundamental points of disagreement) in a form that would direct attention to precisely those points. In such cases, referee anonymity would be absolutely essential, since it would permit referees to express their arguments freely, without fear of unnecessary personal controversy, and it would force any subsequent discussion to concentrate on the disputed points without being distracted by the identity of the referee. Journals wishing to place special emphasis on the seriousness of optional published refereeing could restrict the publication of contested manuscripts in many ways, such as by regulating the subject matter allowed, the frequency of publication, the form of the referees' comments, and so on. In most cases, however, such extra precautions would be unnecessary since few authors, whether from well-known institutions or not, would lightly decide to publish a manuscript accompanied by a thorough, carefully reasoned criticism or reservation.

In summary, inconsistencies or errors in the refereeing process are inevitable since present refereeing systems must occasionally bias decisions for the acceptance or rejection of manuscripts in favor of authors or referees, none of whom can ever be considered to be even approximately infallible. Furthermore, formal attempts to eliminate presumed bias or halo effects (Adair 1981), apart from being extremely difficult or impossible to control in practice, are in fact completely irrelevant to the essential basis of peer review, namely the actual content of the referee's report, regardless of presumed bias and the impossibility of any completely certain resolution of fundamental author-referee disagreements without the participation of the interested scientific public. A simple system of optional published refereeing could readily eliminate the worst features of present refereeing systems without changing present refereeing procedures; it merits a realistic trial period by journals in both the social and the physical sciences.

It is a useful exercise to contrast this commentator's proposal - "optional published refereeing" - with the open peer commentary service provided by this journal. Optional published refereeing would involve the publication of (some) unfavorably reviewed manuscripts, along with the (anonymous) negative referee reports, without rebuttal from the author. (In the editorial notes following the commentary of J. S. Armstrong I point out that several variants of this practice have occasionally been implemented.) Open peer commentary, on the other hand, is only accorded to refereed, accepted articles, with commentators identified and authors formally responding. The system is viewed as a complement *to peer review, not an alternative for it.*

It seems to be an empirical question whether publication quality control - really a filtering system to increase scientists' and scholars' confidence in what they read - and the advancement of research and knowledge would be better served by publishing questionable material together with dissenting critiques or by rejecting it altogether. Put otherwise, the real question seems to be whether editorial judgment *can be given a sufficiently reliable basis (by strengthening the peer-review practices under discussion in this issue) to be validly exercised on the reader's behalf, or whether the reader will have to exercise this judgment on his own, without the help of gatekeepers. But even if the empirical answer turned out to be the latter, certain kinds of questions would necessarily continue to call for editorial judgment, namely,* how *marginal a paper would one still be willing to publish, and when with, when without dissent?*

BBS has explicitly opted for the other end of the quality spectrum (except in a few special cases), reserving the elaborate service of international, interdisciplinary Commentary only for the best and most important work in the field, as judged by rigorous prior refereeing. Ed.

Judging document content versus social functions of refereeing: Possible and impossible tasks

Belver C. Griffith
School of Library and Information Science, Drexel University, Philadelphia, Pa. 19104

Journal refereeing: The present study and its predecessors. Clearly Peters & Ceci's (P & C's) study makes the point, perhaps inadvertently, that research on editorial refereeing is difficult to do and that payoffs in both original data and convincing evidence of their validity are low. Their data, in entirety, are several judgments, apiece, of only nine documents; the age of the reported research, the selection of high-prestige authors' and institutions' papers as guinea pigs (therefore, less rigorously reviewed on initial submission?), and the use of fictitious, and not, to my mind, innocuous institutional names are all confounded. In contrast to P & C, I am encouraged that 25% of the papers were detected as having

already been published. I have the nagging suspicion that such "remembered papers" are probably the only papers with substantial scientific impact.

A variety of problems must, almost necessarily, prevent perfect control in research on journal refereeing; a scientific paper is a complex intellectual product whose meaning and value derive from a complex, highly dynamic intellectual environment. To this is added a variable, and occasionally casually managed, social system of the editor(s), referees, and publishers. Numerous previous studies, usually small in scale, either of refereeing or of judgments of scientific "quality," have encountered similar problems; however, they all support the hypothesis that reputable scientists vary greatly, among themselves, in judging the quality or acceptability for publication of individual scientific papers. (The earliest reasonably good study is Orr and Kassab 1965. A carefully done investigation of the judgment of quality, Virgo 1974, leads one to believe that judges share, at best, only about 50% of variance.)

How closely should referees agree? The basic question is, To what extent do scientists agree in evaluating the contents of a scientific document? This question is variously approached: Acceptability for publication? Quality? Relevance to scientists' current scientific work? The last has been intensively and systematically studied because of its importance to the practical problem of determining whether an information system furnishes documents that are of use to scientists (see Saracevic 1975 for an excellent review). This kind of research is perhaps the most straightforward and simple of all investigations of scientists' judgments of document content; it shows that the judgment of a document's relevance to a scientist's own work is unstable. If a scientist cannot agree with himself on whether the content of a document relates to his own work, is it surprising that he cannot agree with others on the quality of the document?

The role of refereeing in scientific communication. Garvey and Griffith (1962) early argued, on the basis of studies of psychology, that scientific communication in a discipline is a system in which any component's function relies on other components and affects other components, often indirectly. The performance of a burly bouncer in a saloon cannot be evaluated on the basis of the absolute number of unruly patrons he ejects. Similarly, there is reason to believe that the existence of a refereeing system, like the presence of the bouncer, has much more of an effect than the referees' direct interaction with authors. Strong indirect evidence of such an effect runs through the Garvey and Griffith research on scientific communication in psychology. They found a plethora of forms of unpublished research reports; perhaps fortunately, only a minority of any form was actually presented for journal publication (see Garvey & Griffith 1971; American Psychological Association 1963-1969).

Another feature, one bearing on the overall effectiveness of communication, is that referees make significant contributions to the quality of reported work. Korten and Griffith's (ca. 1970) intensive unpublished study found vivid examples of the importance of referees' comments to authors. The most extreme, a direct quote from one respondent, follows:

> "I have had two articles accepted by the *Journal of Experimental Psychology* with no reviews sent to me and no revisions suggested. I dislike that. No article is that good. I want reviews so that I can make improvements."

Last, with regard to the role of the journal referee in the system of scientific communication, there is the social function emphasized by Zuckerman and Merton (1973) of "certifying" knowledge. I believe that as readers we always approach with misgiving the several loosely refereed journals in the social and behavioral sciences, even when the articles seem directly pertinent.

The puzzle of social and behavioral sciences' literatures. The bouncer's tossing of a patron from the saloon represents something of a failure of socialization. Could we regard manuscript rejection as a similar failure in scientific socialization? That is, are not authors and referees in the same scientific community; should they not share the same standards? Rejection rates in the range of 70-85%, as reported by Zuckerman and Merton (1973) and by P & C, seem to represent a great deal of fundamental disagreement.

Several writers have raised a variety of complex questions regarding the effectiveness of communication mechanisms in the social and behavioral sciences. (Defects in the scientific communication system are discussed by Garvey, Lin, and Nelson (1970); defects in the use of earlier literatures are discussed by Price 1970 and by Griffith and Small 1976; the last are particularly concerned about the inability of the social and behavioral sciences to purge older material.) High rejection rates, and the community's tolerance of those rates, are, in my view, another symptom of the low value placed on the literature of the social and behavioral sciences. The first parts of my commentary argue that low interjudge reliability probably cannot be changed; the system becomes a strange one when such low reliability is coupled with very high levels of rejection. Reluctantly, one must conclude that the contents of prestigious behavioral science journals are largely chance determined, which cannot result in a high-quality product, and which raises questions as to what fails, in such a system, ever to see the light of day.

Scientific communication: So where do we go from here?

James Hartley
Department of Psychology, University of Keele, Staffordshire, ST5 5BG, England

Peters & Ceci's (P & C's) paper should be warmly welcomed. Findings such as these force people to comment, to argue, and to defend the present publishing system - or at least to discuss how it might be improved.

It seems that unreliability in peer-review judgment is inevitable and perhaps even desirable. There are no firm yardsticks for evaluating the worth of other people's contributions. Editors (and referees) are busy people. They work unpaid and give up large amounts of precious time. Unreliability arises unintentionally and inevitably, as it does in the marking of examination scripts. Occasionally there may be blatant discrimination (and even fraud), but in general most people feel that the system operates reasonably well, relying as it does on the good faith of all concerned.

Nonetheless, few would dispute that there is room for improvement. It is surprising how little is known about different editorial practices for different journals. This variety is not discussed by P & C but is well illustrated by M. D. Gordon (1980). More discussion among editors from different disciplines would surely lead to some improvements. The good features of some journals (such as the editor sending each referee's report to the author *and* to the other referees) could be more widely used. Other improvements are discussed by P & C.

The key question that arises is whether or not the current anonymity in the system is necessary. There is a place for someone adventurous to start an open journal, one that names its referees, and one that might include the comments of named referees with the accepted papers. If such a journal were to be started, we could at least find out what the difficulties are in practice as opposed to speculation. [*See commentaries of J. S. Armstrong, C. Belshaw, M. D. Gordon and editorial notes passim. Ed.*]

Such an approach is unlikely to cause much change in the format of most journals as we already know them. But other

methods are available and should be explored. The Open Peer Commentary system used in this journal is one of these. It works by capitalising on the inconsistency between commentators. There is more scope for journals that publish only indexes or abstracts: Interested readers can contact the authors for the details if they want them. Similarly, electronic journals and computer-based retrieval systems can provide details - from the computer or the author - on request. New technologies such as these may lead to new formats for journal papers. Findings, for example, may be summarised in a one-page "information map" (following Horn 1980). In principle there is no need for any of this material to be refereed, although undoubtedly most of it will be in practice. Refereeing is likely to preserve certain standards (whatever P & C say) and to hold back the literature at the floodgates a little longer.

The rejection of a journal article, while painful in itself, is not the end of the road for the author. Articles can always be resubmitted elsewhere. Indeed, one implication of P & C's paper (which I am ashamed to say that I have tested with success) is that if an article is rejected then it can be revised and resubmitted to the same journal some two or three years later. (In my own case I waited until the first editor had retired.)

It is more conventional, however, to resubmit one's work elsewhere. If the article is good, someone somewhere will publish it. The news will be picked up on the grapevine and passed along (as in Garcia's case; see Revusky 1977). Here new abstracting services and publications such as *Current Contents* are a boon to present-day authors. Such journals not only draw attention to literature in your field, they also draw other people's attention to your work.

P & C conclude their article with a call for more research into peer reviewing. I would like to recommend a wider focus. There is a need for more research into the whole process of writing and publishing scientific papers - of which peer reviewing is but a part. To be successful this research will demand a variety of strategies. Despite my enthusiasm for their paper, I personally would not wish to associate myself with their methodology.

The insufficiencies of methodological inadequacy

Robert Hogan
Department of Psychology, Johns Hopkins University, Baltimore, Md. 21218

The paper by Peters & Ceci (P & C) is an interesting and unflattering analysis of the journal review process. I doubt, however, that it will have much impact on the behavior of the journal system - largely because the lesson to be learned from the paper is hard to determine. The following, briefly, are my views on what the paper may mean. It makes three points: (1) editors and reviewers failed to recognize manuscripts that had been previously published; (2) most of these previously published papers were rejected the second time around; and (3) the papers were rejected largely on methodological grounds.

With regard to the first point, it is not surprising that the editors didn't recognize the papers. As an editor, I usually recognize redundant papers in my particular specialty, but there are areas in psychology where all the manuscripts sound alike to me. I am a bit surprised that the reviewers didn't notice the redundancy. That probably indicates that, like everyone else, the reviewers were suffering from information overload.

With regard to the second point, it may mean that reviewers have a bias against unknown authors from obscure institutions. The more plausible explanation, in my judgment, is that the publication process is a random walk. This in turn suggests that (a) the key to success is to do a lot of research and submit a lot of papers; and (b) the better journals do not necessarily have the highest rejection rates.

The third point is the most disturbing. I have never seen a piece of psychological research that could not be faulted on methodological grounds. This means that methodological inadequacy is always a matter of degree. The P & C paper suggests, in addition, that it is the wastebasket category into which manuscripts are sorted when no other grounds for rejection can be found. Academic psychology seems peculiarly prone to what medieval scholars called the fallacy of dogmatic methodism - that is, when a problem is analyzed by the proper method, truth will somehow inevitably emerge. Methodological rigor won't save us; actually nothing will. But better conceptual analysis will improve the quality of the life of the mind for everyone, and may even promote progress in the psychological sciences.

R. Hogan is editor of the Journal of Personality and Social Psychology. *Ed.*

Peer review in the physical sciences: An editor's view

William M. Honig
Western Australian Institute of Technology, S. Bentley, 6102, Western Australia

My comments on the Peters & Ceci (P & C) paper are in three parts: (1) my background in this field; (2) direct comment on the P & C paper; (3) general comments relevant to the physical sciences.

I founded *Speculations in Science and Technology* in 1978 as a journal devoted to speculative papers in the physical sciences, engineering, mathematics, and the biological sciences (all the hard sciences) with a standard review procedure for all papers. About 60% of accepted papers are from authors of orthodox backgrounds (universities, research laboratories) and 40% from those listing home addresses or self-defined groups. Initial submissions, however, are 30% and 70% from the above groups respectively. When I started this journal I had the opinion that the accepted review processes suffered from the attitudes of a calcified establishment with a built-in preference for papers supporting current paradigms or coming from the elite universities. I have written many editorials on the subject of the review process (Honig 1980a; 1980b; 1980c) and have devoted one issue to this specific subject (Honig 1980c). I think that many of my earlier published remarks are relevant to the general subject of peer review, although they are not directly relevant to the issues raised in the P & C paper. I shall summarise these views in the latter part of my commentary.

Directly commenting on the P & C paper, I make six remarks:

1. Because their procedure is clearly unethical, the journals involved may have detected this and replied in kind with devious or misleading replies.

2. In my experience, and that of my reviewers, there have been many papers that were thought "old hat." This was rarely mentioned in the reviewer's or editor's reports although relevant references to similar work were sometimes mentioned. The major reason for this behaviour on the part of reviewers initially surprised and angered me, but I eventually came to share this attitude. The reason was that direct remarks almost always trigger effusive, detailed replies (usually friendly and discursive), and raise many, many more questions (which are quite relevant) and points of interest. My own time and that of the reviewers was simply too limited to engage in such activities. Such replies also triggered guilt feelings because our own work prevents us from exchanging views or acting as informal advisers and colleagues. I think we found it simpler to concentrate on the paper; and if it was inadequate because it wasn't new, the usual reply would point out some obvious flaws or beg off in some other way. I have been involved in many

direct replies culminating in increasing anger on the part of authors, along with interminable correspondence. This is my comment on the second paragraph of P & C's "Discussion."

3. I do agree, and have myself noticed, that younger or newer individuals in a field, affiliated with nonelite universities, seem to make the best reviewers; they have the time, make a greater effort, and give more detailed, constructive reviews.

4. I also notice that nonexplicit but negative-sounding replies are classed as rejections by P & C. On the basis of my own experience in such matters, I disagree with this classification. I have found that when the author replies with a spirited and detailed response listing many additional relevant supporting references, it usually sways reviewers toward ultimate positive decisions. If the authors terminate correspondence at this point, it is equivalent to a rejection, of course, but one decided upon by the author.

As the Ruderfer (1980) paper shows, a spirited and detailed reply defending the author's thesis with many extensions and current references militates against rejection, or, in Ruderfer's case, results in the strengthening of the arguments. This makes it more likely that submission to another journal would be successful.

5. I do agree, however, with the conclusions of P & C; on the basis of my own experience, I find that there is indeed a definite preference and prejudice for papers from elite groups. This may be even more prevalent in a field like psychology than it is in the hard sciences. I also agree that this is caused by the human reaction of reviewers, but the major cause, I think, is that reviewers (particularly senior ones) are simply pressed for time and cannot devote their efforts to the kind of intense effort that would match that of all authors. I have a relatively weak and somewhat unsatisfactory suggestion to make on this matter (Honig 1980c). I have suggested that there be independent consultants specifically devoted to evaluating papers, and that such consultants should be paid by the author. After going through such a process, the author might submit his final paper, together with the consultant's remarks, to a journal.

6. My general comment on the P & C paper is that they have put their efforts into establishing a relatively minor fault in the review process at the cost of an unethical procedure they have used. I think the procedure could have been, or might in the future be, considered for somehow testing the acceptance of paradigm-breaking suggestions. Submission need not be to the prime journals but, even if acceptance is finally secured, the readership itself is, I think, strongly prejudiced toward allotting its reading time to papers of authors in elite groups.

Finally, with respect to speculative papers in the hard sciences, with which I have been personally concerned, I make the following remarks:

1. The greatest reason for rejection is the authors' poor preparation of papers and the psychological damage from which such authors suffer (see "'They Laughed at Columbus' and Other Author Syndromes" in Honig 1980a).

2. In physics I have found papers particularly contentious and polemical, particularly those concerned with the axiomatic foundations of quantum mechanics and special relativity.

In the case of special relativity, I have had hundreds of submissions; I devoted three issues of my journal to this one subject before finally restricting such papers to discussions of experimental tests. This bears on the three fundamental constraints of science, as mentioned by Harnad (1979); these may be listed as

a. logical consistency;
b. testability;
c. being subject to ongoing self-corrective discussions.

The special problem with quantum mechanics and special relativity paradigm-breaking proposals has concerned (a), logical consistency, although the establishment view has shown that such logical consistency is not experimentally evident with quantum mechanical and special relativistic considerations. Abstract discussions do indeed reveal logical inconsistencies. The view I had to take was that unless and until logically consistent theories (more general than the present ones) result in confirmed experiments, supporting new paradigms and not derivable from current ones, the extreme behaviour of opponents of the quantum mechanical and special relativistic paradigms must be discouraged.

I suppose it is because I have been concerned with speculative papers, usually of a fundamental nature, that my reviewers and I have given little or no notice to the standing of an author's institution, although more than half of our submissions contain sarcastic remarks about the establishment conspiracy and other such attitudes.

W. M. Honig is editor of Speculations in Science and Technology. *Ed.*

Peer review: A philosophically faulty concept which is proving disastrous for science

David F. Horrobin
Efamol Research Institute, Kentville, Nova Scotia, Canada B4N 4H8

Peer review is the procedure that governs access to publication and money in modern science. Its function is to identify and reject poor work, to improve and accept good work, and to let the best through unimpeded. It performs the first two of these purposes reasonably well but fails disastrously with the third.

Peer-review processes have often been accused of being biased; almost equally often they have been defended by distinguished spokesmen for science. The only two *experimental* studies of which I am aware have, however, both upheld the accusation. A number of women complained to the Modern Language Association in the United States that there were surprisingly few articles by women in the association's journal, compared to what would be expected from the number of women members. It was suggested that the review processes were biased. The association vigorously denied this but under pressure instituted a blind reviewing procedure under which the names of the authors and their institutional affiliations were omitted from the material sent to the reviewer. The result was unequivocal: There was a dramatic rise in the acceptance of papers by female authors.

Now Peters & Ceci (P & C), operating in a quite different field, tell a surprisingly similar story. Journal editors and referees were not only found to be biased, but also to be astonishingly ignorant of what had recently been published in their own journals. Obviously the crucial question is whether these are isolated examples or typical of the general situation across academia. I have done no experimental investigations, but I do edit two biomedical journals (*Prostaglandins and Medicine* and *Medical Hypotheses*), and I do not see any convincing evidence that the medical sciences are free of the problems of reviewer bias and competence. In my own experience about one-third of referees' reports are accurate, comment on important issues, and are fair in their recommendations; about one-third are accurate but obsessed with the trivial and recommend revision or rejection on inadequate grounds; and about one-third are inaccurate and can be demonstrated to be so on objective grounds. What constantly astonishes me is the intemperate language in which many reports in the last two categories are couched. The lack of sound judgment among people who have the fate of science and the lives of others in their hands is appalling.

P & C have provided unusually sound evidence for something that those concerned with the reality rather than the public image of science have known for some time. The referee system as currently constituted is a disaster. What is most disastrous is its built-in bias against highly innovative work.

The towering achievements of science for the most part have their origins in brilliant individual minds. These minds are exceptionally rare. The concept of peer review is based on two myths. The first is that all scientists are peers, that is, people who are roughly equal in ability. The second myth is that in those rare instances in which someone who is exceptional does appear, the ordinary scientist always instantly recognizes genius and smooths its path. No one who knows anything at all about the history of science can believe for one second in either myth. Most scientists are not the peers of the very best, and most scientists follow the crowd when it comes to the recognition of brilliance. The *concept* of peer review is philosophically faulty at its core. Ordinary scientists consistently fight against or ignore the truly innovative. The defects in the *functioning* of peer review, such as those revealed by P & C, compound this fundamental fault.

The most important lesson to be drawn from the P & C work relates not to journals at all but to research funding. If one journal rejects an article there are dozens more to which it can be submitted. It would be surprising if even moderately competent experimental work could not eventually be published. This is not necessarily true of theoretical work, which at present has an exceptionally rough passage in the biomedical sciences (which is why I founded *Medical Hypotheses*). But with grants the situation is completely different. In any country there are likely to be only two or three major sources of funding in any field, and those two or three are quite likely to use the same reviewers. There is no reason to believe that these reviewers are any more competent than journal reviewers. There are strong reasons to believe that because the stakes are so high many of them are dishonest and deliberately shoot down work that is in any way threatening to their own personal research lines. The system assumes perfect honesty and integrity and therefore gives a built-in advantage to the many scientists who fail to meet those standards. Peer review is an open invitation to the crooked, which may be one reason why in many areas of the biological sciences there is a lack of substantial progress. Something must be wrong when in spite of all the media hype about medical progress, there are virtually no diseases for which one is likely to be better off receiving the best 1981 rather than the best 1951 medical care. The lack of practical advances may indicate that the problems are exceptionally difficult. It may also indicate that our approaches are fundamentally wrong and therefore cannot have practical results.

I believe that as far as journals are concerned, some form of review is necessary because of the abysmal quality of many submitted papers. It must, however, be controlled by a strong and open-minded editor who is prepared to edit and not blindly act on the recommendations of referees. As far as research funding is concerned, however, I believe that the review system has such faults that it is beyond rescue. Purely destructive comment is of little value, so if we do abandon peer review for grants how do we decide who should get the money? It is my view that not only should the review system be abolished but the grant system should be thrown overboard also. Instead, we should revert to a system in which it is the function of universities and other institutions to support researchers and not the function of researchers to support their institutions. The money at present spent on grants should be given out in two ways: (1) All accredited universities should receive capital grants based solely on the number of their students in order to provide the equipment necessary for a sound research infrastructure. (2) All academics who receive a university post should receive a small grant of $20,000 or so per year which would enable them to employ one technician and do research, provided that they were willing to get personally involved. The big research teams, the vast paraphernalia of grant giving, and the dishonesty involved in peer review would all disappear, and all those with an inclination to do science would be able to do what they wanted.

But if people were given money in this way, without any commitment as to the problems they were to tackle, how on earth would society get answers to the problems it wants to solve? In the 18th century, the British Navy, wrestling with the problem of mapping the oceans of the world, desperately needed an accurate way to determine longitude. They set up a prize of 10,000 pounds sterling, a truly astronomical sum, for the solution, and before very long the answer was found. Science should operate by carrots instead of sticks. Governments should work out what it would be worth to them to solve practical problems, such as, for example, curing schizophrenia or a particular type of cancer. They should then offer a graded series of large tax-free cash prizes for practical solutions to those problems. The cash carrot of $100,000 or $1 million would far more effectively stimulate interest in research in that topic than the most elaborate system of peer-reviewed grants. The scientists who wanted to work on "blue sky" problems would be able to do so without the corruption of having to say that their work would be relevant to this or that problem. The state of science would be very considerably healthier – and we might even get quick solutions to some of the problems governments urgently want to solve.

D. F. Horrobin is editor of Prostaglandins and Medicine *and* Medical Hypotheses. *Ed.*

Peer reviewing: Improve or be rejected

Michael J. A. Howe
Department of Psychology, Washington Singer Laboratories, University of Exeter, Exeter EX4 4QG, England

Prior to saying anything at all about an article that speaks darkly of institutional affiliation, editor-author friendship, and old-boy networks, I must hasten to come clean and admit to having been the doctoral supervisor for the second author, Steve Ceci. Pausing only for the briefest pursing of the supervisory lips (just what, in "Procedure," might "superficial detection" be?), I shall proceed in a fashion that excludes evaluative commentary.

Peters & Ceci's (P & C's) findings reveal a grim state of affairs. Neither the limitations of the small sample nor the fact that the exact causes of the results are yet unknown can diminish the seriousness of a situation in which highly prestigious journals reject a very large portion of (undetected) articles they have recently accepted. It seems likely that systematic bias is involved, related to authors' institutional affiliations and, in some instances, personal reputations, but the present data do not, of course, permit separation of the possible effects of bias from other determinants of unreliability.

Reprehensible though bias may be, it should at least be possible to eliminate it. Blind reviewing is not a complete solution, authors being formidably ingenious at finding ways to reveal their identity if they wish to do so, but it is heartening to reflect that a blind reviewing system offers almost as much scope to writers who wish to create the illusion that their manuscripts come from the pen of more prestigious authors as it offers to those contributors who wish to lay bare their true identity. (My thanks, at this point, for many useful discussions with my long-valued famous friend and colleague, X.)

To the extent that the reported results are *not* caused by bias, the situation seems much harder to ameliorate. If there is a large element of randomness in the reviewing process, attributable to some extent to incompetence among the highly selected and well-regarded scientists chosen for the job, it is difficult to imagine what practicable steps would quickly

produce a marked improvement. P & C suggest a system in which referees themselves are submitted to formal evaluation by "judges, authors, and editors." But who is to review *them?* If the currently accepted experts in their fields are to be evaluated by the supposedly *more* expert (whoever they might be), the outcome can only be to give more control to an increasingly smaller establishment of authorities, hardly a state of affairs that P & C would want to bring about.

Granted that their target article raises more questions than it answers, P & C have done a necessary service in laying bare a serious deficiency in current practice. It is fair to point out that the authors encountered a considerable amount of opposition to the research they report in this article, amounting virtually to a "hands off" reaction from certain quarters. The findings make it all too clear that they were right to persist.

M. J. A. Howe is editor of Human Learning: Journal of Practical Research and Applications. *Ed.*

Interreferee agreement and acceptance rates in physics

David Lazarus
Department of Physics, University of Illinois, Urbana, Ill. 61801

I am neither surprised nor dismayed at the Peters & Ceci (P & C) finding - only somewhat put off by their righteous indignation at the peer-review system's being of finite value, particularly when used deceptively. Even in my science, physics, which is by common consent (however misplaced!) regarded as far less "subjective" than psychology, there is no way that we can run a journal with even far higher acceptance rates (45% for *Physical Review Letters*) without encountering enormous discrepancies between the opinions of different referees. In only about 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! Remove these, or distort them the way P & C have done, and I have no doubt that the "accept" concurrences would drop. But so what? Since when aren't a person's institution and reputation legitimate measures of the value of his work? Good science is not and never has been "objective," except to those who have never practiced it!

When I scale our experience with physics journals to the realm of the journals mentioned with 80% rejection rates, it boggles the mind that anyone could ever imagine that "objective" selection criteria could exist. I just hope we never find ourselves in such a situation in physics. We have enough problems with excellent journals (such as the *Physical Review* - which is arguably the world's most distinguished physics journal) which have 75% acceptance rates. We rely on the honesty and integrity of our authors - and their own self-selection of the quality of papers they send us - as much as on our referees and editors, to ensure the quality of our journals. I hope we can always depend on our authors to provide high-quality physics; without that, there is no "objective" way to publish a quality journal.

D. Lazarus is editor-in-chief for the American Physical Society, which publishes the Physical Review, Physical Review Letters, *and* Reviews of Modern Physics. *Ed.*

Peer review: Prediction of the future or judgment of the past?

Richard T. Louttit
Division of Behavioral and Neural Sciences, National Science Foundation, Washington, D.C. 20550

In their empirical study of reacceptance of once-published articles, Peters & Ceci (P & C) make only casual reference to "peer-review practices in publication *and funding*" (emphasis added). There is no further discussion of the peer review involved in the selection of projects to receive financial support from such federal research-supporting agencies as the National Science Foundation. The implication is left that the two processes - evaluation of research project proposals yet to be conducted, and evaluation of the output of completed research - are the same. I would submit that they are not. Peer review of proposed research involves a prediction of the potential value of the proposed project to the advancement of knowledge in a particular field of science. The ability of principal investigators to conduct the research effectively in their institutional settings and to analyze and interpret the results meaningfully is an essential feature of the grant peer-review process. Studies of this process have been conducted periodically since 1965 (NIH Study Committee 1965). No strong recommendation to modify the process substantially has been forthcoming. The major finding of the special oversight hearings on the NSF peer-review system conducted by the Subcommittee on Science, Research, and Technology of the U.S. House of Representatives in July 1975 was: "The National Science Foundation peer review evaluation systems appear basically sound." No example of bias on the part of reviewers was presented at these hearings. Subsequent studies (see, e.g., Cole, Rubin & Cole 1977; Hensler 1976; NIH Grants Peer Review Study Team 1978) have led to similar conclusions. Some specific changes have been made over the years in an effort to help investigators improve their own research plans, and therefore benefit the entire scientific enterprise (for example, NSF returns to principal investigators anonymous verbatim comments of peer reviewers).

To focus specifically on the P & C study, two possible conclusions from their data are not separable since necessary controls were not used. In addition to the "bias" conclusion, which they favor, the unreliability of the evaluation system that involved only two reviewers per article might have produced their results regardless of the prestige of the submitting institutions. In contrast, review of proposals in the Division of Behavioral and Neural Sciences at NSF typically involves four to six written reviews. These, together with the proposal, are reviewed and discussed by a panel of five to 10 scientists, increasing the reliability of the process substantially. To separate the "bias" and "unreliability" hypotheses the P & C procedure could have been used in journals that employ blind review. If the results were comparable to those reported in this article, then unreliability rather than bias would have been the only possible conclusion.

The use of completely fictitious institution names, rather than those of lower-prestige academic institutions, presents, I believe, serious problems for drawing useful conclusions from the study. The Tri-Valley Center for Human Potential will, of course, not be recognized by reviewers. More important, it does not sound like an institution with a major stake in scientific research. The authors' research design virtually ensures an initial negative affect toward the article on the part of reviewers, thus leading to the conclusion they expected - that reviewer judgments are biased on the basis of institutional prestige. They could instead have sent the articles with real names from lower-prestige universities, such as the University of North Dakota, thus testing bias in relation to prestige level of real academic institutions. In their study of NSF peer review, Cole et al. (1977) concluded that "in general reviewers from high-ranked departments were not disproportionately favoring proposals from applicants in similarly high-ranked departments" (p. 36). Their results also "showed no significant tendency... for eminent scientists to favor the proposals of eminent scientists over the proposals of less eminent scientists" (p. 37).

While I don't find the P & C study particularly definitive,

the continuing effort to assure fair and objective evaluation of scientific research, both before the fact, as in grant proposal review, and after the fact, as in journal review, is and should be of concern to all of us in the community who serve as authors and reviewers.

R. T. Louttit is director, Division of Behavioral and Neural Sciences of the National Science Foundation. See also the latest NSF peer review studies; Cole and Cole 1981 and Cole, Cole, and Simon 1981. Ed.

Publication, politics, and scientific progress

Michael J. Mahoney
Department of Psychology, Pennsylvania State University, University Park, Pa. 16802

In his 1968 analysis of the sociology of science John Ziman wrote that "the [journal] referee is the linchpin about which the whole business of science is pivoted" (p. 111). A busy linchpin, too, with some 40,000 scientific journals currently rendering a new article every 35 seconds (Mahoney 1976). Continuing studies in the epistemology and psychology of science leave little room to doubt the centrality of publication in scientific development. This situation might change in some futuristic society of home computers and broader communication networks, but the role of public certification in knowledge would still remain. As Ziman and others have noted, knowledge is a matter of consensus. What we "know" changes with the nonlinear growth of a paradigm. The knowledge of any given era can be viewed as a mixture of competing viewpoints, some with more ostensible authority than others. This authority probably comes from a variety of cognitive-rhetorical sources, including a coherent explanation, corroborative evidence, wide acceptance by experts, and public endorsement by recognized authorities.

While journal referees are seldom public and rarely endorsing, they serve one of the most important functions in contemporary science. In Diana Crane's (1967) apt term, reviewers serve as "the gatekeepers of science." A published idea or finding is much more persuasive than "anecdotal evidence" or a "personal communication." Likewise, one can influence the probability of publication by reminding journal reviewers that one has previously survived the peer-review system. When the variable of prior publication was manipulated in a study on the parameters of publication (Mahoney, Kazdin & Kenigsberg 1978), authors who cited "in press" material as being their own were more favorably reviewed than authors who cited the same material as being in press by another person. Presumably, the more respected the journal, the more credible the contention.

It is clear, then, that the journal referee participates as an epistemological authority in the evaluation of knowledge claims. To get published is, in a very real sense, to have one's work certified as worthy of attention. The importance of studying this evaluative process hardly needs reiteration. We are sorely ignorant of the system to which we entrust our ideas, innovations, and careers. Peters & Ceci (P & C) have offered a valuable illustration of this important issue.

In many ways the study reported by P & C is one of the most sophisticated and rigorous investigations thus far conducted. Their attention to descriptive detail and concerns of validity is impressive. Rather than pick nits in the fabric of experimental design and experimental procedure, I shall confine my remarks to the meaning of their inquiry. In the final paragraph of their report P & C question the assumption that "the review process is basically objective and reliable." Their data are offered as empirical corroboration of potential reviewer bias, incompetence, and unreliability. The focal intent of their paper is to challenge our tacit acceptance of "business as usual" in the pursuit and procurement of publication.

My first response is one of hearty agreement. The phenomenology is not unlike what occurs when you read some clear-cut hard data on a phenomenon you have discussed for years with friends and students. I have yet to meet an aspiring or struggling author who blindly trusted in the objectivity and justice of peer review. Having served as a journal editor, I have also had the first-hand opportunity to explore some of the tacit biases that can creep into such things as the first impression of a manuscript and the assignment of referees. I was surprised, for example, to find myself half-consciously influenced by such things as the prestige of an author and the "clean-ness" of the typed copy - factors that ideally should not have been foremost in the competition for my attention.

But compete they did, and I compensated by increasing my acceptance rate and allowing my readers to pass their own judgments. When I received a submission from an "elite" author, I found myself more cautious in assigning reviewers. I knew that some would be so honored to review such a person's work that they might be blind to its limitations. Still others were likely to challenge the big guns and capitalize on an opportunity to spread their own plumage in secret. With unknown authors I was more often influenced by such factors as writing style, creativity, and timeliness. If the manuscript was clearly marginal or an unlikely candidate I was reluctant to assign it to my most conscientious referees. They would often choke on the volume of its flaws and return it to me with a clear message of disapproval. Many would voice their anger over time wasted on the outliers.

The point here is that bias, incompetence, and unreliability - in other words, human factors - are unquestionably present in our current system of certifying knowledge. I have no qualms with P & C's main thesis. Our thinking might diverge, however, in relation to the implications of this assertion. Where they seem to imply that we should be striving to increase objectivity and reliability in the peer-review process, I would argue that such goals may entrench us in a *degenerating* (rather than *progressive*) problem shift (Lakatos 1970). A progressive problem shift is one that has excess theoretical and empirical content such that it predicts and facilitates novel ideas and findings. The retrenchment in objectivity and reliability is, in one sense, an appeal to the justificational metatheories that are responsible for some of the conceptual bankruptcy in modern science (see Lakatos & Musgrave 1970; Mahoney 1976; in press; Weimer 1979).

Without belaboring some of the fine points, I would like to point out that *objectivity* and *reliability* are concepts that derive from an epistemological metatheory that assigns a relatively passive role to the human brain. "Objectivity" usually refers to a state independent of the mind, and reliability refers to a consensual agreement on some referent. Both concepts relegate reality to a realm distinct from human knowing processes. Reality is something out there to be publicly explored and democratically defined. The mind and its senses may attempt to map reality, but the territory is independent of the surveyor. This notion is a familiar one to philosophers and epistemologists. It is, however, a naive one in light of recent developments in cognitive psychology, psychobiology, and the psychology of science (Davidson & Davidson 1980; Mahoney 1976; 1982; Shaw & Bransford 1977; Weimer 1977; Weimer & Palermo 1981). While institutionalized science may attempt to emulate the precision and order of logic, its actual development seems to be more adequately captured by perspectives that acknowledge its inherently subjective, "psycho-logical" dimensions (see Mitroff 1974; Weimer 1979). We are doomed to deny or bemoan the problems of "objective science" until we appreciate the naiveté of that assumption.

P & C have communicated valuable information on the role of

arational processes in the certification of knowledge. Papers that were previously published by recognized authorities from respected institutions have been almost uniformly rejected when resubmitted under different names and affiliations. To what do we attribute the variance in their dispositions? Time? Referee sampling? Author's visibility? Institutional affiliation? To what extent did P & C instantiate their own invocation of personal biases in the conduct of science? What are we to believe? This is, in my opinion, the bottom line of the research theme they have expanded. It might be more accurate to say, *How* are we to believe? in that the question is most basically of an epistemological nature.

As conscientious scientists we try to remain current on developments in the field. We read our favorite journals and compare our experience with the received views. We are generally most comfortable when the discrepancy between these two is minimal. Sometimes, in our busy schedule, we have only the time to read titles and abstracts. Our attention is drawn by themes that are of current relevance in our lives. We half-consciously presume that the abstract of an article should be a crystalline residue of its condensation. For others of us, attention is more focused on method than message. We immediately inspect the tables and scrutinize statistical credentials. The alleged data must pass through formal decision rules in their personal accreditation. In both extremes the variable that is overlooked is that which is most obvious – namely, that *the contents of any contemporary psychological journal are likely to be a very selective (20–30%) sample of the evidence and ideas offered for publication*. The parameters of that selection process are seldom confronted, partly because the figure of accredited knowledge is seldom contrasted with the ground of suspended or rejected assertion. It is naive to forget or minimize the power and responsibility assigned to the "gatekeepers" of archival knowledge.

There will always be human factors in human knowledge. It is an active, participatory process. A more basic question might be, Are we aware and accepting of the factors that are most influential in the public dissemination of scientific work? While we may readily acknowledge the presence of paradigm politics and personal prejudices in the operation of a scientific journal, are we really aware of the magnitude and dynamics of their influence? What can be done to improve the system? As P & C aptly note, research on the peer-review system requires considerable time, persistence, and a tolerance for lack of cooperation (if not wrath) from journal editors and referees. In an earlier study of the peer-review system (Mahoney 1977), the emotional intensity and resistance of several participants were expressed in the form of charges of ethical misconduct and attempts to have me fired. Several editors later informed me that correspondence from my office was given special scrutiny for some time thereafter to ascertain whether I was secretly studying certain parameters of their operation. The emotional intensity that surrounds research on the peer-review system should not be surprising if we recall its role in the certification of knowledge claims. When tacit authority structures are first identified and scrutinized, the initial response is rarely one of welcome.

The issues raised by P & C are important in dimensions far removed from the perfection of a reliable and objective system of communication. Communication can no longer be viewed as a passive process. We are active participants in constructing the personal and theoretical realities to which we respond. Objectivity and reliability may be poor guides in our nurturance of truly progressive scientific development.[1] From a more ecumenical perspective, the peer-review process becomes less of a culprit and more of an instantiation of our tacit assumptions about knowledge. We ultimately trust in certain authorities (Bartley 1962), and we seldom question the warrant for our commitment. We accept the probability of human factors influencing the information to which we have access, but we rarely contemplate the filtration processes that may be diluting our experience. We subserve the pragmatic demands of publication rather than confront the magnitude of the power we are willing to concede to the publishing industry.

Let me conclude by saying that the issue of subjectivity in publication is one of the most timely questions posed by contemporary students of human behavior. The content of our knowledge must necessarily reflect its process, and the latter remains one of the most fascinating mysteries of psychological inquiry. The role of personal boundaries in information processing will probably continue to stimulate research for some time to come. My essential comment on P & C is therefore more of an extension than a critique. I would argue that their data are significant and demanding. I would also argue that attempts to objectify the review process are likely to purchase reliability at the price of innovative quality. The challenge before us is not to improve the current system so much as to consciously reappraise the purpose and functions of that system. We must be willing to scrutinize closely and reflect upon contemporary processes of knowledge accreditation and dissemination. And, if we value the scientific spirit of exploration and development, we must remain open to yet unknown demands for progressive changes in the institution of publication. The "publish or perish" maxim has long noted the connection between the survival of scientists or their work and the demand for public performance. In this decade perhaps we will have an opportunity to reexamine that maxim and to explore less dichotomous options in the nurturance of knowledge.

NOTE

1. One can, for example, conceive of the possibility that more heuristic communication in our scientific journals could be achieved by publishing those manuscripts that elicit the *least* reliable ratings (i.e., the lowest interrater agreement). This would favor those manuscripts on which one reviewer was enthusiastic and another flatly rejecting. Such divergence (as opposed to consensus) could reflect the presence of a robust issue that elicits dichotomous responses and inherent contrasts in our constructions of reality.

M. J. Mahoney was formerly editor of Cognitive Therapy and Research. *Ed.*

Reform peer review: The Peters and Ceci study in the context of other current studies of scientific evaluation

Clyde Manwell and C. M. Ann Baker
Department of Zoology, University of Adelaide, South Australia 5001

Regression below the mean - ascription or "flawed masterpiece" hypothesis? Peters & Ceci's (P & C's) paradoxical result is that, upon resubmission, eight of the nine (already published) papers were rejected. They favour the explanation that there is systematic bias based on *ascription*, that is, that referees and editors are more likely to judge a manuscript favourably if its author is recognized as being famous, or at least is located at a prestigious institution.

As such, P & C's results would be an additional example of what the eminent sociologist of science Robert K. Merton (1968a) called the "Matthew Effect," based on the biblical aphorism: "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."

While we consider ascription to be the most likely explanation of P & C's paradoxical result, we believe that an alternative hypothesis should be considered. For the sake of simplicity we call this alternative hypothesis the "flawed masterpiece" hypothesis and point out that it and ascription are *not* mutually exclusive. Some scientific papers are "safe" or "low risk"; such

papers are uncontroversial and unlikely to provoke strong feelings in editors or referees. Other scientific papers are "high risk"; such papers break new ground, perhaps by presenting a new method or a new (and possibly extreme) viewpoint, or by questioning a previously accepted conclusion. It is reasonable to assume that high-risk papers will contain a higher frequency of actual errors, or at least results or statements about which referees or editors will argue, than low-risk papers, many of which will involve straightforward application of accepted techniques or paradigms.

Thus, in many high-risk papers there will be a combination of *both* innovation and error (or at least minor imperfections) – in other words "flawed masterpieces." In a sense, nothing ventured, nothing gained.

In "Procedure" P & C present evidence that the 12 papers chosen for resubmission were above average in their citation levels in *Social Science Citation Index*. Accordingly, theirs was not a random sample of published papers. The above-average citation score might just be more evidence for ascription: The papers attracted above-average attention because of the eminence of their authors. However, the above-average citation levels might mean that the papers were recognized as at least minor masterpieces. Accordingly, it may well be that these papers also included some errors, or at least were controversial, thereby provoking an unfavourable response from referees and editors upon resubmission.

P & C realize that their experiment was not perfectly controlled. Through no fault of theirs it was impossible to ascertain the fate of resubmitted manuscripts that had previously been rejected. However, other aspects of experimental design could be modified. One might deliberately seek out low-risk and high-risk papers and resubmit them, comparing referee response.

Another alternative would be a 2 x 2 experimental design, allowing one to separate ascription from other variables. Consider the following four categories: 1-2, already published papers by eminent individuals are resubmitted with fictitious (and therefore low status) names (which is the experiment P & C performed). 2-2, already published papers by low-status individuals are resubmitted with fictitious names. 1-1, already published papers by eminent individuals are resubmitted with names of *other* eminent individuals. 2-1, already published papers by low-status individuals are resubmitted with the names of eminent individuals.

The reason for this elaboration of experimental design is that there is already evidence that changes in the characteristics of a submitted paper can influence the behaviour of referees and editors. Mahoney (1976; 1979) sent manuscripts to 75 referees for the *Journal of Applied Behavioral Analysis*. These manuscripts had identical introductions, methods, and citations. The manuscripts differed in how (or whether) data were presented, and whether the conclusions agreed with, or were contrary to, the data. For a given manuscript, reviewers showed poor agreement (as P & C noted in their literature review). However, of pertinence to the present discussion is Mahoney's finding that if the data demonstrated inconclusive results, the manuscript was not favourably received. Worse, manuscripts *without* data were favourably received. Not too surprisingly, manuscripts showing positive results with behaviour modification were evaluated favourably by referees for the *Journal of Applied Behavioral Analysis*.

In summary, Mahoney's (1976; 1979) contributions show that referee behaviour is influenced by relatively small changes in the structure of a manuscript and suggest that general paradigm orientation also affects judgment. It is thus not unreasonable to assume that the more innovative or imaginative manuscripts might evoke more extreme responses from referees and editors. Since such special contributions are also more likely to contain some genuine errors, or at least points with which referees might be predisposed to argue, these high-risk papers run a greater chance of rejection. If this flawed-masterpiece hypothesis explains a significant amount of P & C's paradoxical results, then it carries an important implication: The probability of rejection of such important papers (as judged by the higher than average citation score) is close to 50%, perhaps higher. It is likely that the importance of ascription is that referees and editors are more willing to overlook shortcomings in otherwise excellent manuscripts if they know that the manuscripts are written by eminent individuals – a conclusion reached also by A. Carl Leopold (1978). Since the majority of scientists are not eminent (the size of an elite is estimated to be proportional to the square root of the total number of individuals), these results suggest that much individual creativity is lost.

The "gatekeepers of science" – and plagiarism. P & C's finding that only three out of 38 referees and editors detected the resubmitted manuscript as one that had already been published is congruent with observations on plagiarism. A recent, widely publicized case involved a Jordanian researcher who pirated a number of already published papers (plus the literature review of a research grant proposal) and "recycled" these items under his own name with otherwise minimal alteration (see references and discussion in Manwell & Baker 1981; also Broad 1981b). The fact that that plagiarist was able to amass a not inconsiderable publication list is prima facie evidence that many referees and editors cannot detect resubmissions based on already published research.

It can be argued that, in the case of the Jordanian researcher, the plagiarized articles were recycled into relatively low-prestige journals, where standards of refereeing might be assumed to be lax. However, there is another, less widely reviewed case of multiple plagiarism where some of the recycled papers did appear in high-quality journals (Editorial note 1965).

This earlier plagiarism case illustrates an additional important complication, suggesting that such detection failures are more common than suspected. By chance, one of us (CM) knew both the plagiarist and one of the whistle blowers who spotted the plagiarism. The whistle blower said that he had very great difficulty in getting the case taken seriously by either the editors or the administrators at the plagiarist's university; furthermore, the whistle blower felt that his own career had been badly damaged because he had rocked the boat – a view expressed by other scientists who have attempted to call attention to cases of fraudulent behaviour (Manwell & Baker 1981). Subsequent to the publication of that article yet another such case has appeared: At the University of Newcastle (New South Wales) the administration kept the plagiarist and sacked the whistle blower (Martin 1981a; additional information is available from the authors of the present commentary).

Since detection of plagiarism can have grievous consequences for the careers of those who are honest enough to protest such theft of information, it is likely that many cases go unreported. It is even conceivable that, in the study by P & C, some of the referees or editors were following a career defence mechanism: They recognized the manuscripts as being recycled under another name (i.e., apparent plagiarism) but preferred to swallow the whistle rather than blow it; hence, they found *other* reasons for rejecting the recycled manuscripts, reasons that were less likely to cause controversy.

Peer review is fair – or can some establishment sociologists of science interpret their own data? A comparison with the Cole, Rubin & Cole study. There is an important omission from the references on peer review in P & C's paper: The large study by Cole, Rubin, and Cole (1978) is widely quoted by administrators as demonstrating that present secret peer-review practices are fair. In their more popular article summarizing the definitive 1978 study, Cole, Rubin, and Cole concluded, without equivocation, that there was "no evidence to substantiate recent public criticisms" of peer review (Cole,

Rubin & Cole 1977, p. 34). In fact, their study has been criticized on several grounds, notably that its experimental design could not detect a mixture of positive and negative intentional bias and that the results did not demonstrate fairness (Chubin 1980; Manwell 1979; Michie 1978; Mitroff & Chubin 1979). We add here three further pieces of evidence concerning the inadequacy of the Cole et al. study which bear on the issues P & C raise about peer review.

1. Positive and negative intentional bias: Favouritism and disfavouritism. Cole et al. tend to ignore the results of the survey done by Hensler (1976) on the peer-review practices of the very same organization that they studied, the National Science Foundation. Hensler found that of 1,068 reviewers 50.3% felt that NSF programme directors did a good job in matching reviewers to research grant proposals - but only 14.8% of the reviewers mentioned that peer review was an "unbiased process." Among the weaknesses in practices of peer review, the referees mentioned that it "allows favoritism towards friends and colleagues," claimed by 10.7% of the reviewers; it "allows bias against professional enemies," claimed by 3.8% of reviewers; it is "biased against innovative proposals," claimed by 6.6% of the reviewers; and it "allows too much opportunity for bias, no special type mentioned," claimed by 11.1% of the reviewers (Table 8 in Hensler 1976, p. 25).

Thus, a significant percentage of a large number of individuals who had personally refereed research grant proposals for the National Science Foundation felt that either positive or negative intentional bias occurred - and they felt this sufficiently strongly to volunteer these specific criticisms in a rather open-ended survey.

Furthermore, Hensler controlled for the "sour grapes" effect. Of 118 peer reviewers who had had an unsuccessful NSF application of their own in the last five years, 33.1% felt the peer-review system was "sound," 59.3% felt it was "acceptable [but had] some weaknesses," and 7.6% felt it had "many weaknesses." Of 477 peer reviewers who had had a successful grant application with the NSF in the last five years, 49.1% felt peer review was "sound," 47.4% felt it was "acceptable [but had] some weaknesses," and 3.6% felt it had "many weaknesses" (Table 7 in Hensler 1976, p. 23). Thus, although opinions are coloured slightly by a reviewer's own success or failure in the system, this only accounts for a modest amount of the dissatisfaction with present peer-review practices.

Cole et al. should have taken these results into account in their study. Since the Cole et al. analysis combined many outcomes in a classical multiple regression procedure, it was possible to test only for unidirectional bias - whether, for example, the likelihood of success was influenced by institutional affiliation. Their procedure could not detect mixtures of positive and negative bias. Furthermore, in studying the behaviour of scientists it is extremely difficult to determine bias itself; for example, Martin's (1979) elegant analysis of how the design and interpretation of experiments in high-altitude photochemistry are influenced by political considerations regarding supersonic transport and the source of research funding. Martin's results, like those of P & C, show the importance of a *particularistic* approach in studying bias. When the barrier of secrecy surrounding peer review has been penetrated, researchers have produced evidence of incompetence and dishonesty (Horrobin 1974; Manwell 1979; 1981; Margulis 1977).

2. Evidence against referee consensus - and also against ascription. In the literature surveyed by P & C several researchers have shown that there is often poor agreement between referees evaluating the same manuscript. Similarly, in the study by Cole et al. (1978) the consensus among referees was often poor, although this varied significantly among disciplines: a relatively narrow standard deviation for referees evaluating research grant proposals in modern algebra, 0.31, or in solid state physics, 0.35. These are "hard" sciences, with strong coherence and agreed-upon paradigms. For "soft" sciences, with both paradigm and personality conflict, the standard deviations are larger: 0.69 for ecology or meteorology. However, even a fairly rigorous science with very little paradigm variation, biochemistry, showed an unexpectedly large standard deviation: 0.60.

When Cole et al. attempted to control for different levels of quality among the applicants, by partitioning the group of biochemistry applicants into quintiles based on their *Science Citation Index* scores, there was no more agreement among reviewers for the most highly cited scientists than for those of intermediate quintiles or the lowest ranks. The wide variance is totally unaffected by the quality or "impact" of the biochemistry applicant. Since the identity and institutional affiliation of applicants are known to the referees, it appears that, in contrast to P & C's results for manuscripts from psychologists, ascription does not explain the lack of agreement between reviewers of research grant proposals in biochemistry.

It may be that the ability to write convincing grant proposals in biochemistry is unrelated to the quality of previous publications as measured by the frequency with which they are cited by colleagues. It may also be that the system is very noisy: The quality of the grant proposal and the applicant's past accomplishments may be obscured by a mixture of positive and negative intentional bias - or just random evaluation, or sheer incompetence. Whatever the explanation, it does not justify Cole et al.'s assertion of fairness. Of particular concern to us is that Cole et al. fail to stress the fundamental contradiction between their more recent results and their earlier studies (Cole & Cole, 1972; 1973). For the entire data aggregate on the research grant proposals in 10 branches of science, the applicants' science citation score explained only 6% of the variation in whether or not research proposals were funded. Yet, in their earlier studies Cole and Cole emphasized that *Science Citation Index* scores are strongly associated with other measures of quality and that the reward system of science is fair. Were that the case, then the citation score would explain a much larger percentage of the variance in success or failure of research grant proposals. Obviously, to resolve such a fundamental discrepancy a more particularistic approach - such as that employed by Mahoney (1976; 1979) or P & C - with the experimenter deliberately manipulating different parts of manuscripts (or research grant proposals) is necessary in order to ascertain just how fair and accurate the peer-review system really is.

3. The role of programme directors and journal editors. The literature on peer review tends to neglect the role of programme directors or journal editors vis-à-vis referees. A referee can recommend acceptance or rejection of a manuscript or research grant proposal, but only editors or programme directors have the real power to effect a decision, and that decision need not necessarily be congruent with the recommendations made by reviewers. Although it hardly supports their assertion that peer-review practices are fair, Cole et al. (1978) provide a valuable case history.

Their "case 10" was a research proposal from a single investigator. The proposal was reviewed by four referees. One rated it "excellent," two rated it "very good," and one rated it "good plus" (Cole et al. 1978, pp. 102-3). Furthermore, there was some agreement among referees that the investigator in question had performed well in the past; "remarkable success" was the comment of one referee. The only criticisms were two rather general ones, and these criticisms were given by only one of the four referees in each of two cases. One referee, while saying that "he [the applicant] is likely to accomplish his primary objectives," did "question the proposed duration of the grant." Another referee wrote: "My only reservation about this proposal is that it is not clear why one should study [details deleted by Cole et al. 1978 in order to provide anonymity]."

Yet, at a time when most research proposals were being funded, and when many research proposals received far more

damning criticism, the programme director rejected the above-described grant proposal outright. Cole et al. (1978, pp. 104–6) present explanations, both from the programme director and from themselves, for the fairness (?) of this decision. First, we quote the programme director himself:

> I remember that [research proposal] well. We play funny games here. Once in a while you get a proposal and you are absolutely certain about how the proposal is going to turn out. . . . What I did with this [research proposal] was send it out to four people, three of them were reviewers who, for one reason or another, were overly generous. So, what I'm saying is that this isn't really a fair application of the review process because in some sense I had a good idea of what the reviews were going to say in advance, or *should have said in advance.* I was using this proposal not to test the reviews of the proposal but to test the reviewers themselves.

Cole et al. (1978, p. 105) do allow that "it might appear to be a perfect illustration of inequity or bias – a case in which the program director disregarded the peer reviews – and it may indeed be such a case." They (Cole et al. 1978, p. 105) provide their own rationalization, which has a special significance for the P & C paper:

> A terse comment by the best man or woman in a field, or one that damns with faint praise, might in fact provide a program director with more useful information than he could possibly get from five or six less-qualified reviewers. What is the program director to do, therefore, when he is faced with conflicting reviews in which the reviewer he believes to be most qualified is in a minority?

This excuse is hardly acceptable given the evidence that there were *no* significantly "conflicting reviews," all four referees rating the proposal "good plus" or better. It is also clear from his own words that the program director had a negative opinion of the applicant *before* he received any referees' reports. What the quote does show is that, although the situation is somewhat different, P & C's hypothesis that ascription plays an important role in peer review has wider generality than many realize. The perceived reputation of referees can influence how programme directors or editors will interpret their recommendations.

Cole et al. (1978, pp. 105–6), in discussing this case history, make an important generalization: "Most scientists who serve on review panels, or who have had occasion to read many referee reports for journal articles, are fully aware that there are code words and modes of negative judgments in such reports." This generalization provides yet another paradox in attempting to understand peer review: Why the frequent use of "code words and modes of negative judgments in such reports" when the supposed justification for secrecy in the peer-review process is to allow referees to make frank and full comments? The danger in the use of "code words and modes of negative judgments" is that not all referees, editors, or programme directors will interpret them in the same manner – introducing an additional element of chance in what is already a rather random process.

The professor of machine intelligence at the University of Edinburgh provides an account which should raise concern (Michie 1978, p. 11):

> A fairly senior official of a non-British [research granting] agency, whose nationality I shall not disclose, once told me that if for any reason he felt justified in short-circuiting the system in order to get a given result, he would make a judicious selection of referees – either the scientist's particular friends or his particular enemies, according to which result he wanted. He needn't have told me. I knew it already. In his heart of hearts so does any scientist who has been in the game any length of time.

Michie's article emphasizes the role of personal prejudice in peer review. We know of no accurate estimates of the frequency with which personal enmities, or friendships ("old boy networks"), influence the peer-review process. If the frequency with which such allegations are made in gossip among scientists is any indication, then both positive and negative intentional bias must be relatively common in secret peer review.

There is agreement among sociologists of science that many scientists are extremely competitive (Hagstrom 1974). Examination of the autobiography of a Nobel Prize-winning molecular biologist who epitomized competitiveness revealed that his inaccurate description of a female scientist arose from a form of victim blaming (Manwell & Baker 1979; Sayre 1975). Such situations are common – and since most peer review is secret, often verbal, individuals cannot be held accountable for deliberate or accidental errors. Feuds among scientists are often conducted "with a remarkable degree of bitchiness," as Arthur Koestler (1971, p. 54) has written.

Accordingly, only a careful particularistic study can reveal the extent to which personality conflict, professional rivalries, paradigm differences, and other sources of bias affect peer review. The Cole et al. study was not designed to test for such biases, or for referee incompetence. The very fact that Cole et al. could only explain such a small amount of the total variation in success or failure of research grant applications suggests that these biases may well be quantitatively important.

Can science survive secret peer-review practices? Peer review, nearly always done in secret, determines who gets the opportunity to do scientific research. Most discussions of peer review have concentrated on referees evaluating manuscripts or research grant proposals, the gatekeepers of science. Far less attention has been given to other important dimensions of peer review, such as the awarding of prizes (Leopold 1978), election to the more prestigious learned societies (Boffey 1975; Moyal 1980), and, perhaps most important of all, the hiring and firing of scientists. For this last category the only useful studies known to us explore peer review in the "academic marketplace" (Caplow & McGee 1958; Lewis 1975), although there are also valuable data in dismissal cases, for these reveal that professional incompetence or dishonesty is rarely even an allegation and that political and personal factors are predominant (Lewis 1972; Martin 1981b). Suspicion of abuse in the peer-review practices involved in getting and keeping academic jobs has reached such a point that a state legislature is considering a bill that would allow candidates access to formerly secret letters of recommendation (California may be sued on secret files 1978). Yet another relatively neglected dimension of peer review concerns the evaluation of scientific "apprentices," the examination of graduate students and their theses (Lovas 1980; Mahoney 1976).

Science has been succinctly defined as "public knowledge" (Ziman 1968). "Communality" (or openness in communication) is one of R. K. Merton's original four norms of science – with secrecy as its antithesis. Yet, secrecy pervades present practices of peer review in science.

This arrangement has been repeatedly justified in terms of the importance of maximizing quality control. However, secret peer review did not protect science from a number of cases of plagiarism and the publication of fraudulent research – cases that occasionally damaged certain scientific disciplines (Manwell & Baker 1981). It is now evident that the damage has spread to the questioning of the integrity of science by members of the public. At the time this commentary on peer review was being written there were three different congressional committees in the United States investigating fraud in science – and questions have been asked about peer review (Broad 1981a).

The dependence of science on secret peer review is to some extent a post-World War II phenomenon. Systematic use of secret peer review was not a common practice of many scientific journals until well into the twentieth century. Because it was common for nineteenth-century scientific journals to publish debates over major scientific issues, much of the peer review in those days took place in public.

It was only after World War II, and even later with Sputnik I, that science became dependent on large-scale financing from government research granting agencies, and thus secret peer review became widespread. Furthermore, the relatively low cost of much pre-World War II scientific research meant that the average scientist was less dependent on his peers for judgment prior to the publication of results.

Thus, it is largely in the last 30 years, roughly only 10 percent of the period encompassing the history of Western science, that secret peer review has become so dominant over all forms of scientific activity.

In addition to widespread suspicion of abuse of peer review, the last few years have seen the emergence of other serious pressures on science, notably the shortage of jobs, the relative reduction in funding, and the increased bureaucratization (which has, in part, arisen from legislation caused by protest over abuse of peer review – as in discrimination against female scientists). Combine this with the evidence for serious demoralization of science students (e.g., Zinberg 1976), "the switch from science" and the numerous "brain drains," and a generally bleak picture emerges for the future of science.

Philip Abelson (1979) has written in an editorial in *Science:*

> During the past 2 months I have had casual conversations with about 20 professors from widely scattered universities. If their attitudes are an indication of the spirit on campus, the long-term future of science in America is in jeopardy. Not one of those 20 conveyed the impression that life is great, science is fun, and that academic research is the best possible of all activities. Rather the majority were gloomy – some were bitter.

Jerome Ravetz (1981), in commenting on the recent outbreak of fraud cases, warns of "the inherent difficulty of the quality-control operation in science" and that "integrity and morale, at the highest professional levels, are more crucial to the health of science than perhaps in any other organized social activity." (p. 7)

It is clear that research on, and reform of, peer review is one of the most urgent tasks facing the scientific community.

After this commentary was submitted, two pertinent new studies of peer review appeared: Cole and Cole 1981 and Cole, Cole, and Simon 1981. Ed.

Making the plausible implausible: A favorable review of Peters and Ceci's target article

Jason Millman
Department of Education, Cornell University, Ithaca, N.Y. 14853

Peters & Ceci's (P & C's) methods are exemplary. Like two detectives, they search for clues and follow leads, discarding hypotheses and eventually building a convincing empirical case against nonblind peer review. Nonblind reviews can be hung on conceptual grounds alone, for unlike grant proposals, which have the track record and potential of the investigators and their affiliations as explicit criteria, scholarly journals publishing original investigations are expected to use only the merit and appropriateness of the work as criteria. Whether nonblind reviews actually interfere with the valid application of the merit and appropriateness criteria is moot as long as a large segment of the community of scientists perceives that such interference results. P & C's paper adds empirical evidence to such a perception, but its more important contribution may be as a noteworthy demonstration of a rational scientific approach to social science research.

A large part of such an approach makes use of the following syllogistic argument:

Major premise: If *A* then *B*.
Minor premise: *B*.
Conclusion: Therefore *A*.

With respect to the reviewer-bias aspect of the P & C study,

A = Reviewers and editors are influenced by the prestige of an author's name or institutional affiliation.
B = Previously published articles submitted again with less prestigious names and institutions will tend to be rejected.

The syllogistic argument is invalid because events other than *A* could account for *B*. In the present case, regression effects, change in editorial policy, and the like are competing hypotheses or explanations for the finding, *B*. Much of good research design consists of identifying the most plausible alternative explanations and evaluating their validity. If these hypotheses are not supported, the original claim becomes more credible.

P & C are adept at identifying and assessing seemingly reasonable rival explanations and, in the present study, demonstrating their implausibility. Many of these contending explanations for the study's findings, and the evidence against them, are cited in Table 1.

P & C's expertise is not limited to searching for and evaluating rival hypotheses. The investigation contains many commendable features, such as:

1. Permission was obtained in writing from all but one of the senior authors to use their original papers as test manuscripts (the exception being the result of a clerical mistake). Permission to conduct the study was also obtained in writing from all eight non-APA (American Psychological Association) publishers of the articles that were used. (Personal communication with S. Ceci.)
2. The questions raised command much interest, and the one about detection could even be labeled daring.
3. The importance of the study is competently discussed.
4. Claims about the reviewing process, previous research, and the like are amply documented. Contrary findings are reported.
5. A natural setting was used.
6. A good rationale was presented as to why the manipulable independent variable was important.
7. The journals, articles, and author institutions were well described. The reader was clearly told what the treatment was.
8. Evidence was presented to document "prestige" affiliation.
9. The authors guarded against superficial detection of the papers without changing meaning or otherwise compromising the treatment.
10. An optimum time lag (18–32 months) was used between the dates of publication and resubmission. The time was short enough so that it would be reasonable to expect detection, yet not so short that reviewers would not have had a chance to read the articles.
11. The authors provided a plausible explanation, that is a description of the mechanism, of why prestige mattered.
12. Convergent evidence in support of the reviewer bias hypothesis was presented. Added to the major premise shown above was an observed event, *B'*; namely, greater agreement among reviewers for nonblind (selective) review journals than among reviews for blind (selective) journals (see "Disscussion," paragraph 13 and note 6).
13. The authors' "conjectures" were called to the readers' attention.
14. Some (but not all!) of the study's limitations were pointed out.
15. Some implications of the research findings were noted and suggestions were made.

This commentary began by baldly stating that manuscript review for scholarly journals should be blind, and that the commendable methodological features of the study may be a more important contribution than its empirical claims. A possible exception is the finding that a sizable majority of the

Table 1 (Millman). *Rival explanations for the high rejection rates of resubmitted articles and the evidence against these alternative accounts*

Rival explanation	Contrary evidence
1. Publication criteria changed (and the content of the article is no longer appropriate).	1. Publication criteria remained the same (see note *a*, Table 3 in target article). All but two of the editors were the same.
2. Rejection rates increased (and the article with the true author and institution identified might not now be accepted).	2. Rejection rates averaged about the same during the two periods (see Table 3 in target article).
3. Reviewers judging the original submission were less qualified, easier, etc.	3. "no major shift in the qualifications or criteria used by the journals to select reviewers for the two periods" ("Discussion," paragraph 7).
4. Reviews were not outright rejections, but really encouragements to resubmit.	4. Rejection statements (see note 4 in target article) do not so indicate; where a review did not explicitly state a decision, six independent judges performed a content analysis and concluded that a negative decision was implied by the review.
5. The content is now well known, so it was reasonable for the article to be rejected now, whereas at the time of the original submission, the content was fresh.	5. An analysis of the reasons for the rejections revealed no instance in which the article was rejected for redundancy, for being old, and the like.
6. Theories, methodologies, or innovations developed after the original article was published, could provide legitimate grounds for rejecting the resubmitted article.	6. An analysis of the reasons for the rejections revealed no instance of a reason that could not have been used at the time of the first review.
7. Regression effects, caused by the use of only previously published articles, account for the findings.	7. In two separate analyses, it was shown that the magnitude of the decrease in acceptance rates could not reasonably be accounted for solely by the regression effect.
8. The manuscripts reviewed at the two times were not the same. The original, preedited manuscripts may not have been written as well as their published (and resubmitted) versions.	8. The force of such differences would be to lower the rejection rates of the resubmitted papers; thus the rival explanation, if true, would strengthen the bias claims found in the paper.

editors and virtually all the reviewers did not detect that the manuscripts were previously published. Reasonable inferences are that editors do not read the manuscripts they reject and that the allegedly expert reviewers are not expert. Such inferences, if valid, describe a distressing situation. Perhaps authors should not be the only recipients of letters of rejection.

The preprints of accepted BBS target articles are circulated to a large population of potential commentators selected from the following sources: (1) specific, computer-aided, international, interdisciplinary literature searches; (2) the BBS Associateship; (3) editorial recommendations; (4) referees' recommendations; and (5) authors' recommendations. The commentaries that actually appear represent the outcome of this sampling process, as constrained by considerations of time, space, quality, and relevance. Please note that the contributors of the preceding and following commentaries (J. Millman and B. Mindick, respectively), were not only recommended by, but are also current institutional colleagues of one of the coauthors of the target article (SJC). This is perfectly appropriate, and is only drawn to the reader's attention to point out the possibility that under such conditions a contribution may not always be as independent of the target article and its authors as the usual commentary. Ed.

When we practice to deceive: The ethics of a metascientific inquiry

Burton Mindick
Department of Human Development and Family Studies, Cornell University, Ithaca, N.Y. 14853

As an academic psychologist as well as an ethicist, I see the Peters & Ceci (P & C) report as exemplifying elements of the dilemma confronting investigators who attempt to carry out nontrivial, ecologically valid research. P & C clearly felt (I suspect rightly so) that to carry out their examination of the peer-review system adequately and without undue reactivity, it was necessary to deceive the journal editors and reviewers who were their research subjects. Deception is of course never ethically attractive, particularly because harm to the defrauded party or exploitation is so often a concomitant of dishonesty.

But despite the initial aversion that we ought to feel toward deception research, a proper assessment of the ethical merits of the study demands the weighing of other factors as well. In pondering the issue, I found myself drawing upon two disciplines in which I have been trained and for which I hold considerable respect: the ethics of psychological research and those of the Jewish homiletic tradition. The two systems speak with remarkable unanimity.

The first source that seems appropriate derives from the Mishnah, the fundamental codex of postbiblical Jewish law. The Mishnah states:

> These are the obligations which have no bounds; leaving the borders of the field to be garnered by the poor, . . . deeds of kindness, and the *study of Torah*. These are the obligations whose satisfaction yields fruits in this world, and whose abiding riches remain for the world to come: hastening to the house of study morning and evening, hospitality to wayfarers, visiting the sick, dowering the bride, accompanying the dead to the grave, careful attention to one's prayers, and bringing peace between each person and his friend. *But the study of Torah is equal to the sum of all the others.* (Mishnah Peah 1:1, translation and italics mine)

This citation from the Talmud, which is also part of each day's morning worship, emphasizes two elements that are central to the ethics of research: (1) all human beings have an obligation to pursue knowledge, and (2) the pursuit of knowledge is so essential that it has a value equal to the sum of many

other ethical responsibilities, even rather major ones, such as "deeds of kindness" and peacemaking.

In strikingly similar fashion, the 1973 Code of Research Ethics of the American Psychological Association (American Psychological Association 1973) emphasizes much the same two points. The code's introduction and summary statement declares: "We begin with the commitment that the distinctive contribution of scientists to human welfare is the development of knowledge and its intelligent application to appropriate problems. Their underlying imperative, thus, is to carry forward their research as well as they know how" (American Psychological Association 1973, p. 7).

In discussing the "balancing of considerations for and against research that raises ethical issues," the code states:

> Whether a particular piece of research is ethically reprehensible, acceptable, or praiseworthy - taking into account the entire context of relevant considerations - is a matter on which the individual investigator is obliged to come to a considered judgment in each case... the investigator needs to take account of the potential benefits likely to flow from the research in conjunction with the possible costs, including those to research participants, that the research procedures entail. (American Psychological Association 1973, p. 11)

Thus, it is evident that morally concerned scholars separated in space by nearly half the globe and in time by two millennia have arrived at the view that it is not just the scientists' privilege to produce and disseminate knowledge, but their obligation to do so, and that any attempt to resolve this dynamic tension between knowledge seeking and other moral obligations must always involve a balancing of the two weights as relative equals, not a total surrender of one to the demands of the other. When science is asked to yield to the alleged moral good, Copernicus and Galileo are required to abandon their heliocentric views of the world in favor of a geocentric view which, it is claimed, will continue to provide hope and comfort to the poor and oppressed. When humanistic and compassionate sensibilities capitulate fully to science, hideous experiments are performed with human subjects in death campus by Nazi doctors.

There is thus no easy way to prejudge scientific versus other humanistic considerations in the abstract. As the APA ethical code cited above indicates, the issue must be resolved on a case-by-case basis, and with due regard to both human costs and human benefits.

The P & C study does have certain costs. In addition to the problem of deception, there is the imposition on the time and energies of reviewers and editors. Persons in both these groups are often busy and scientifically productive individuals whose efforts in the review process are largely voluntary and unpaid. Finally, there is also the discomfort and perhaps embarrassment that will be felt by these individuals when they read the findings, which they cannot find flattering.

As for the study's benefits, these are potentially considerable. The research is essentially "metascientific," since it deals with the science of the way we do science. Specifically, we are here concerned with the question of how knowledge, once produced, becomes the property of other scientists. And, if we agree with Ziman (1968) that the essence of science is the *publication* of new knowledge, then there can be little question that this research goes to the very heart of the scientific endeavor, not just in psychology but in other disciplines as well. If, as the study suggests, publication is in significant measure a function of the prestige of investigators or their institutional affiliations, we have little assurance that what is called "good research" is indeed good, and what is termed "bad research" is indeed bad. Under these circumstances, what Kuhn (1970) calls a "paradigm" may be nothing more than an orthodoxy based partly on faddishness, partly on sometimes unwarranted elitism, and, to a degree that is unknown, partly on genuine merit. The corollary of this conclusion would be that important contributions to knowledge in the form of paradigmatic shifts may often be lost until and unless they are proposed by the "right people."

What happens to science when certain doctrines become established as orthodoxy and contradictory views are seen as heresy? Lysenkoism is one example of the kind of aborted science that results when government defines what will and what will not be taught. But we all object vigorously and self-righteously to such external influences. The greater danger, perhaps, lies in biases that are part of the very fabric of the hierarchical academic establishment. Such were the forces that confronted Mendel as he carried out his epic genetic research with peas. After being pressured by Carl von Nageli, an eminent botanist of his time, to turn instead to studying hawkweed, Mendel pursued hawkweed research fruitlessly and to his monumental discouragement for the rest of his scientific life (McGuigan 1968). Similarly, it was the Aristotelian professors of Italy who, seeing a threat to their favorite cosmological doctrines from Galileo and his Copernican teachings, first denounced him to the Dominicans, and were thus the first (if not the formal) cause of Galileo's woes.

The benefits of the research we are discussing here are not trivial. Making certain that merit is the criterion for publication is what is most likely to lead to the advance of our knowledge. In this benefit, all of us share, even (uncharacteristically enough) the reviewers and editors who served as subjects. As for embarrassment, this is hardly a standard by which research should be measured. Can you imagine Galileo saying to Cardinal Bellarmine, "Your Eminence, I will gladly withdraw my teachings about the sun, if they in any way offend or embarrass the Church." The benefits of the P & C research will become tangible, however, only if specific steps are taken to minimize the now-recognized abuses of the peer-review system. It is to this task that we need to address ourselves, thus obviating the *need* for this kind of research in the future.

B. Mindick is a past Fellow of the Herbert H. Lehman Ethics Institute, Jewish Theological Seminary. Ed.

Designing peer review for the subjective as well as the objective side of science

Ian I. Mitroff

Department of Management and Policy Sciences, Graduate School of Business, University of Southern California, Los Angeles, Calif. 90007

Make no mistake about it; Peters & Ceci's (P & C's) target article is highly disturbing. As a result, it deserves the most serious attention.

Only the most resistant and diehard defenders of the positivistic view of science will take issue with their findings. True, the sample could have been larger, the design expanded to include fictitious high-prestige authors and institutions, and so on. But can we really believe that the findings would have been substantially different? I don't think we can, because for too long too much evidence has been pointing in the same direction. While the P & C study may not be the definitive capstone study on the subject (is there ever a definitive study when it comes to social phenomena?), it is close enough for all practical purposes.

In a word, all the evidence of which I am aware (much of which is cited by P & C) points to the fact that in even the seemingly most rigorously performed experiments - in the physical as well as in the social sciences - there is an undeniable elment of taste, style, judgment, or aesthetics - call it what you will. If this is the case, then no wonder that in general there is so little agreement between reviewers. Why should we expect people to agree exactly in matters of style, taste, judgment, or aesthetics? However, does this make one's work

any less professional or scientific? I believe not. What it does do is make our traditional concepts of science, which have either denied, repressed, or wished away such phenomena, increasingly out of touch with present-day realities, and as a result, less able to cope with the management of a complex social enterprise.

The traditional coping mechanisms of the past were based largely on a central unspoken assumption: that because natural phenomena could for the most part be described and evaluated in impersonal terms, a scientist's work could also be judged in those same impersonal terms. To be clear, I am not suggesting that the impersonal evaluation of a scientist's work according to relatively fixed rules applied impartially to all is wrong or outmoded per se or serves no useful purpose in science whatsoever. Such procedures are vitally necessary to the very concept of professional evaluation, not just to science itself. However, I believe it is clear that such procedures are no longer sufficient in themselves.

It is undoubtedly true that for most scientists their primary energy is object centered, that is, directed toward physical and social phenomena considered as objects. Scientists are not primarily subject centered, that is, self-reflective or interested in interpersonal behavior (Mitroff 1974), except perhaps in their personal lives, or when interpersonal behavior happens to have an impact on their careers. Even in the latter cases, their interest is not likely to be professional, that is, based on what we have learned about human behavior from social science. As a result, it is not surprising to find that scientists tend to want to think of their own behavior and that of others in the very same impersonal terms in which they think of the phenomena they study. If scientists are often accused of depriving their subjects of their humanity, they are consistent at least in that they are inclined to do the same to themselves.

My point is that if one accepts, as I obviously do, the substance and the importance of the P & C study, then the critical question is, What do we do? The natural tendency is always to specify more of the same: tighter review procedures, and the like. However, doesn't the P & C work show the futility of this? Why should new procedures of essentially the same kind as the old be expected to alleviate the problem, especially when one begins to suspect that the very form of the procedures may themselves be part (or at least a reflection) of the problem?

I believe that the situation calls for some new and bold *policy* experiments in science, not just more experiments directed toward "establishing the phenomenon." To that end, I would like to share the following suggestion. (I do not pretend in the least that it is feasible or that it would be better than what it is intended to cure.)

If the intuitions and feelings of scientists play such an indispensable role in what they decide to investigate and how they go about doing it (experimental design, techniques, etc.), then we must allow scientists to share these sides of themselves with us as well. Such expressions should be considered an integral part of their work and its reporting, so that their work can be properly judged in its entire context.

For a long time, I have thought of writing a paper with the following format. Each page of the paper would have a line down the middle. On the right half of each page (perhaps corresponding to what is currently being attributed to the left half of the brain) would appear the standard, traditional account of the most tightly controlled inquiry I am capable of doing. On the left half (right brain) would appear a blow-by-blow, stream of consciousness account of what I thought, felt, and so on, as I went through the pain, sweat, and joys of doing the study. There would be no attempt to tie these two sides together. They would merely sit there together, existing with and without the other. No comment would be given.

The time is way, way overdue to acknowledge both the "right" and the "left" halves of science (cf. Bruner 1962, pp. 2–5). The time is even more overdue to do something about bringing them together.

Rejecting published work: It couldn't happen in physics! (or could it?)

Michael J. Moravcsik
Institute of Theoretical Science, University of Oregon, Eugene, Ore. 97403

Being invited to contribute a commentary gives one a rare chance to go out on a limb. In doing so, one takes a chance of having, at some future time, to eat one's words. It would be delightful and instructive indeed if an experiment similar to that described in Peter & Ceci's (P & C's) target article were carried out in physics. In the absence of this, however, let me suggest reasons why the situation described in the paper is highly unlikely to arise in physics.

Let me make two preliminary (perhaps unnecessary) comments: First, my suggestions do not imply a claim to, or an exhibition of, the "superiority" of physics over, say, psychology. Fields of knowledge and communities of scientists all have their own characteristics, different from each other, and all subject to change with time. To recognize these is not tantamount to ranking fields.

Second, the suggestions do not imply that the peer-review system is ideal and perfect in physics. On the contrary, its operation is constantly discussed, and modifications in it are always being suggested, and sometimes even undertaken. At the same time, it seems (I believe to most physicists) that the peer-review system in physics is consistent enough so that the enormous internal inconsistencies suggested by P & C's experiment in psychology are highly unlikely to develop in physics.

I want to mention five reasons why this is so, because I believe they are illuminating and stimulating from the point of view of the science of science.

The first is, on the surface, a purely mathematical reason: The rejection rate in the best physics journals is more like 20–30% and not 80%. Therefore, using the same statistical arguments with which P & C exclude a purely stochastic explanation of the phenomenon in psychology, one can easily see that a similar situation in physics is unlikely purely on statistical grounds. It is interesting to remark, however, that the low rejection rate in physics is not an accident, but results from a much firmer consensus in physics concerning what is "right" and what is "wrong," and to the fact that physics is much more cumulative, so that even small contributions are judged useful additions to the structure of knowledge.

The second reason is sociological: The subfields of physics are well delineated, its research specialties are well formed, and the community within each is highly interactive. Hence bogus people with bogus institutional affiliations would be much more readily recognizable as fake. Counteracting this trend, however, is the fact that physics is a more international field than some others, and hence an article could more easily be resubmitted using fake authors from an institution in a faraway (and scientifically undeveloped) land, since these faraway places tend to be very much isolated from the center of gravity of the scientific community.

The third reason may be called material: On the whole, research in physics requires more resources than many other fields of human inquiry and hence the institutions at which research in a given subfield could possibly be carried out are fewer and better defined. This is particularly true for experimental work in physics, which is equipment intensive (especially in specialties like my own, namely elementary particle and nuclear physics), and where often a given type of experiment could only be performed in two or three places in the

world. Even in theoretical physics geographical concentration is sufficiently marked that an unknown name coupled with an unknown institution would prompt a referee to investigate its existence, if for no other reason than to make contact with the new researcher.

The fourth reason is epistemological: Since physics is more cumulative and the arrow of its evolution more discernable, a research topic from three years ago would, in many fields of physics, be significantly out of date, and the field substantially closed. The smaller the subdivision of fields and research topics we make, the faster this outdating becomes. It is therefore more likely that the resubmitted articles would have aroused suspicion merely because of their focus of interest, and their omission of the most recent references.

The fifth and final reason pertains to communication among physicists: The professional "grapevine," often formalized in preprints distributed to workers in a subfield, is so well developed in physics that one rarely sees a paper published in a major journal that one had not learned about some months previously, through preprints, over the telephone, at meetings, or from seminar or colloquium speakers.

None of these effects is so overwhelming that exceptions could not occur to it. Yet, it would be astounding if one single experiment would hit upon such a rare exception. I still hope, though, that somebody will feel challenged to try to prove me wrong in this prediction! [*See also Moravcsik 1980. Ed.*]

Reliability, bias, or quality: What is the issue?

Katherine Nelson
Department of Psychology, City University of New York Graduate Center, New York, N.Y. 10036

There are three levels at which one can discuss the findings of Peters & Ceci (P & C): the level of unreliability, the level of bias, and the level of quality.

In my experience as an associate editor (for *Developmental Psychology*), I have found that the distribution of quality of submitted manuscripts is much like that in many other lines of endeavor, for example, student admissions to college or grading of class performance. There are a few outstanding, clearly publishable pieces, a few that are clearly unpublishable in any journal, and then the majority, which fall in the middle where a decision could go either way. Some of these are "clean" methodologically and stylistically but are of limited interest. Others are "dirty" - they contain methodological flaws, or are poorly presented - but of potential interest to many. In either case the vast majority are likely to find a journal outlet eventually. The only question for the referees and editor is whether they can be made suitable for *this* journal. It is well known that for many journals an article that is initially rejected can eventually be published in that same journal if the author is persistent and attentive to the reviewer's criticisms. There is clearly no single standard of publishable versus not publishable.

Thus with respect to unreliability P & C's findings do not greatly surprise me. In the great middle range of manuscripts judgments may shift in either a positive or negative direction depending upon the referees. The fact that the shifts here were overwhelmingly negative I would attribute to the ability of referees to find something to complain about in every manuscript they receive.

With respect to bias, however, the implication is that the manuscripts were rejected the second time around because their institutional affiliations (and possibly individual identifications) had been changed. This is precisely the situation that blind reviewing was devised for. I work for a journal that practices blind reviewing, and am favorably impressed by its success. Occasionally an author's identity will be guessed (especially if the line of work is well known), but more often it is successfully masked. More important, the bias *against* unknowns cannot surface here because it is difficult to establish any author's identity.

Unfortunately, grant proposals, unlike journal reviews, cannot use blind reviewing since part of the question is the grantee's competence to carry out the proposed research. Other solutions need to be sought, and vigilance against bias must be constant.

To me the most revealing aspect of this report concerns the quality of published manuscripts. At first glance, one would conclude that editors and reviewers must be deeply chagrined at the extent of their ignorance of recently published work in their specialty areas. But from a different perspective, one might derive a different conclusion.

The focus of P & C's report is on the "false negatives," that is, the fate of publishable articles that are judged unacceptable. But it could just as well (and more appropriately, I believe) have focused on the "false positives," the publication of unacceptable manuscripts. Unfortunately, given the push to publish for purposes of promotion, and given the proliferation of journals, even a high rejection rate does little to increase the probability of high quality in even the best journals. Most published manuscripts will be soon forgotten by editors, reviewers, and the general reader simply because they are eminently forgettable. To locate the few important and lasting contributions among the many of little value has become a major barrier to scholarship, one that is little helped by computer searches or abstract services. When trivia become the norm, what is "publishable" declines in value for all concerned. This, I think, is the most important ramification of this study.

While this conclusion may seem unduly harsh and derogatory with respect to the field as a whole, I do not mean it to be. Most researchers, I believe, are truly interested in their work and value it above its promotion potential. However, the structure of psychological research ensures that much of the work that comes out of our laboratories is addressed to questions that are of interest only to others working in the same narrow vein. As with a crossword puzzle, the problems may be fascinating in their own right and the solutions satisfying, but the results may be of no significance to anyone in the real world. Unfortunately, editors cannot create significance on their own; but without it, it is unlikely that the results of test number 6 of hypothesis X from laboratory Y, manipulating an as yet untested variable will be long remembered even by those whose job it is to keep up with the field.

This very basic problem has been discussed from time to time in the *American Psychologist* and other psychological journals. I expect that it is not unique to psychology. It will not, unfortunately, be resolved by any attempts to improve the reviewing process itself.

What is the source of bias in peer review?

Ray Over
Department of Psychology, La Trobe University, Bundoora, Australia 3083

The demonstration by Peters & Ceci (P & C) that editorial evaluation of manuscripts is biased by the information available to reviewers about the identity and affiliation of authors will confirm dark suspicions that are prevalent within the scientific community (see Lindsey 1978; Mahoney 1976). P & C join with Mahoney and his associates (Mahoney 1977; Mahoney, Kazdin & Kenigsberg 1978) in focusing attention on the validity, and not simply the reliability, of reviewing. They provide a neat demonstration that agreement between reviewers as to the merit of a manuscript provides no guarantee that the assessment is correct. In their study, the reviewers who evaluated a manuscript when it was first submitted presumably agreed in

favoring publication, just as the reviewers who evaluated the same manuscript when it was resubmitted agreed in recommending rejection.

In view of what is known about halo effects (for example, Nisbett & Wilson 1977), it would be surprising if reviewers were able to disregard indirect information on the research standing of authors when assessing manuscripts. At the same time, although the data reported by P & C call into question the validity of editorial judgment, the limited design that they employed makes it difficult to determine which factors led to manuscripts being accepted on one submission but rejected on another. For example, is it the author's actual or perceived research productivity and impact that is important (see Mahoney et al. 1978)? How influential are invisible collegial connections relative to university affiliation and status? In fact, it is not necessary to assume from the results reported by P & C that access to information about author characteristics produces reviewer bias.

In discussing their findings in signal detection terms, P & C imply that information about the author causes reviewers to shift their criteria for acceptance. However, editors neither distribute manuscripts randomly to reviewers, nor necessarily give equal or consistent weighting to recommendations made by all reviewers (see Lindsey 1978). The identity, status, and affiliation of the author may be important determinants not only of who will be asked to review the manuscript, but of the action that the editor will take when recommendations have been received. The primary concern of the reviewer is with the standard of a single paper. The editor not only has a commitment to quality control, but a vested interest in the prestige of the journal. Just as authors attract kudos in terms of where they publish, so journals undoubtedly gain prestige in terms of whom as well as what they publish. The bias noted by P & C may reflect to a greater degree the selective practices of editors than a criterion shift on the part of individual reviewers. The reviewers of the resubmitted manuscripts might well have recommended rejection of the originals, if in fact the originals had been sent to them for evaluation.

Incidentally, if there is reviewer bias, the results reported by P & C are better conceptualized in signal detection terms as discriminability rather than as criterion effects. Gottfredson (1978) found that reviewers agree as to the characteristics a manuscript must possess to be recommended for publication. Prominent characteristics include reporting a sound research design, methodology, and data analysis, the primary grounds on which the manuscripts resubmitted by P & C were rejected. In practice it may be that reviewers attend to these characteristics more when evaluating low-status than high-status authors (discriminability), rather than noting the defects equally but deciding only for low-status authors that they constitute grounds for rejection (criterion).

Although further research employing experimental and quasi-experimental designs is to be encouraged, editors' files already contain potentially important information on the review process. It should be possible to establish in a rigorous manner whether the steps taken by editors in processing manuscripts (distribution to reviewers, follow-up of recommendations) have favored specific categories of authors. Attempts also need to be made to identify factors that determine the composition and dynamics of an editorial board, as well as the recruitment of reviewers. P & C cast doubt on the expert status of reviewers, since in their study the majority failed to detect that the manuscript under review had recently been published in the same journal. The fault may instead lie with editors who have mistaken beliefs about the specialized research interests of those from whom they seek reviews. Again, the issue of the competence and appropriateness of reviewers can be assessed through analysis of records on file.

P & C concentrated on bias attributable to reputation and affiliation. In the 1970s there was a marked increase in the proportion of papers written by women that were published in psychology journals. Women were also recruited in large numbers to editorial boards (Over 1981). The possibility that assessments made by men and women reviewers differ as a function of manuscript and author characteristics needs to be considered. Although there have been reports that men and women judge the performance of women differently, evaluations tend to be influenced by task, situational, and perceived personal characteristics in interaction with the sex of the person who is being judged (see Levenson, Burford, Bonno & Davis 1975; Pheterson, Kiesler & Goldberg 1971). However, in an analysis of book reviewing, Moore (1978) found that reviewers tended to be more favorable toward books written by own-sex authors than toward books written by other-sex authors, although reviewers of both sexes expressed more positive opinions about books written by women than about those written by men.

The remedies for review bias noted by P & C need to be carefully evaluated. For example, blind reviewing has obvious virtues, but typically the editor who distributes manuscripts and later acts on recommendations from reviewers is not blind (see Lindsey 1978). Further, the method is potentially corruptible, since skilled writers can indirectly reveal their identity, or hint that they are someone (a high-status scientist) who they are not. Although increasing consensus between reviewers is a desirable objective, reliability must not be improved at the expense of validity. Despite its limitations in design and data, P & C's paper will function as a significant catalyst by focusing attention on important but poorly researched questions relating to the validity of editorial evaluation.

Biases, decisions and auctorial rebuttal in the peer-review process

David S. Palermo
Department of Psychology, Pennsylvania State University, University Park, Pa. 16802

I have but a few comments to make on the Peters & Ceci (P & C) paper. First, I am both impressed with the research and chagrined with the results. I am chagrined because I have often been impressed with the ability of my colleagues in the field of psychology to set aside their own personal biases and theoretical viewpoints when evaluating the work of others who have applied for grants and fellowships or submitted papers for editorial review. It is clear to me that this ability has been demonstrated in many instances by many individuals. Furthermore, I know that persons at distinguished universities do have their grants and papers rejected at times, for I have been a party to such decisions. I also know that I, and some other editors, have a policy of applying more lenient criteria for first submissions by young authors, regardless of their institutions. The research of P & C, however, suggests to me that my observations may have involved a bit of selective perception or been based on a rather biased sample. It is discouraging to me to find clear evidence that irrelevant factors may play such an important role in these decisions.

Second, after reading an earlier version of the P & C manuscript, I suggested the possibility that the research reported in the resubmitted papers might by the second review have been incorporated into the apperceptive mass of the reviewer and, therefore, rejected because it was old hat. P & C have incorporated that hypothesis into their discussion and note that there is no evidence to support my hypothesis in the comments of the reviewers who rejected the manuscripts. Furthermore, it is disappointing that the care with which editors try to select reviewers fails to the extent that the reviewers in this research gave little evidence of recognizing published articles in their areas of expertise.

It should be noted, on the other hand, that one of the problems in making a decision that is based primarily upon an underlying tacit knowledge of what makes a significant contribution to the scientific literature is that one must develop a conscious rationale for the decision. The review reflects the reviewer's efforts to justify a decision made in terms of a system of rules of which the reviewer is unaware. Frequently, the conscious reasons cited in reviews have little to do with the underlying cognitive processes that brought about the decision. Thus, editors are often provided reviews that are in agreement with respect to the decision but in which the different reviewers' rationales focus upon disparate aspects of the paper. Unfortunately, P & C's research suggests that the tacit system of the reviewers may take into account the institutional affiliation of the author as a part of the decision-making process.

Finally, I would note that the journal I edit gives authors the option of blind review but does not require blind review. Only a very small percentage of authors ever takes advantage of that opportunity. Clearly authors from less prestigious institutions should choose blind review if they take P & C's research seriously. Perhaps I will need to insist on blind review for the benefit of the journal. This matter seems particularly important at this time, when many very competent young scientists are taking academic positions at institutions they would not have considered a few years ago.

As an editor, I have always tried to keep in mind the fallacies of the peer-review system as I thought I knew them. Like most editors, I usually reevaluate papers when authors inform me that they think injustice has been done. Sometimes authors do not take advantage of that part of the system. Sometimes they should, for I have had the opportunity of correcting some of my errors. I do not wish to play God because I know the system is fallible. The article by P & C makes clear the humanness of the enterprise. Let us hope that it results in more humane decisions in the future - our science may depend upon it.

D. S. Palermo is editor of Journal of Experimental Child Psychology. *Ed.*

Reviewer "bias": Do Peters and Ceci protest too much?

Daniel Perlman
Department of Psychology, University of Manitoba, Winnipeg, Canada R3T 2N2

Peters & Ceci (P & C) address the issues of reliability and bias in the journal reviewing process. Changing the names and institutional affiliations of the authors (from high to low status), they resubmitted 12 previously published articles. Their ploy was only detected in three cases. In eight of the nine remaining cases, the resubmitted manuscripts were reviewed and rejected.

It is easy to hurl criticisms at their efforts. (a) Given that the sample of reviewed articles was very small, one wonders if the results are really representative. If another set of previously published articles was resubmitted, would a higher proportion be judged acceptable during the second evaluation? (b) Certainly we have known for some time that interrater reliabilities are far from perfect and that biases affect reviewers. Have P & C demonstrated anything that novel? (c) Undoubtedly, a better research design could have been employed. By starting with already published articles, P & C may have included some "false positives" (i.e., poor manuscripts that were erroneously accepted during the first review). Thus, one would expect some of these manuscripts to be rejected during the second review cycle. P & C also changed the names and affiliations on the papers before resubmitting them. These changes were confounded with the time at which the manuscripts were submitted plus modest expository alterations. All in all, one might better call P & C's endeavors a demonstration rather than an experiment in the classical sense. A relatively simple improvement in the design would be to submit each in a series of unpublished manuscripts to two comparable journals (i.e., once identifying the paper as written by a high-status author; once identifying the paper as written by a low-status author. (d) Finally, one wonders whether P & C's ploy was ethical.

But despite these criticisms, P & C's article is a provocative piece. It will undoubtedly be often cited and much discussed. Why? It is a striking demonstration; the authors are articulate; and the issues they address are of personal concern to many scholars.

One must agree with P & C's view that reviewer bias, incompetence, and unreliability are disquieting. As they suggest, ways of reducing these effects should be sought. Young scholars as well as those from less prestigious institutions should have fair access to journal space. Blind reviews are a step in the right direction. Another innovation for improving interrater reliabilities may be to have reviewers simultaneously rate sets of papers rather than periodically reviewing individual papers over a long time span. Simultaneous consideration of several manuscripts provides reviewers with a salient, comparative framework. In at least one previous study of reviewer reliabilities (McReynolds 1971) this procedure worked well.

In agreeing with much of P & C's argument, one might reach the conclusion that undue favoritism is shown toward authors from high-status institutions. It is this conclusion that I wish to examine further. Is preferential treatment of authors affiliated with high-status institutions an unwarranted form of favoritism? Or is it rational behavior?

Let us begin with the assumption that there are differences in the quality of submissions to journals. An editor's task is to decide which papers, among the several manuscripts available, are the best for publication. Editors typically rely heavily on reviewers' assessments in deciding which papers to accept. When reviewers are asked what criteria they use in evaluating manuscripts, one of the most important dimensions they cite is the author's identity. If reviewers believe authors have "justifiably strong reputations," they are positively influenced toward accepting their papers (Rowney & Zenisek 1980).

To determine whether this explicitly acknowledged influence is an undue form of favoritism, one needs a criterion measure. Impact, as assessed via the *Social Science Citation Index* (*SSCI*), is such a measure. Suppose papers submitted by scholars from prestigious institutions are, in fact, typically better than articles submitted by scholars affiliated with less prestigious institutions. Then the papers submitted by scholars affiliated with high-status institutions should eventually be cited more frequently. Furthermore, accepting a higher proportion of papers from these institutions would *not* reflect unfair preferential treatment. Instead, it would reflect justly deserved preferential treatment and be indicative of rational behavior on an editor's part.

To test the key aspect of this analysis, 60 articles from the *Journal of Abnormal Psychology* (circa 1975) were selected. This journal has a policy of nonblind reviewing (see Table 1). Half the articles were by scholars affiliated with high-status institutions, half by scholars affiliated with low-status institutions. Institutional status was defined in terms of the Roose-Anderson (1970) ratings of U.S. psychology departments. The high-status institutions were the same as those used by P & C plus the Universities of Michigan and Pennsylvania. The low-status institutions were those rated below 2.5 or not rated at all.

To assess the impact of these articles, the number of times they were cited in 1980 was determined via the *SSCI*. The results were clear-cut. The articles by scholars affiliated with high-status institutions were cited considerably more often

than the articles by scholars at low-status institutions. The means were 3.77 and 1.40, respectively. To avoid the influence that a few very frequently cited articles might have on the means, the articles were classified into those cited two times or less and those cited more than two times. A chi square analysis ($\chi^2 = 6.23$) indicated that the results were statistically significant. Comparable results (means 9.52 and 2.56, respectively; $\chi^2 = 5.55$) were obtained using a second sample drawn from the *Journal of Personality and Social Psychology*, a journal with a policy of blind reviewing (see Table 1).

One might maintain that scholars at elite institutions form an invisible network such that the high citation rate of their articles reflects an in-group, self-citation phenomenon. This is unlikely, however. Despite the importance of scholars at elite institutions, the vast majority of published articles are written by people with lesser affiliations. Thus, the frequent citation of scholars at elite institutions undoubtedly reflects a pluralistic judgment of the value of work done at high-status institutions. Therefore, it appears that an institution's prestige is a valid predictor, and editors may be justified in using this as a factor in their decision making.

Advocates of blind review, however, may still object to using either institutional affiliation or an individual's reputation as criteria in selecting articles. They could claim that the excellence of a manuscript should not only be apparent over time, it should also be immediately apparent without the aid of status cues. Thus, even with a blind review process, assessors should identify a higher proportion of items submitted by scholars at prestigious institutions as worthy of publication.

I find this a compelling rebuttal for many editorial situations. It is especially compelling when one is judging a completed manuscript. In these instances, blind processing plus every effort to achieve valid, reliable reviews should result in high-status scholars having a higher probability of success yet all contributors being treated fairly.

But there are other editorial conditions under which no finished paper is available in advance for judging. These include BBS's practice of soliciting commentaries on articles, as well as book editors' invitations to colleagues to contribute chapters. In situations of uncertainty such as these, reputation and institutional affiliation appear to be among the sensible criteria to use in selecting potential contributors. [*See editorial note following this commentary. Ed.*]

In conclusion, P & C protest too much. If journals were showing undue favoritism to scholars affiliated with prestigious institutions, then many of their published articles (especially in journals with nonblind reviewing) would be false positives. Carrying P & C's analysis to an extreme, one would then expect the articles by this set of scholars to be less good and therefore eventually to have less impact on the field. Certainly the data in the present study show this is not the case.

BBS's quality control policy for its solicited commentaries may be best illustrated by the following excerpts from our Instructions: *"Please note that although commentaries are solicited and most will appear, acceptance cannot, of course, be guaranteed . . . all original data will be refereed in order to ensure the archival validity of BBS commentaries . . . BBS also reserves the right to edit commentaries for relevance and style . . . portions of commentaries redundant with the target article or with other accepted commentaries may have to be deleted by the editor [with] . . . priority . . . assigned in terms of order of receipt." Ed.*

Table 1 (Perlman). *Article citation rate as a function of author's institutional status*

Author's institutional status	% and number of articles cited: More than two times (high citation rate)	Two times or less (low citation rate)	
Journal of Abnormal Psychology (nonblind review)			
High (N = 30)	47% (14)	53% (16)	
Low (N = 30)	17% (5)	83% (25)	$\chi^2 = 6.23$ ($p < .025$)
Journal of Personality and Social Psychology (blind review)			
High (N = 30)	57% (17)	43% (13)	
Low (N = 30)	27% (8)	73% (22)	$\chi^2 = 5.55$ ($p < .025$)

Improving research on and policies for peer-review practices

Richard M. Perloff[a] and Robert Perloff[b]

[a] *Department of Communication, Cleveland State University, Cleveland, Ohio 44115 and* [b] *Graduate School of Business, University of Pittsburgh, Pittsburgh, Pa. 15260*

It should be stressed at the outset that Peters & Ceci (P & C) have conducted a well-planned and executed study, written clearly and persuasively, and addressed to a problem of considerable significance not only to psychology and the other behavioral sciences, but probably to all disciplines and fields of inquiry whose research and scholarly output find their way into the scientific, technical, and learned literature as well. Moreover, their fundamental hypothesis that impressions and judgments are influenced by a variety of background factors, including prestige, is sound, and supported by the social psychological literature. Hence, the results they report are reasonable, at least intuitively, even though the sheer magnitude of their results (the rejection of eight out of nine manuscripts originally accepted and published by the same journals to which the papers were resubmitted) is mind-boggling and, on the face of it, an embarrassing indictment, if not of the peer-review system at large, certainly of the system used by the eight journals whose editors were duped and perhaps humiliated by P & C.

Unfortunately, however, there are several methodological problems that may cast doubt on the validity and generality of the findings. P & C argue that the change in reviewer recommendations reflects a bias in favor of prestigious institutions. There may be reason to question this interpretation. First, one cannot know whether the results reflect a "status bias" until one includes experimental conditions such as the following: (1) articles accepted from low-status institutions resubmitted under the by-line of equally low-status universities; and (2) articles accepted from high-status institutions resubmitted under the by-line of high-status universities. Without such controls, it is impossible to argue that the findings reflect the status bias P & C suggest. If, for example, articles written by psychologists at the Tri-Valley Center for Human Potential were accepted initially and then rejected by a second set of reviewers the second time around, it would be plausible to argue that individual differences in reviewer preferences and biases were operating; that is, for good or ill, and for whatever crotchety reason, reviewers may nowadays be much tougher cookies than they were three or four years ago. These and other controls are suggested in Table 1.

P & C appear to be aware of such a possibility when they suggest that their results may be attributable to individual reviewer differences from one to the other review period. However, they dismiss this too quickly.

Table 1 (Perloff and Perloff). *A fuller and more interpretable design for a Peters & Ceci-type study of peer-review bias*

	Articles previously rejected		Articles previously accepted (and published)	
	Original source of submission			
	Nonprestigious	Prestigious	Nonprestigious	Prestigious
Prestigious resubmission	I[a]	II[a]	V[b]	VI[b]
Nonprestigious resubmission	III[a]	IV[a]	VII[b]	VIII[c]

[a] These conditions would be sensitive, awkward, and difficult, but not impossible to attempt.
[b] Reasonable and appropriate conditions to attempt.
[c] Only condition examined by Peters & Ceci.

While individual differences in reviewer *competence* may not be operating, it is plausible that the second set of reviewers came from higher-status institutions than the first set, a state of affairs that P & C themselves acknowledge in their quotation from M. D. Gordon (1980), but to which they apparently give too little credence. Such reviewers might be more critical and wary of manuscripts and have more confidence in their own reviewing skills. The important point is that no checks were provided to ensure that both sets of reviewers were equivalent in demographics, competence, or status of institutional affiliation. Without such information, one cannot reject the explanation that the current crop of reviewers was more critical than the first. Despite P & C's plea to the contrary, it would seem as if the small reported agreement among reviewers would support rather than refute this interpretation (Watkins 1979).

Finally, it is interesting to speculate about what the pattern of findings would have been if the original authors' institutional prestige had been more fully varied. That is, reviewers may be resentful, jealous, or envious of psychologists from a Harvard or a Stanford, thereby making an extraordinary effort to be critical in their reviews. On the other hand, when reviewing articles from schools of slightly less - though still respectably high - prestige, reviewers may not feel any resentment, and may, therefore, apply a more lenient standard, as P & C suggest. In any event, it is quite possible that the relationship between institutional prestige and reviewer bias is not at all linear, but rather curvilinear, or at least more complicated than P & C suggest.

Recommendations for improving peer-review practices.

Blind reviews. As P & C themselves recommend, blind reviews, wherein the reviewer does not know whose manuscript is being reviewed, are clearly called for if it can be shown (as P & C may in fact have shown, the criticisms proposed in this commentary notwithstanding) that identifying the author does not add to or may even detract from the review's validity.

No reviews. Not entirely facetiously, it could be proposed that one way to remove the bias associated with the reviewer's awareness of the author's identity and affiliation, is to have no review at all. This caveat emptor approach might be viewed as a nod to the free market of ideas. Let millions of flowers bloom. All one need do to get published is to write an article, submit it for publication, and pay for its publication. In this way, all individuals, whether from recognized or unrecognized institutions, would be assured of having their words immortalized. Those articles that catch fire and are cited might come from beggars, thieves, princes, or future Nobel laureates. Let it all hang out: the garbage, mediocrity, and the crown jewels. One could argue that all people are "created equal," endowed with such inalienable rights as the pursuit of truth via totally unrestricted opportunities to publish what they wish.

Paid reviews. But perhaps the most promising route for improving peer-review practices, for funding as well as for publication, is to pay reviewers adequately for their time. If reviewers are paid, they should feel a greater sense of responsibility to review thoroughly and impartially whatever manuscript they are asked, and are competent, to review. Because currently the better reviewers are probably the busier ones, more frequently invited to prepare reviews, they are quite naturally likely, consciously or otherwise, to look for shortcuts; they may tend to spend less of this unpaid time closely scrutinizing articles from high-status authors, or they may examine manuscripts from lower-status authors primarily to find obvious methodological flaws, perhaps overlooking the strengths in some of these papers.

Where would the money come from to support such a paid review system? Where it always comes from: authors' institutions, their research funding, or their personal resources. This would cost authors, to be sure, but would it not be worth the cost if it assured them a fairer shake?

2004: A scenario of peer review in the future

Alan L. Porter
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Ga. 30332

Surely, Peters & Ceci's (P & C's) article was one of the stimuli for the radical restructuring of the scientific information system over the last two decades of the twentieth century. Looking back it is hard to imagine the days when anonymous referees acted as a scientific inquisition, deciding what would and would not appear in the "open" literature. Recognition that the referee system was both unreliable and biased combined with constricted budgets and new technology to undermine the system. It may be useful to sketch the evolution of our present system for those who have not directly experienced it.

The 1980s saw growing pressures to change the scientific information system. Journals had multiplied to fill any emergent subfield niches, putting grave pressure on library and individual budgets, in turn yielding many low-circulation operations. Yet, novel or interdisciplinary work was often difficult to get published as invisible colleges protected their well-traveled research trails (Chubin & Connolly 1982). New information technology, such as microfiche and home com-

puter terminals with data-base access, was also becoming hard to ignore. The challenges to the sanctity of peer review tipped the scales.

In 1989 the American Psychological Association (APA) announced its intention to cease publication of primary research journals, substituting abstract publication with full papers copied on request. That led to the premier scientific politics event of the century with the attempted impeachment of the APA officers. As the abstracting scheme got under way, non-APA journals experienced an initial bonanza as prospective authors rushed toward them, and "readers" sought their wares. However, the new system had key advantages over the old: It saved money and improved speed and access to information. Inexorably, the APA system subsumed its competitors, resulting in consolidation with subfield and cross-subfield abstracting publications.

From that point, developments emerged rapidly. The refereeing process was improved by requiring authors to submit data-base searches, along with their manuscripts, covering keywords and citations to references, with provision of abstracts of references. Papers were made available on microfiche to libraries to ease the access problem. The nice touch of providing "reprints" to the authors for direct dissemination to core peer groups smoothed the information exchange process further. Then came the next improvement: "PSYCH."

PSYCH was the new on-line psychological data base, accessible through most any terminal at modest cost. Papers could now be read directly without undue delay (or publication as such). With direct access came the interactive review process. Given the lack of economic constraints, it was no longer necessary to screen manuscripts before publication; the processes of review and dissemination could be fused. The emergent system involved only the barest editorial review before a paper was accepted into the "new research" category with dissemination of its abstract. Automated counting took note of how frequently a paper was requested for reading, and a commentary system allowed direct feedback to the author upon reading of a paper. The author had the opportunity to respond; he could even improve the original paper, properly crediting the feedback received. Both updated and original versions of the papers were maintained, along with a full record of comments received, tagged by anonymous terminal designations.

Secondary journals emerged as a distinctive force as editors sought out important and controversial primary research. Sometimes replications (negative as well as positive) were combined with original results. Joint publications involving authors of original pieces plus key commentators became possible (terminal identity could be recovered on mutual agreement) to synthesize the essential findings. Cross-disciplinary science grew as "publication" ceased to be a severe barrier and computer aids facilitated the joining of research streams.

An improved base on which to evaluate scientific performance also emerged in conjunction with the better information exchange. Tallies could be made on how many people accessed one's articles and what comments were made; and commentators could be queried on the individual's work with provision of a full file for them to reconsider.

As we all know, the net effect has been dramatic. Scientific dialogue has replaced curriculum vitae building as the theme of the information system. The incentive for quality work is the open peer interaction process; more valid measures are available for assessing scientific performance; and tremendous savings are made in time and resources by abandoning the old refereeing, resubmission, and multiple publication outlets schema. There is even talk nowadays of modifying the peer-review process for research funding.

A. L. Porter is coeditor-in-chief of Bulletin of the International Association for Impact Assessment.

BBS plans to begin to experiment with receiving and processing commentaries (and eventually also articles) in machine-readable form via various telecommunications networks. The 1,000-word format of the commentaries and the rapidity with which they are processed make them especially suitable for pilot studies of this sort. The worldwide circulation of the "preprint" of the target article would be a logical next step, as would the generation of type-set quality hard copy from the machine files. With machine readability and telecommunication naturally mediating the Commentary process in this way, the transition to an exclusive soft copy format could then readily be made, if and when the archival scientific literature ever actually elects to do so. Ed.

Reviewer reliability: Confusing random error with systematic error or bias

Stanley Presser
Survey Research Center, University of Michigan, Ann Arbor, Mich. 48109

Peters & Ceci's (P & C's) target article addresses one of the liveliest issues in the sociology of science. In a study that has already received attention (Holden 1980), P & C changed prestigious author affiliations to unknown ones on a sample of psychology journal articles and found that eight out of nine papers were then rejected upon resubmission to the very journals that had published them shortly before. Since all the articles were evaluated by referees aware of the authors' supposed identity (only journals using nonblind reviews were selected), P & C conclude that the results demonstrate a systematic bias on the part of referees in favor of authors from high-status institutions and against those from low-status ones. There is, however, reason to doubt this interpretation.

To begin with, the inference is inconsistent with the only true experimental investigation of this issue with which I am familiar. Mahoney, Kazdin, and Kenigsberg (1978) varied institutional affiliation on an experimental manuscript sent to 68 reviewers for two "behavioristic journals," with half the manuscripts allegedly by an author at a "prestigious university" and half by someone from an "unknown college." They report that institutional prestige had no effect on any of the referee measures (including summary recommendation). By contrast, the M. D. Gordon (1980) study that P & C cite is a nonexperimental one from which no clear conclusion can be drawn. Gordon found that major university referees for two British physical science journals were more favorable to papers from major universities than to those from minor universities, whereas this was untrue for referees at minor universities. Yet this may simply mean that the editors sent different mixes of papers to the two types of reviewers. (For example, the editors may have sent the highest-quality papers - presumably more likely to be by major university authors - mainly to referees at major universities. This would account for the fact that major university referees were more favorable to papers from major universities.) Moreover, Zuckerman and Merton (1973, p. 491), who examined this issue with the files of an American physics journal, found that "referees were applying much the same standards to papers, whatever their source.... the relative status of referee and author had no perceptible influence on patterns of evaluation."[1]

But there is a more compelling reason to discount bias as the explanation for P & C's results: They confuse random error with systematic error. Assuming publication decisions were free of bias, the nature of the evaluative judgments called for would still produce random error. Some good papers would be rejected, and some bad papers would be published. The work of Stinchcombe and Ofshe (1969) provides estimates of the likelihood of these events. Their model of the review process depends on a number of assumptions. They assume that the

quality of articles submitted to a journal is distributed normally. Therefore, if a journal had a 16% acceptance rate, and referees were error free, only papers at least one standard deviation above the mean would be accepted. Of course, referees are far from error free. Stinchcombe and Ofshe assume an interreferee reliability coefficient of .50. In addition, they assume that referees are unbiased, and thus that the validity coefficient (square root of reliability) is .70. With these assumptions, they are then able to estimate the proportion of papers at different quality levels that will be judged to be one standard deviation or more above the mean (and therefore accepted). To take two examples, they conclude that 18% of the papers between the mean and +1 standard deviation will be accepted, whereas 84% of the papers between 2 and 3 standard deviations above the mean will be accepted. The higher the quality of a paper, the greater the probability of acceptance. Because of the very large number of average quality papers, however, a large fraction of the acceptances will consist of papers lower in quality than +1 standard deviation. Stinchcombe and Ofshe estimate this fraction at 7/16. In other words, almost half the papers published will be below the supposed cutoff level.

These results should apply to P & C's work, since they report the average rejection rate of their journals to be 80%, quite similar to the 84% used in Stinchcombe and Ofshe's model. It is likely, then, that of their nine published papers many were lower-quality papers that should have been rejected initially. Moreover, some of the higher-quality papers would be rejected on resubmission simply owing to the unreliability of the review process. Indeed, applying the Stinchcombe-Ofshe model, one would expect only 3.4 of the nine papers to be accepted on resubmission, *without* any alteration in the papers themselves.[2]

In evaluating the difference between 3.4 and 1.0 (P & C's observed result) two factors are important. First, the observed result is based on a very small sample and is thus subject to a fairly large sampling error. (I estimate that the upper bound of the 95% confidence interval is 2.9.) Second, the expected value (3.4) is based on the assumption of an interreferee reliability of .5. Yet as P & C observe, this is at the upper limit of the range of reliabilities found in studies of the matter, the typical figure being considerably lower. Assuming a lower reliability, the expected number of acceptances would be smaller than 3.4. Taken together, these factors strongly suggest that P & C have mistaken random error for bias, and that they would have obtained the same results if they had resubmitted their papers without altering authors' affiliations. It is possible that referees are affected by status considerations (and it is surely sensible to prefer blind reviews to nonblind ones), but P & C's research does not tell us whether status does, in fact, play a role in the reviewing process.

ACKNOWLEDGMENTS

This commentary was written while the author was at the Institute for Research in Social Science, University of North Carolina. Sahni Hamilton provided helpful editorial suggestions.

NOTES

1. None of the other references cited by P & C provides data on how referees evaluate articles written by authors from institutions differing in status.

2. The distribution of the 9 papers and the probabilities of acceptance are:

Interval	No.	Prob.
−1.0 to 0	1	.028
0 to +1.0	3	.180
+1.0 to +2.0	4	.503
+2.0 to +3.0	1	.843

Carrying out the arithmetic, 1(.028) + 3(.180) + 4(.503) + 1(.843) = 3.42.

Reliability and bias in peer-review practices

Robert Rosenthal

Department of Psychology, Harvard University, Cambridge, Mass. 02138

More than 20 years ago I was a young faculty member at the University of North Dakota (UND), the same university with which the first author of the Peters & Ceci (P & C) target article was affiliated 20 years later. The interesting contribution he and his coauthor have made took me back in time. It reminded me of the 15 to 20 articles I had written while at UND that I was not able to publish in mainstream psychological journals. After I had been at Harvard a few years, most of those same articles *were* published in mainstream journals. My anecdote does not demonstrate that journal editors were biased against papers from UND and biased toward papers from Harvard. There are plausible rival hypotheses that cannot be ruled out. My belief, however, is that location status bias may well have played some role in the change in publishability of my stack of papers. That belief is not, however, greatly increased by the evidence of the target article, which I view as an anecdote about bias of only slightly greater utility than my own.

What the target article does show. There are two important lessons to be learned from the empirical components of the target article. The first is that the chances are excellent that previously accepted papers will not be recognized if resubmitted under a less prestigious name. The second is that the chances are excellent that previously accepted papers will not be accepted if resubmitted under a less prestigious name. Both of those results are interesting and important, and I am grateful for the authors' having made them available to the field.

What the target article does not show. There are two things, however, that the target article claims to do, strongly in some places, more mildly in others, that it does not do. It does not enlighten us very much about the biasing effect of authors' prestige, and it does not enlighten us much about the degree of reliability in the peer-review system that was studied.

Bias. P & C imply that their work involved the "direct experimental testing of bias." Any direct experimental test requires assessing the effects on a dependent variable of a manipulated independent variable. The authors *have not employed any independent variable!* At the very least they should have sent out an equivalent number of papers ostensibly written by high-prestige authors. Without that obvious control, conclusions about the effects of authors' prestige are simply not warranted. The prestige-of-author bias hypothesis could have been tested with articles that had previously been (a) accepted, (b) rejected, (c) not even submitted, or any combination of (a), (b), and (c). The most interesting and informative design would have employed all three types of articles, but always with half allegedly written by high- and half by low-status authors.

Reliability. P & C imply that their work examines review reliability. The assessment of reliability of dichotomous decisions (accept vs. reject) requires determining the relation between the variable of decision 1 (e.g., the original decision) and the variable of decision 2 (e.g., the subsequent decision). The authors have not used any variable of decision 1. Although they recognize the lack of articles that had been rejected earlier, that does not change the fact that one cannot compute reliability from only two entries in a fourfold table.

An additional problem complicates the issue of reliability of review: The original measure of acceptance was not the same as the later measure of acceptance. The earlier measure was defined by appearance in the journal; the later measure was defined by a letter of acceptance or rejection. A more proper test of reliability would have required not only the missing data of earlier rejected papers noted above but also the use of the original letters of acceptance or rejection as the measures of acceptance. For many published papers the initial letters were so-called letters of rejection.

Improving peer review. I applaud P & C's suggestion for a systematic evaluation of referees by judges, authors, and editors, and for blind reviewing of articles, an idea I have long found attractive for reasons that should be clear from my opening paragraph, and one that was advocated even earlier by Gardner Lindzey and Kenneth MacCorquodale (Rosenthal 1966).

An ethical question. I feel that the work of the target article was sufficiently important to warrant what I do regard as questionable ethical behavior. The referees and editors had to do a lot of extra work (in total, not so much individually) with no opportunity to opt out if they had wanted to spend their time in some other way. At the very least, I believe that uninformed participants in this type of research should be offered an honorarium of some sort for their professional time investment along with their debriefing. I also believe that ethical transgressions of this type are more warranted when the research has greater potential for yielding answers to the questions posed. Thus I think there would be less of a cost-benefit issue had the research been better designed.

Conclusion: A publication decision. If some future investigator were to send me this published target article to review, I would illustrate further the "unreliability" of the peer review process; that is, I would not accept this article in its present form because it does not do what it promises to do, though it does do other things quite well.

ACKNOWLEDGMENT

Preparation of this commentary was facilitated by support from the National Science Foundation.

Rejecting published work: Similar fate for fiction

Chuck Ross

Santa Monica, Calif. 90404

In 1977 I did an experiment similar to Peters & Ceci's (P & C's) but in the world of mainstream fiction. I typed up Jerzy Kosinski's (1968) *Steps* and submitted it, untitled, to 14 major publishing houses and 13 literary agents. To another 13 agents I sent a letter of inquiry. The highly acclaimed novel, which had won the prestigious National Book Award for fiction in 1969, was rejected by all (including Random House, its original publisher). No one recognized the work, and no one thought it deserved to be published (C. Ross 1979).

It is disheartening that P & C report comparable results with their resubmission of published articles to psychological journals. It is disheartening not only to the mainstream and scientific authors whose quality work is rejected, but also to the rest of us who are consequently deprived of many fresh concepts or new perspectives on old ones. It also brings up questions of how information is disseminated in our society. What are the standards, and should they be reevaluated?

No doubt the answer is yes, a rethinking is in order. One of the problems similar to journal and mainstream fiction publishing is that of response bias. *Steps*, the polished, published novel, was resubmitted under the pseudonym Erik Demos. At the time of resubmission, Houghton Mifflin was Kosinski's publisher. Its response to the manuscript was:

> "Several of us read your untitled novel here with admiration for writing and style. Jerzy Kosinski comes to mind as a point of comparison when reading the stark, chilly episodic incidents you have set down. The drawback to the manuscript, as it stands, is that it doesn't add up to a satisfactory whole. It has some very impressive moments, but gives the impression of sketchiness and incompleteness." (C. Ross 1979, p. 40)

The question I asked was how come Kosinski was no longer as good as Kosinski when Demos was the name on the envelope. Perhaps blind reviewing is also needed at publishing houses that print fiction.

Just as disquieting is P & C's report that they encountered editorial resistance from some of the journals in the course of their research. Couched in this resistance may be a reluctance to change, even when confronted with a system that is not working very well. For example, I concluded from my experiment that the method for looking at unsolicited manuscripts was not working and needed reevaluation. However, much to my surprise, *none* of the publishers agreed with me. A typical response was that of Tom Wallace, then editor in chief at Holt, Rinehart & Winston, and now a senior editor and vice-president at Simon & Schuster. He said that publishers are always looking for good manuscripts, and that readers of manuscripts start off idealistic, but that most of the material is "just gibberish" (C. Ross 1980). Unfortunately, the author can rarely find out who the reader or reviewer is, or what qualifies the reader or reviewer to judge what is good and what is "just gibberish."

In 1976 Viking published *Ordinary People*, the first unsolicited novel it had published in 27 years. Surely there was at least one other work among the thousands of unsolicited novels it had received over that time span that was worth publishing.

Furthermore, none of the publishers thought the issue was important. One editor in chief thought my experiment was "silly, a frivolous exercise." Another executive found it "amusing, a clever trick." Still a third thought the experiment was "kind of naughty" (C. Ross 1980).

One hopes that the editors of scientific journals (psychological and otherwise) that use nonblind refereeing will not toss off P & C's findings with this same air of nonchalance. The issues P & C raise are indeed important.

Rejection, rebuttal, revision: Some flexible features of peer review

Donald B. Rubin

Mathematics Research Center, University of Wisconsin, Madison, Wisc. 53706

Peters & Ceci (P & C) are to be commended for providing an interesting and provocative article on the practice of professional publication. As the authors are fully aware, however, their sample of 12 articles is so small as to provide more a source for anecdotes than a basis for a substantial scientific study. Moreover, the anecdotes would have been even more stimulating if P & C had been able to supply some history for the 12 articles: Had they been previously rejected by other journals before being accepted where published, how many revisions were required before final acceptance, and what were the reasons for initial rejection? All who routinely submit articles for publication realize the Monte Carlo nature of reviews, and thus take advantage of existing opportunities to rebut and respond to criticism as well as pursue alternative outlets for publication.

I find this article particularly interesting because I am currently coordinating and applications editor for the *Journal of the American Statistical Association* (JASA). As a journal editor, I have two primary reactions: first, surprise at the combined ignorance of reviewers as to the recent content of their own journals, and second, concern with the general thesis of the article, which, I believe, tends to be misleading.

Regarding my first reaction, I am surprised that nine of 12 articles were not detected because I doubt that JASA would fail to detect a published JASA paper being resubmitted. I have a board of 10 to 15 associate editors, and after my initial screening of a manuscript to determine whether it is implausi-

ble (in the sense that the content or level of the paper is obviously inappropriate for publication in JASA), the manuscript is sent to an associate editor with interests in the topic of the paper. If the associate editor considers the paper to be plausible, it is sent out for a full review, which means typically two or more additional readers. We often detect redundancy with previously published work, or even sometimes with work submitted to other journals, and I'd be surprised and dismayed if we did not detect a submission that was actually a recently published JASA article. My impression is that the 12 psychology journals involved in the P & C study do not use boards of specialty associate editors; perhaps the use of such boards would alleviate the problem of unfamiliarity with the journals' contents.

Regarding my second reaction, I think that this study may be misleading because it doesn't reflect the true nature of the submission-rejection-revision process that eventually leads to publication. With JASA, as with other journals, eventual publication nearly always involves an iterative process between the author and the editors and reviewers and includes compromises and agreements. Judging from my experience, it is likely that none of these 12 articles was accepted as originally submitted, but rather that the final published versions used in this study were honed to meet the specific criticisms of specific reviewers. It is not at all surprising then, that in order for the articles to be perfectly acceptable to other reviewers, different honing would be required. Thus an important question, not adequately addressed by P & C, is whether the eight initial rejections of the nine submitted versions of the manuscript were terminal rejections or tentative rejections leaving open the possibility that revisions addressing or rebutting the reviewers' concerns might be publishable. Although it is hard to judge precisely from the information in note 4, which briefly describes the reasons for rejection, my impression is that few of the rejections were terminal. Consequently, I expect that if experienced authors (like the ones who actually wrote the papers) had actually submitted these manuscripts, they might eventually have gotten most of them published in the journals to which they were submitted. Of course, these new published versions would be somewhat different from the versions actually published since they would have had to address the concerns of new reviewers. I imagine, however, that as a consequence of responding to additional criticism, the new published versions would be better articles than the versions that were actually published. Articles can always be improved, and part of the subtle agreement between authors and editors is that at some point a decision is made that the author has expended enough energy and that the paper is a contribution worth publishing without further iteration. If the final version had been the original submission, the editor would feel free to demand more, but it certainly would be unfair after several revisions addressed to the concerns of two initial reviewers to ask for revisions addressed to the concerns of two new reviewers, unless these new concerns were quite condemning.

D. B. Rubin is coordinating and applications editor for the Journal of the American Statistical Association.

Anosmic peer review: A rose by another name is evidently not a rose

Sandra Scarr
Department of Psychology, Yale University, New Haven, Conn. 06520

A rose grown in major universities seems to smell sweeter than the same variety submitted to journals from less halcyon fields. Reviewers' senses of smell are rendered questionable by their inability to detect the same scent resubmitted to them and to react with similar pleasure to the experience.

Among the many possible interpretations of these phenomena offered by Peters & Ceci (P & C) is one to which I subscribe: that blind review is a partial remedy for such biases in perception. I want to object, however, to proposed remedies that attempt to make the criteria for editorial decisions more mechanically combinatorial. Like all human judgments, recommendations by reviewers for acceptance, rejection, or revision of manuscripts, I argue, are based on complex weightings of criteria that can be only partially specified.

As an editor (*Developmental Psychology*) over the past two years and an associate editor (*American Psychologist*) for the preceding three, I have found that blind review removes many of the biases in institutional affiliation, professional status, and gender to which P & C's results may be largely attributed. (A new perfume by Coco Chanel doubtless receives more favorable reviews than one introduced by Sam Smith.) In the two journals mentioned, both of which use blind review, more than half - perhaps two-thirds - of the manuscripts are genuinely blind to the reviewers, as evidenced by the frequent tell-tale mistakes they make about the experience, gender, and presumed theoretical positions of the authors. Young reviewers mistake senior investigators for novices, to whom they lecture; many reviewers refer to authors as "he" or "he/she" or "they," inappropriately; and some reviewers argue for a more extreme version of the author's own theoretical position, as though the author might need coaxing. Blind review may create its own problems, such as casting doubt on the methodological abilities of authors with demonstrated competence, who in an identified author system would not have to specify their procedures in such excruciating detail; but it is manifestly more fair to lesser known investigators in lesser known institutions.

My major point of disagreement with P & C is in the recommendation that reviewers and editors be required to specify exactly the criteria for their decisions about a manuscript. Please understand that I try to specify to authors why a manuscript fails to meet my criteria for publication, be it the easy methodological complaints or the more difficult judgments about theoretical issues and importance to the field. Both reviewers' and editors' decisions are based, I argue, on complex human judgments that include weightings of many criteria, much like judgments of personal attractiveness, ratings in wine tastings, and preferences among floral scents. Although we may not like to categorize scientific reports with ambiguous matters such as wine tasting, personal attraction, and perfume preferences, I think that there are important similarities. The final recommendations of reviewers and decisions by editors are not based on any simple sum of component parts (so much smile, body, or sample size added to so much legs, color, or statistical sophistication). Rather, judgments in the three examples are weighted sums, with the weights adjusted for the criteria that apply especially to an individual case, and the priorities of the rater. Personnel judgments, too, depend on weighted sums of how well employees get along with others, process information, apply themselves to the job, produce the expected work, and so forth. Employees and manuscripts can have different profiles of excellence, because they make different contributions to the organization - be it business or research - and no simple list of criteria will capture that weighted set of judgments.

Can it be otherwise? One can specify some flaws that disqualify manuscripts, wines, scents, and persons from approval, flaws so glaring that any trained reviewer would recognize the problems. Thus arises the greater agreement on rejection than acceptance. It is far more difficult to specify the combination of virtues that qualifies a manuscript for acceptance. If we attempted to have a prescribed set of ticked-off

criteria that were to be summed to an acceptance or reject decision, would we not publish the dullest, most conventional articles? Many say that this is what our journals do now, but I suppose it could get worse. Checklists, like other coercive criteria, serve the mediocre.

The qualifications of reviewers, by knowledge and experience in a field, are also called into question by the P & C's data on editorial decisions. How could they not have recognized previously published articles in their own fields? Any of us who write frequently know that one cannot also read all of the material in one's field, defined an inch beyond one's current research problem. This is not an excuse but a lament. It is also the case that many research reports are redundant with others in an area, another small parametric variation on a familiar theme, and hence indistinguishable for all but the authors and other investigators concerned with that particular problem. Reviewers are not selected because they are the leading experts in a particular problem. If editors followed the practice of selecting as reviewers those who were most closely identified with a research problem, the journals would be even more parochial than they are. Rather, most editors, I think, choose reviewers who have some perspective on a field from a modest distance, which also means that they may not recognize a particular article as a resubmitted one.

P & C raise the further question about whether reviewers ought to be anonymous (and irresponsible, which is another matter entirely). Like all editors, I am opposed to abusive, illogical, and unreasoned reviews. I have reprimanded reviewers for occasional ad hominem remarks, for being too picky about details, and for being less critical than they ought to have been. Editors learn, of course, which reviewers to trust. Identifying reviewers to authors might reduce the occasional lapse into abuse, but I think that the loss of anonymity brings with it more trouble than good. Young people in a field - often the best and most willing reviewers - cannot afford to criticize more senior colleagues to whose institutions they may need to go and to whom their grant proposals may be sent for review. I have received stinging and justified critiques of senior colleagues' research reports from junior investigators, whose careers would not be advanced by a loss of anonymity. There is an important protection in a reviewer's anonymity. Some senior people, including me, most often sign their reviews, however critical, but we have little to lose thereby.

My complaints about the editorial process are that there is too little adventure in the thinking and research I see, although I cannot specify exactly what my criteria for such judgments are. The variance of editorial recommendations increases as a function of the increase in unconventionality of the manuscript. This is where the editor plays a crucial role. Let us recall that reviewers' recommendations about publication are advisory, not binding, and that editors make the decisions about which reviewers to consult and which recommendations to accept. Although most of P & C's complaints are lodged against the reviewers of their resubmitted manuscripts, more attention ought to be given to the editors. Naturally, I am not eager to focus all of the heat on my colleagues and myself, but, as Harry Truman said, "If you can't stand the heat, get out of the kitchen." The editorial kitchen is where Truman's buck stops, and a strong editor makes decisions that are not merely the sum of reviewers' recommendations. There are no "editor-proof" journals, just as there are no "teacher-proof" curricula. Indeed, one would not want the monsters of mediocrity that would be produced by committee vote. The dark side of editors' power is the idiosyncratic nature of some editorial decisions; some roses smell sweeter to some editors than others, which puts democracy between, not within, journals. Pluralism among journals protects the publication system.

S. Scarr is an editor of Developmental Psychology *and a former associate editor of* American Psychologist.

Referee report on an earlier draft of Peters and Ceci's target article

William A. Scott

Department of Psychology, The Australian National University, Canberra, A.C.T. 2600, Australia

I do not detect any pseudonymy in this manuscript [*see editorial note below, Ed.*] - only an occasional superficial reference in a stunningly superficial study. One would think that a novel contribution to this problem would entail a sample of adequate size embodied in a reasonably informative design. As the authors recognize, they did not manipulate the imputed independent variable (institutional prestige), but set this as a parameter of their study. Given the known unreliability of manuscript appraisal, it is hardly surprising that there should be substantial regression toward the mean in a resampling from one extreme of the distribution. What is once again demonstrated is that humans (including journal referees) are all too fallible.

After discarding the three manuscripts that were detected as fraudulent and the one for which journal policy had changed, the authors are left with a sample of just 7 manuscripts. (The 18 appraisals of them are, of course, not independent.) [*In P & C's final draft these figures were increased to 8 and 22, respectively. Ed.*] It is hard, with a sample of this size, to reject statistically certain plausible rival hypotheses - for instance, that the manuscripts came from a population with a 60% probability of being recommended for acceptance.

Would an adequate sample size have made a difference? I don't think so, given the fatal defect in design. Would an adequate design have made the research acceptable for publication? Maybe, in a much shorter note. But it would only have made a limited point, of which editors are already aware, namely, that referees, like most other people, are susceptible to prestige suggestion. The problem remains, What to do about it? And I do not find the authors' suggestions particularly compelling. In my experience, the biggest problem for an editor was to find enough competent, conscientious referees to appraise the manuscripts submitted within time limits that authors can reasonably expect. Some means is required of extending the editor's horizons of search. Unfortunately, far too many of the initially appealing prospects turn out to be either peremptory and unhelpful to authors or extraordinarily naive.

W. A. Scott was one of the referees for American Psychologist (AP), *which rejected an earlier draft of P & C's paper. This commentary is the text of that referee report. Dr. Scott has requested that* BBS *quote from his response to* AP *editor C. A. Kiesler's decision letter: "I am glad to know the outcome [rejection], but appalled that some of your reviewers were so tolerant. It just goes to show that we are far from a shared culture when it comes to appraising either scientific importance or methodological adequacy." In a cover letter to* BBS, *Dr. Scott repeated his belief "that far too much attention has been given to this [P & C's] preliminary study" and asked that it be pointed out that he himself had declined to submit for publication the report of a follow-up study on lack of methodological agreement among referees for the* Journal of Personality and Social Psychology *(Scott 1977; see also Scott 1974) because of "the belief that no new information had been uncovered in my study." Ed.*

Responsibility in reviewing and research

Sol Tax[a] and Robert A. Rubinstein[b]

[a]*Department of Anthropology, University of Chicago, Chicago, Ill. 60637 and* [b]*School of Public Health, University of Illinois at the Medical Center, Chicago, Ill. 60680*

Certainly any bias that causes original papers, proposals for research, or funding requests to be evaluated on some basis

other than the merit of individual submissions must be considered pernicious. We agree that the scientific community ought to do all it can to eliminate such biases. Ideally, before we do that, we will have recognized precisely the nature of the biases involved so that we can set out strategies for dealing with them properly. Peters & Ceci (P & C) report data that they interpret as revealing a systematic bias on the part of reviewers for nonblind refereed journals toward giving papers by authors affiliated with "prestige" institutions more sympathetic readings than those by authors not so affiliated.

P & C's data do not seem to us to support their conclusion as strongly as they do an alternative the authors discount. That is, because P & C chose to use as their bogus affiliations names such as Northern Plains Center for Human Potential, which are so far from mainstream psychology institutions, it seems to us that they have demonstrated the action of a strong bias against materials originating outside "appropriate" institutions, rather than a bias in favor of submissions from "prestige" institutions. Unfortunately, the inference they wish to make cannot, in our view, be deemed strongly supported because the data do not include instances of the fate of resubmitted published papers originating at rather "ordinary" institutions.

This is not to say, of course, that we need not be concerned and alarmed by P & C's results. In fact, we think a bias that prevents competent work from entering the arena of community scrutiny is much more damaging to individuals, institutions, and the scientific community than a bias that lets mediocre work slip through. After all, once an article or research project is presented to the scientific community it can be discussed, and, where proper, even discredited. The flaws of a poor but published study will become known. A project that is discriminated against, however, never enters the domain of community discourse. If it is a bad project, it is never subjected to the criticism of the researchers' peers, and the researchers themselves do not benefit from the growth that accompanies being shown where one has gone wrong. Likewise, it is the scientific community's loss if the results of a well-done project are not made available for discussion because of a bias against its origin. This is especially damaging if the project presents new interpretations or data.

Further, even if the bias against nonmainstream institutions was ever well grounded, it is certainly no measure of the quality of researchers today. Because of the present employment situation in most academic disciplines, even people trained at "prestige" institutions are just as likely to be found at "ordinary" places (and perhaps also at some "weird" places) as those not so trained.

The larger problem is the widespread tendency to use as measures of quality the number of an individual's or institution's publications, citations, grants, and so on. Everyone in the scholarly community ought to hold (but many obviously do not) a principled objection to using such quantitative data alone for making such judgments. The merits and quality of a colleague's work should only be evaluated through careful and thorough study. This is part of being a responsible member of the scientific scholarly community; yet because it is time consuming and often difficult, it is a responsibility that is widely neglected. Because this creates the climate in which simple counts of publications, citations, grants, and the like, rather than careful evaluations of quality, become paramount (and pernicious), it is this basic situation that must be forcefully attacked.

P & C consider briefly several ways in which the journal review process might be improved. Their suggestions focus chiefly on ways of improving interreviewer reliability and ensuring greater author-reviewer accountability. Some of the options they discuss, such as maintaining lists of "journal shoppers" or catalogs of author evaluations of reviews, strike us as unwieldy, undesirable, or both.

Maintaining a list of "journal shoppers" would not only create a context in which authors would be interested in preserving their reputations by not being branded "publication losers" (as P & C suggest), but would also create another solely quantitative measure that could be used to discriminate against individuals. As we said before, such negative biases ought to be viewed as even more pernicious than the sorts of positive bias P & C attempt to demonstrate in their target article. Further, there are times when journal shopping may not be entirely inappropriate. Journals, after all, have different missions and may view the same material differently. Assuming too that authors revise their rejected papers in response to the critiques supplied by editors, resubmissions ought to be substantially improved, or, at least, different papers. So, while journal shopping may reflect poor professional judgment in some cases, it ought to play a role in the socialization of members of the scholarly community and in the honing of scholarly judgments.

While author evaluations of reviews would supply interesting information, it would surely be more time consuming and difficult to maintain such files than, we suspect, would be justified by their usefulness.

We argue that multiple blind reviewing of papers, followed by the kind of open peer commentary found in *Current Anthropology* and in the *Behavioral and Brain Sciences*, is one way of facilitating responsible consideration of the data and ideas presented in accepted papers. Even these procedures, however, do not by themselves ensure fairness and accountability in the refereeing process *prior* to acceptance. After all, authorship and other characteristics of a paper's origin can very often be inferred from internal cues in the paper under review.

Our view is that the best way to ensure fairness and responsibility in reviewing is for editors to foster preacceptance interaction between authors and reviewers. This would mean that journal editors would send out papers (blind or not) to reviewers. Reviewers' signed comments would be sent to authors. Authors and reviewers would then be free to correspond directly, with copies sent to the editor to make sure that the originator of new conceptual material was properly credited. Authors might choose to argue the adequacy of their submitted presentation, to revise in correspondence with the reviewers, or to withdraw their contribution. When authors persisted, editors would then need to judge at what point, if any, papers met standards of acceptability for their journal, and to exercise their editorial office in deciding on the basis of the developing correspondence what the disposition of the paper ought to be. This editor-involved exchange would also go a long way toward furthering the interests of the scientific scholarly community, which, after all, rest in ensuring the fair, wide-ranging discussion of ideas and data presented by our colleagues.

We turn now briefly to an ethical and methodological point raised by P & C's article. While the results of their study are interesting, they do not, we think, justify the deception used to obtain them. Although standards of acceptability for research methods vary within and between disciplines, as anthropologists we want to assert our preference for adhering strictly to standards of honesty and responsibility in the design and conduct of research, especially when dealing with people (including journal editors). Researchers also ought to keep in the forefront of their considerations the possible effects of their research procedures and the publication of their results on their informants. This means that researchers must inform those they work with about what they intend to do and why. If that results in direct objections from the people one wishes to work with, then the research probably ought not be done.

On the face of it, P & C's study harms no one. But a knowledgeable person could probably discover the journals included in their article. Hence, some editors may well find their own efforts impugned (fairly or not) as a result of the present study.

The general point is that when conducting research we ought to behave in an honest, open, and responsible manner. When our research interests conflict with our ability to follow such standards we should not collect data through deception. Such styles of data collection diminish the scientific scholarly community just as surely as does the exclusion of ideas from discussion.

ACKNOWLEDGMENT

Robert A. Rubinstein's work for this commentary was supported by National Institute of Mental Health grant #MH-15589-03.

S. Tax is the founder of Current Anthropology (CA), *the experiment in scientific communication on which the* BBS *project is modeled. It was the participatory democratic practices of certain American Indian nations that first suggested to Professor Tax the idea of the "CA Treatment," which we have attempted to emulate under the name of "open peer commentary."* BBS *gratefully acknowledges its lasting indebtedness to Professor Tax for providing our model and for offering much generous assistance in our formative years. The proposal he makes in this commentary concerning prepublication collaboration between authors and referees is very much in the* CA *spirit. Ed.*

Perhaps it was right to reject the resubmitted manuscripts

Garth J. Thomas

Center for Brain Research, University of Rochester, Rochester, N.Y. 14642

Having been on the receiving end of referees' critiques for some time and having also been on the interpretation end of referee critiques during my stint as editor of the *Journal of Comparative and Physiological Psychology* (JCPP), I offer a different interpretation of Peters & Ceci's (P & C's) results.

I conjecture that the referees and editors might have recognized P&C's resubmitted manuscripts as very like something they had seen before. However, they would probably not have had the necessary library facilities at hand to verify this because they would have been in their study at home or in their cabin in the woods (or whatever), and unlikely to have the time to devote to the problem (generally being very busy at their own benches). Thus they would not have raised the problem of plagiarism. Such an accusation is serious (as serious as accusing a colleague of fudging data). Both kinds of accusation would need precise documentation. Therefore the reviewers and editors in P&C's study probably fell back on statistical criticisms, as psychologists are wont to do, to justify a negative feeling about the manuscript. The result is that they made the right recommendations (reject), but they did not give compelling reasons. In fact, their "reasons" tended to sound pretty picky. There was one exception, Journal I (I think it was JCPP), whose wrong decision was to accept the manuscript, which reported results identical to data that had been published previously. The verbal reasons given by most editors and reviewers for their rejections may have been inadequate (uncompelling), but the right decision happens to have been made by most of the journals.

For example, the one case I encountered of any irregularity in a JCPP manuscript was left unmentioned in the referee's formal critique, but in a separate note to me, the referee said (in effect), "I am very suspicious of this paper. There might be some data-fudging in it for the following reasons.... It should be checked out." His reasons were persuasive, but the possible data fudging could not be checked out in a compelling way, one that would "stand up in court," so I simply rejected the paper with the vague statement that it did not unequivocally represent a substantial enough contribution (which was also true).

P&C suggest that their findings indicate a bias against unknown researchers from less prestigious institutions. I suggest that their findings merely reflect the reluctance of reviewers and editors to open a can of worms (plagiarism), which led them to give less than compelling reasons for their recommendations.

No one can argue that reviewers and editors should do a *poorer* job than they do now in selecting papers for publication, but I think we should keep in mind that there are two distinct roles for evaluative criticism. One is administrative (an editorial function): Should a paper be accepted or rejected? The other role of critical comments concerns the advancement of science, and is exemplified by BBS's open peer commentary. Obviously, the reasons given in support of administrative decisions should be persuasive but their being so depends on the skill and time that those involved in the editorial process can devote to the task - and this is usually not much. As I have said before, editors and reviewers tend to be individuals with some reputation as productive researchers. They tend to be busy at their own benches and reluctant to devote very much time to working at those of others. It is an unwise policy to expect editors to "educate" would-be authors.

P&C suggest that blind reviewing would help solve the problem of reviewer bias. Such evidence as I know indicates that such procedures do not make much difference in terms of actual administrative decisions (publish or reject). However, blind reviewing may be a desirable public relations maneuver as a reaction to the apparently increasing belief among scientists that the editorial process is shot through with bias and injustice. I would suppose that mistrust of the editorial process will continue to grow as adversary confrontations tend to grow in our society. And, of course, editorial mistakes are made. From the fact that there are always more senders (would-be authors wanting to publish) than there are receivers (readers who want to go through all that stuff), journals are bound to be conservative, and there is bound to be an adversary relation between aspiring authors and editors.

Anything really novel is likely to be given a hard time in the publication process. If research findings persuade one that the emperor has no clothes (and all other "competent observers" think he is clothed) one needs patience and many persuasive data in order to publish! It is difficult to discriminate truly novel advances in science from inadequate research. If the editor is too accepting, the journal gets to be a vanity press, and it fills with trash. Readership and subscriptions fall off. On the other hand, if the editor is too conservative and cautious, the journal might reject the first paper describing taste-aversion learning (for instance; see Revusky 1977).

P&C also recommend that all referee reports be signed. That is certainly necessary for the second function of commentary and criticism mentioned above (as in BBS), but I think that in the prior administrative function it would merely substitute one set of potential biases for another. Think of young scientists criticizing the paper of an older and well-established scientist: Might not they be overly circumspect if they had to sign their critique? This is a "judgment call," because we lack hard data, but I suspect one gets less bias with anonymous reviews than with signed ones.

My own suggestion is modest and would take a long time to have any effect. I suggest that all departments granting Ph.D.s require students to take a seminar in critique writing. Perhaps when the students get out into the world and are asked to review a paper for a journal, they will tend to give more cogent and compelling reasons for their administrative recommendation of acceptance or rejection.

G. J. Thomas was formerly editor of the Journal of Comparative and Physiological Psychology. *Ed.*

Some procedural obscurities in Peters and Ceci's peer-review study

Murray J. White
Department of Psychology, Victoria University of Wellington, Wellington, New Zealand

Are the authors of this target article authentic, and was the study actually conducted, or is there some insidious conspiracy afoot even here? No matter. What surprises me most is not what Peters & Ceci's (P & C's) paper reports, but rather, how it came to be accepted in its present form. The discussion about the wherefores of the study leaves a lot to be desired.

1. If the results for one article, submitted to Journal M, were excluded prima facie because of a shift in that journal's publication criteria, why were two other articles (submitted to Journals C and F) not excluded? (See Table 3 and note 4.) From what I can find, the editors of journals C and F rejected the resubmitted papers as *unsuitable* for their journals, that is their publication criteria had changed.

2. Just how "prestigious" and "impactful" were the journals surveyed? According to P & C, "10 of our journals were ranked in the top 20 for impact out of a list of 77 source journals of psychology." These data were taken from Garfield (1979a) who, in turn, makes it quite clear that they were for the year 1969 (a point not mentioned by P & C). A look at Garfield's data shows that of these top 20 journals, only five had 1979 impact factors *under* 1.15; the mean and median 1979 impact factors for the 20 journals were 2.0 and 1.5 respectively. Furthermore, the 1979 *Journal Citation Reports* (Garfield 1979b) cited by P & C shows 26 psychology journals with impact factors greater than 1.15. Mathematically, the discussion about selection of journals in "Method" just does not make sense, and I am forced to conclude that some of the journals tested were not at all top-flight.

3. We are told that the *mean* citation frequency for the 12 articles was 1.5 in each of the two years following publication. This figure is not that impressive given the impact factors of 8.5 for *Psychological Review*, 5.6 for *Cognitive Psychology*, 3.2 for *Psychological Bulletin*, and so on. Anyway, perhaps three of the 12 articles had mean citation rates of six, and nine had rates of zero, giving an overall mean of 1.5. It is impossible to tell from the data reported, and it is therefore reasonable to advance the hypothesis that it was the three highly cited articles that were detected. The nine nondetections may accordingly have been borderline cases in the first place, with few people having bothered to cite them since. A fractional shift in publication criteria would be sufficient to doom these nine articles (seven out of the original 13, if those submitted to Journals C and F are excluded) a second time around. In this respect, it should also be noted that Journals D and E had substantially increased their rejection rate (see Table 3). I would like to know what the story was about those articles (if any) that were submitted to *Psychological Review, Cognitive Psychology, Journal of Experimental Psychology, Psychological Bulletin*, and the *Journal of Verbal Learning and Verbal Behaviour* (for example).

4. P & C have obviously not read some of the references they cite. The paper by White and White (1977) referred to in "Procedure" had nothing to do with "institutions."

All in all, P & C's findings have to be treated with considerable scepticism.

The quandary of manuscript reviewing

Grover J. Whitehurst
Department of Psychology, State University of New York at Stony Brook, Stony Brook, N. Y. 11794

Humans are likely to differ in their judgments of the meaning and value of complex events. Two classic tasks of psychology have been to understand the reasons for this diversity and, in many contexts, to attempt to reduce it. We have learned much, for example, about the factors that influence judgments about the suitability of applicants for jobs and training, and methods have been developed to increase the reliability of the selection process (Anastasi 1958).

Given the complexity represented by journal manuscripts, we should not be surprised that reviewers will differ in their judgments of adequacy. Given the dearth of research on the review process, we should not be surprised at the lack of proven methods for reducing the variability in these judgments.

Peters & Ceci. Peters & Ceci (P & C) have provided a valuable service in emphasizing with dramatic data the issue of reliability in peer review. I have only two reservations about their study, neither of which leads me to doubt their basic conclusions.

It is unfortunate that P & C confounded the basic issue of reviewer reliability with the more specific question of bias based on institutional affiliation. Ideally, one would have preferred to see institutional affiliation varied orthogonally. Failing that, the more appropriate first study would have involved resubmission of manuscripts with prestigious institutional affiliations. As it is, we do not have the appropriate control group to assess the contention that systematic bias based on authors' status and institutional affiliation accounts for the high level of rejection of the resubmitted manuscripts.

P & C rely heavily on the bias argument to account for what seems on its surface to be an improbable result: The reviewers of all eight manuscripts that received two or more independent reviews were in complete agreement on accept versus reject judgments; seven of these agreements were in the reject category, one in the accept category. It is striking that an article that has as its essential point the unreliability of peer review has demonstrated unheard of levels of reliability. A kappa statistic (Cohen 1968) or an intraclass correlation coefficient for these data would be +1.00.

Scarr and Weber (1978; also see Cicchetti 1980) have reported an R_1 of .54 and a weighted kappa of .52 for the reliability of reviews of manuscripts submitted to the *American Psychologist*. These are among the highest reliabilities on record, with intraclass correlation coefficients in the .20s being the rule in other reports.

Take the Scarr and Weber data as most favorable to a test of the probability of P & C's findings: 57 of 87 manuscripts considered by Scarr and Weber were in complete agreement, for a nonchance corrected agreement level of .655. If this is considered the upper range of the probability that two independent observers will agree on their categorical judgments of the publishability of a manuscript, then the chance of obtaining exact agreement on eight independent manuscripts is $.655^8 = .03$. P & C argue that the high reliability of the reviewers in their sample is made more plausible by their frequent use of the reject category; Cicchetti (1980) is cited as indicating that the reject category is the most reliable one among raters. Actually Cicchetti demonstrated that the "reject but encourage resubmission" category and the "accept as is" category are more reliable than "reject." But even if one focuses on the reject category, P & C's findings are improbable. Scarr and Weber found 33 of 54 uses of the reject category to be in agreement, for a nonchance corrected agreement level of .611. The chance of seven manuscripts involving at least one reject decision being placed in the reject category by independent pairs of reviewers is $.611^7 = .03$. Given that P & C's results are highly improbable using the most generous previous estimates of reviewer reliability, either their results are due to sampling error, or the biasing effects of author status and institutional affiliation are much more potent than previous studies have demonstrated.

Additional reliability data. Cicchetti (1980) reports that only

two studies prior to his own analysis of the data of Scarr and Weber had "(a) investigated the review process for psychological journals and (b) applied statistics that measure agreement rather than mere association" (p. 300). Clearly, additional data are needed. Data on attempts to improve the review process would be particularly interesting.

The *Merrill-Palmer Quarterly of Behavior and Development* is a journal in developmental psychology dealing with a full range of empirical and theoretical manuscripts. From its inception in 1954 through 1980, it was published by the Merrill-Palmer Institute. It has ranked high in measures based on citation frequency. *Developmental Review* is a new journal in developmental psychology that focuses on theoretical and conceptual issues. It is published by Academic Press; the first issue appeared in 1981. All manuscript reviews for the period September 1979 to August 1980 for the *Merrill-Palmer Quarterly* and September 1980 to August 1981 for *Developmental Review* were surveyed. The editor and the reviewing forms and procedures were identical for the two journals during these adjacent periods. There was a 62% overlap between the two editorial boards. Both journals used an optional blind review system that was instituted only at the request of an author; blind review was requested very infrequently. Data are reported for all manuscripts that (a) received two independent reviews, and (b) were rated by both reviewers on a 4-point summary scale (accept as is, accept with revisions, reject but encourage resubmission, reject) and on 10 7-point scales concerned with specific evaluative criteria.

The intraclass correlation coefficient (Bartko 1966) for 73 manuscripts from both journals was .29. This figure is higher than the .21 reported by Hendrick (1976) for *Personality and Social Psychology Bulletin* and the .26 reported by Scott (1974) for the *Journal of Personality and Social Psychology*, but much lower than the .54 reported by Cicchetti (1980) for the *American Psychologist*.

P & C list among their suggestions for improving the journal review system the use of a standard rating form with explicit criteria. Reviewers for *Developmental Review* and the *Merrill-Palmer Quarterly* used 7-point Likert scales to rate manuscripts on 10 dimensions. The dimensions, with their associated intraclass correlation coefficients, are as follows: overall quality (.25), impact as measured by citations or controversy (.20), conceptualization of problem (.09), importance of topic (−.01), originality of treatment (.29), quality of writing (.15), theoreticality (.31), reliability of results (.20), validity of conlusions (.22), breadth of interested audience (.01).

None of these scales is significantly more reliable than the 4-point summary judgment. Most are not as reliable. Some are completely unreliable. This particular attempt to improve the journal review system does not seem promising, though it is arguable that different scales might prove more reliable.

P & C (after Hall 1979) suggest that a procedure of having authors rate reviewers on dimensions of fairness, carefulness, and constructiveness might improve the reliability of reviews by making reviewers more accountable. A rating form with these dimensions was mailed to all authors submitting manuscripts to the *Merrill-Palmer Quarterly* for a nine-month period in 1980. Less than 20% of the authors completed and returned the forms. That compounded the already serious problem of obtaining sufficient numbers of reviews on an individual reviewer to ensure fairness and anonymity, and the experiment was abandoned.

The role of an editor. Neither P & C's article nor the other literature on manuscript reviewing speaks to the role of the editor in influencing the review process. Some editors take a more active role than others, but given relatively low levels of reviewer reliability, even the most passive editor is often faced with split decisions. For instance, in the data from *Developmental Review* and the *Merrill-Palmer Quarterly*, 77% of the manuscripts received at least one review in the "accept as is" to "reject but encourage resubmission" categories. This was true of 47% of the *American Psychologist* manuscripts (Cicchetti 1980). Obviously editors have to pick and choose among many manuscripts that have at least one somewhat favorable review. A number of uninvestigated issues surround this process. (1) Is the editor reliable? If editors deal with mixed reviews in a highly reliable way, then lower levels of reviewer reliability may not be a cause for so much concern. After all, reviewers not only make summary judgments, they also write reviews. It is the editor's chore to evaluate their comments. I expect, but have no relevant evidence, that editors exercise considerable and reliable influence in this process. (2) What is the function of author intervention with the editor? As an editor I have accepted initially rejected manuscripts because of persuasive feedback from an author. As an author I have changed editorial decisions on my own work. Does such input increase the fairness of the editorial process or does it only reward assertiveness? (3) Does the editor bias the outcome of the review process in the choice of reviewers? This can range from what I expect is the virtually unanimous practice of assigning the most promising manuscripts to the strongest reviewers, to the more pernicious practice of assigning friends' work to reviewers who are likely to be lenient.

In making judgments about complex events, the person in control often feels most certain of reliability and validity. This is no less true of journal editors than of job interviewers. The intraclass correlation coefficients for *Developmental Review* and the *Merrill-Palmer Quarterly* are not nearly as high as one would like, yet as editor of those journals, I seldom had the feeling that authors were being subjected to a capricious process. Yes, reviewers often disagreed somewhat on summary recommendations even though their substantive comments might have had the same flavor; less frequently they disagreed diametrically, and occasionally I disagreed with unanimous reviewers. But in each case, I felt my decision was reliable and fair. I expect I am not so different from other editors in that respect. That is the quandary of manuscript reviewing. What is probably an unreliable process does not seem so to the people who control it. More research like P & C's, but directed at the editor's role, is most likely to change what may be, but does not feel like, a flawed endeavor.

G. J. Whitehurst is editor of Developmental Review *and former editor of the* Merrill-Palmer Quarterly.

Research on peer-review practices: Problems of interpretation, application, and propriety

William A. Wilson, Jr.
Department of Psychology, University of Connecticut, Storrs, Conn. 06268

Peters & Ceci (P & C) have presented persuasive evidence for the existence of response bias in the review procedures of psychological journals. The results are very interesting, but I don't want to encourage the continuation of research of this kind. An explanation of that negative opinion follows some questions and comments.

1. Is the bias related to the investigator and his reputation or to the prestige of his institution? Perhaps favoritism shown to an individual who has previously contributed significantly to a given field will enhance the overall accuracy of an editorial decision, for such a person is less likely than others to have made the kinds of research errors that are not detectable in the manuscripts of either experienced or inexperienced researchers. I would hazard the guess that positive bias in the

journal review process *is*, in fact, largely related to investigators and not institutions.

2. Is the relationship between prestige and the amount of favoritism monotonically positive? Perhaps not. Many editors give extra consideration to a manuscript received from a "developing nation" or from someone identified as a member of a "disadvantaged group." Special consideration is certainly shown in the time and effort given to improving an acceptable paper; special consideration may in fact sometimes lead to the acceptance of a manuscript that would otherwise be judged marginally unacceptable. Should practices of this sort be condemned or should they be encouraged? Of course, they would be less likely with blind refereeing.

3. Minor comments about the research report:

a. To demonstrate the quality and prominence of the articles selected, the authors point out that the articles received 1.5 citations per year in the two years after publication. Especially considering publication lags, it seems likely that many of these were self-citations, which would *not* be convincing evidence.

b. The editors represented in Table 1 apparently included both an editor and an associate editor for five of the journals. Although practices differ across journals, it is likely that in some cases the editor in chief saw very little more than the title of the newly submitted manuscript.

c. Obviously we must still be concerned about those editors and reviewers who did not realize that the information they were reading was already available in the literature. Perhaps these findings mean that we need even larger stables of reviewers, because investigators know the latest research only in their most narrow areas of specialization. Perhaps it means that editors are shortsighted in appointing prominent research scientists to their consulting staffs; many of the most active researchers go to the literature only to add to their reference lists a few other names to accompany their own. Perhaps editors should deliberately seek out people who might be more conversant with the literature, such as writers of textbooks or *Psychological Bulletin* articles, and the like.

4. Improving the review process:

a. I agree with the implied criticism of editors' cover letters that emphasize the expertise and wisdom of their consultants. This is often the only resort of an editor who has received reviews that do not themselves provide evidence of the consultants' knowledge and judgment. All this is understandable when a manuscript is truly unredeemable, and neither consultants nor editors wish to spend the time to explain why – understandable, but not excusable, because editors and reviewers should realize that an educational role is part of their duties.

b. The general suggestion that referees receive more information from editors, including training materials that might increase the reliability and validity of reviews, should certainly be endorsed. A few people who feel very important might consider this insulting and resign, but the total effect would be positive.

c. The specific suggestion that authors review referees reminds me uncomfortably of student evaluations of teachers. Problems concerning the validity of referees' comments would be as nothing compared to the difficulty in assessing authors' reactions (but see the next point).

e. Most editors have tried to make a difficult job somewhat easier by ruling that "the decision of the judges is final." Couldn't this policy be relaxed somewhat, so that rebuttal and a call for reconsideration would come from authors more as a function of how much they were wronged than as a function of the intransigence of their personality?

f. Perhaps there is relevant opposing research, but I believe that a system that called for *signed* reviews would be more costly than beneficial. Some referees now insist on signing their reviews and yet offer frank critical comments, but I don't think most human beings (even psychologists) would do so in an unbiased manner. One practical outcome might be an increase in the number of consultants who would decline to submit a review, being unwilling to submit either a dishonest positive recommendation or an unpopular negative recommendation about the paper of a friend or prestigious colleague.

5. My major reaction to the research reported by P & C is concerned not with its findings, interpretation, or application, however, but with its propriety; the deception involved in conducting the research casts obvious doubt upon its ethical nature. When an author submits a manuscript to a journal, he knows that he will receive reviewers' comments on it. He hopes that they will be favorable; he certainly hopes that they will be fair. At the very least he knows that they will be genuine, and he could properly complain if some of the comments were faked instead and he found himself involved in an experiment without his consent. I hope that editors and publishers have not tolerated such practices, and will not do so in the future.

The primary problem with the present study is that editors and reviewers were experimented upon without their informed consent. The time spent on the bogus papers was practically stolen from them. The research led to a conclusion whose most obvious interpretation is unfavorable to the ability or integrity of the "subjects." Clearly this violates ethical standards to some extent. The authors of this report have been quite responsive to comments about their research design and interpretation; I hope they will now respond more fully concerning the ethical problems raised.

Experimenter and reviewer bias

Joseph C. Witt[a] and Michael J. Hannafin[b]

[a] *Department of Psychology, Colorado State University, Fort Collins, Colo. 80523; and* [b] *Division of Psychoeducational Studies, University of Colorado, Boulder, Colo. 80302*

In this very stimulating article Peters & Ceci (P & C) address a most important question, but travel too far on limited data. The trap they set for 12 journals was a clever one but does little to improve our understanding of peer-review practices employed by modern research journals. Unfortunately the most poignant and telling aspect of this type of study may be the display of the extent to which research ethics and experimental methodology are often compromised in order to produce evidence to support a hypothesis.

Like most authors who have experienced the agony of a rejected manuscript, we found it easy to agree with the theme of the article. Article acceptance seems to be mediated by arbitrary forces. It was with great expectations, but unfortunately with different results, that we reviewed the same literature as P & C. We found their literature review somewhat selective not only in terms of the articles they chose to discuss, but also in terms of what data contained in the articles they chose to consider. For example, P & C note that interrater agreement between reviewers of manuscripts "is typically reported as low to moderate, with intraclass correlation coefficients of 0.55 at best." They go on to cite various studies that presumably document this statement (Scarr & Weber 1978; Watkins 1979). However, inspection of the data from Scarr and Weber reveals that, on a 5-point scale, there was perfect agreement for 57 of 89 manuscripts, and raters only differed by one category on an additional 12 manuscripts. The fact that 79% of the reviewers were substantially in agreement led Scarr and Weinberg (1978, p. 935) to conclude that "inter-rater reliability is quite high and gives us new faith in ourselves as casual observers." Obviously Scarr and Weber arrived at

conclusions from their data markedly different from those of P & C. In another instance, Watkins (1979) is cited as finding "a near total lack of interrater agreement for reviews of manuscripts submitted to the *Personality and Social Psychology Bulletin.*" P & C fail to mention that in the same article Watkins reported a kappa statistic of .49 for interreviewer agreement for articles submitted to another well-known psychology journal, prompting him to write that "reviewers agreed at a level substantially greater than would have been expected by chance alone." This game of reporting statements taken out of context could go on and on, but such instances of bias call into question the degree to which authors such as P & C remain objective, open minded, and unbiased in the pursuit of "truth" about review bias.

Aside from possible bias in their literature review, P & C's methodology contained sources of potential experimenter bias. To disguise the articles, P & C made "minimal and purely cosmetic" alterations of the originals. This may represent a source of bias, because professional journals are as concerned with the technical adequacy of the writing as with the content. By altering the word and sentence order, P & C may have made the flow of the text appear choppy, or simply disorganized to reviewers. To say that the content was unaffected by alteration of word order is to discount the relationship between information and expression. It is likely that changing some of the "nontechnical" words of even a fictional story would sabotage the meaning and flow of the text. For example, changing *The Old Man and the Sea* to *The Old Guy and the Water* or *Gone with the Wind* to *Off with the Breeze* might have resulted in different reviews for those works. [*Cf. Ross, this* Commentary. *Ed.*] Also, the displayed data that P & C converted into tables and vice versa may have ended up in an inappropriate format. For example, most data in the *Journal of Applied Behavior Analysis* represent the rate of response over time. These data are published exclusively in the form of figures, and it would be clearly inappropriate to convert to a table format. Because publishing in quality research journals is very competitive, such "minimal" changes may call into question the author's competence and affect markedly the probability that the article will be published.

A further concern pertains to the scope of P & C's inference and speculation, given the nature of their study. For example, they imply that a primary source of variance in their study is institutional affiliation. This hypothesis could have been tested directly by simply including recognized productive and prestigious institutions as a within-study control measure as well as the fictitious institutional affiliations. Since several presumably cosmetic changes were made in the manuscripts themselves, the use of previous acceptance data no longer constituted a truly valid baseline. In a methodological sense, too much was compromised to permit such sweeping inferences.

A final concern centers on the possible ethical questions surrounding P & C's methodology. They did not indicate whether or not editorial cooperation was solicited prior to the study. If the issue is as important as the authors contend, surely editors would cooperate; if not, that alone would be information worth reporting. If editorial support could have been elicited, P & C's methods could have been more direct and their inferences less speculative. For example, the editor could have sent a rating form that included key review dimensions affecting recommendations for rejection or acceptance. In this manner hypotheses regarding why papers were rejected could have been tested in a more systematic way. The deception precluded any systematic analysis of the reasons for rejection except as broadly inferred.

We hope that these comments will not be perceived as petty or picayune - because the subject matter is far too important. Instead, we offer them as constructive criticism, to stimulate additional research and discussion in this crucial area.

Competency testing for reviewers and editors

Rosalyn S. Yalow
Veterans Administration Medical Center, Bronx, N. Y. 10468; Montefiore Hospital and Medical Center, Bronx, N. Y. 10467

There are many problems with the peer-review system. Perhaps the most significant is that the truly imaginative are not being judged by their peers. They have none! However, this is not the issue addressed in the article by Peters & Ceci (P & C). Their contention is that papers prepared by fictitious authors working in fictitious institutions that were virtually identical with previously published articles by known authors from major institutions were rejected because of reviewer bias against an unknown author. What was most disturbing about the P & C paper was the finding of the failure of reviewers and editors from respected psychology journals to have recognized the resubmission of previously accepted papers. Does it mean that they do not read or remember papers from their own journals? It certainly reflects on their competence.

However, I am in full sympathy with rejecting papers from unknown authors working in unknown institutions. How does one know that the data are not fabricated? Those of us who publish establish some kind of track record. If our papers stand the test of time and are shown to be valid through confirmation by other investigators, it can be expected that we have acquired expertise in scientific methodology. Admittedly this is not always so. Some investigators with established reputations have subsequently been shown to have fabricated data, but these cases are rare. Even established investigators do make mistakes and on occasion write erroneous papers. Nonetheless, on the average, 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.

As a reviewer, I read not only the paper to be reviewed, but also previous papers by the same authors, and I also attempt to determine whether their work is cited by others in the field, and the like. Thus, it is most unlikely that I would have failed to notice that the resubmitted papers were fraudulent, or that they came from nonexistent authors in nonexistent institutions. I think what has been demonstrated by this study is not reviewer bias, but rather reviewer and editorial incompetence. I certainly agree that better training of the referees for those psychology journals would be in order.

It is the editor's responsibility to determine whether the reviewer's reports or the rebuttals by the authors have more merit. In the cases cited the problem was not in the review process per se, but in the lack of editorial competence. On occasion I have written to remind editors that their office is not simply a letter drop, that theirs is the responsibility to act as judges between the reviewer's report and the author's response. In fact, in my Nobel lecture (Yalow 1978), I published the initial letter of rejection by the *Journal of Clinical Investigation* of work that was to prove to be of fundamental importance to the development of radioimmunoassay. Eventually we reached a compromise with the editor, and the paper was published. I have since had the opportunity of writing to other editors who rejected our papers saying, "You may not become as famous as [the editor] in being identified in a Nobel lecture, but you are on the right track." Nonetheless, I must admit that we have never failed to publish our worthwhile papers eventually, and there have been times that I have been very appreciative of reviews that I initially resented. The idea of an author review of journal reviewers is an excellent one. It would encourage less aggressive authors to protest inadequate or inappropriate reviewing.

The problems associated with the peer-review system of research funding differ from those involved in refereeing papers for publication. I believe a prospective review system

for research funding should be replaced with a retrospective system for established investigators and a less restrictive mechanism for getting young investigators started. Refereeing papers is retrospective - one examines what has been accomplished. Even when it comes from lesser institutions, good work is usually recognized, although there may be some delay. It is not without interest that the Veterans Administration Hospital in the Bronx, during a time when it had no close medical school affiliation, produced three investigators who were elected to the National Academy of Sciences and several who have received other major distinctions. Given the resources to initiate their careers, talented investigators will surface. The P & C article is really not relevant to the broader issue of peer review for research funding.

The data presented in this article are not sufficient to support P & C's hypothesis that reviewer bias per se was the primary factor in the rejection of these papers. It might simply have been that the reviewers and editors were unfamiliar with the purported investigators or their institutions and chose to use other excuses for rejecting the papers. This article would have been of more interest to me if the authors had confronted the editors with the hoax and attempted to elicit the real reasons for rejection.

R. S. Yalow is a Nobel Laureate in physiology/medicine. Ed.

Reliability and validity of peer review

David Zeaman
Department of Psychology, University of Connecticut, Storrs, Conn. 06268

Peters & Ceci (P & C) interpret their results as evidence for weakness in the *reliability* of the manuscript review process. It may seem paradoxical to suggest that their results may also be interpreted as evidence for high *validity* of the review process, but a valid disposition of the 12 resubmitted manuscripts should have been rejection on the grounds of prior publication, and 11 of the 12 were rejected. A validity batting average of .92 is not bad.

A counterargument is obvious. Only three of the 11 manuscripts were rejected on these grounds. However, the lack of currency may have tilted negatively the judgments of the other reviewers even though they did not say so. P & C did not present a thorough and objective account of the reasons reviewers chose for rejection so there is no way to evaluate this interpretation. It would have helped if reviewers had been asked to rate the currency or redundancy of the resubmitted manuscripts to see whether lack of currency contributed to the overwhelming correctness of the reviewers' judgments.

A general comment about the literature on the reliability of the peer-review process: Not a single study has attempted to measure the reliability of this process by taking repeated measures on randomly selected manuscripts where the repeated measures cover the entire editorial process (including judgments of a couple of reviewers and an editor, and the possibility of correspondence with the authors). This study is no exception. There are simply no ecologically valid data on the reliability of the whole review process as it currently works. It is probably too difficult to collect such data.

D. Zeaman is editor-elect of Psychological Bulletin. *Ed.*

Bias, incompetence, or bad management?

John Ziman
H. H. Wills Physics Laboratory, Bristol BS8 1TL England

The results of the experiment reported in Peters & Ceci's (P & C's) paper are so counterintuitive that they cannot be accepted as demonstrated fact until the research has been replicated, with appropriate variations, by other independent investigators. So much for scholarly caution and scepticism!

On the face of it, however, this experiment has been carefully conducted, with adequate precautions against obvious pitfalls of method or phenomenological interpretation. It presents evidence of a gravely pathological situation, calling for further serious inquiry and radical remedy.

A gross "anomaly" such as this is not merely difficult to explain, it is difficult to talk about at all. It puts at risk the whole conceptual framework within which we are accustomed to make observations and construct theories. 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.

How, then, can we think about these astonishing results, which suggest that the whole notion of an agreed standard of "publishability" is a chimera? The outcome of the laborious procedure of accepting papers for publication seems no better than random selection: The one paper that was accepted from the nine that were successfully resubmitted is well within the statistical expectations for a random process with an average acceptance rate of 20%. The consensus of reviewers *against* the resubmitted papers suggests something worse than total chaos - but that is bad enough to make nonsense of the whole system. The peer-review process seems not merely imperfect: It is an entirely useless, if not positively harmful activity, based upon quite erroneous assumptions.

I recoil from this conclusion, not because it is inconceivable but because it would take a very long time to imagine what to say or do next. Yet I am not convinced that institutional-prestige bias is strong enough to explain the whole effect, and that the pathology could be remedied in practice by some modest change of procedure, such as the use of blind reviews. The phenomenon reported by P & C is much more extreme than could be explained from the evidence of previous deliberate investigations on this particular point. If this were the whole story, then it is hard to imagine how any papers without authors from prestigious institutions could ever get published at all! To assess this factor, we must have control data from further experiments - for example, what happens to published papers by unprestigious authors when they are resubmitted under fictitious names of similar low standing?

What stands out is, to put it crudely, the gross incompetence of the expert reviewers at the jobs they were asked to do. They were so ignorant of their subjects that at least 75% of them did not even know that the very same work had been done before: Those who had accepted the papers originally had evidently overlooked serious methodological errors that were at once obvious to other reviewers the second time round. I must say, from personal experience as an author, as an editor, and as an adjudicator between authors and reviewers, I have never come across anything like such widespread incompetence or irresponsibility. That was in physics: Is the situation then so

entirely different in psychology? This question, also, needs to be tested empirically.

Or were reviewers being asked by "editors" to do a job that was beyond their real competence? Presumably the prestigious journals that involuntarily participated in this experiment are quite large, and mainly run by professional editorial staff handling hundreds of papers a year. Are they sufficiently well informed of the intricacies of the myriad specialties of their subdisciplines to choose properly qualified reviewers for the typescripts they receive? It is a familiar experience for recognized scientific authorities to be asked to review papers, comment on research proposals, or examine doctoral dissertations, that lie a little outside their own little cabbage patch - the few dozen scientists and the few hundred papers with which they are really well acquainted. Knowing who knows about what is itself a personal resource that comes from active participation in a research field; it cannot be transferred to a card index or computer file.

Paradoxically, it may be that the implicit standard of "publishability" is too low, and too difficult to assess. The resubmitted papers are said to be "above average" in quality, although this statistic has little significance in the very skewed distribution of merit found in all scientific literature. In principle they ought to be far superior to most of the 80% of papers that were rejected by these particular journals. And yet they contained very obvious deficiencies, which were probably apparent to the original reviewers who accepted them. In other words (as everyone knows!), a large portion of the scientific papers that are getting published are not very sound, and are not even very convincing to scientists who understand well enough what they are about. It might be easier to establish an intersubjectively agreed-upon standard of scientific credibility at a somewhat higher level, where such superficial sloppiness was known to be intolerable to editors and reviewers - and hence would already have been eliminated from submitted typescripts. The results reported by P & C could thus justify a radical reform in this direction (with all that this would imply for academic disciplines and professions) just as well as a contrary campaign to abolish the "useless" practice of peer review altogether. All that we can be sure of is that present practices are deeply flawed: For this blow to complacency P & C are to be gratefully thanked.

J. Ziman is editor of Science Progress. *Ed.*

Commentaries submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Ed.

Authors' Response

Peer-review research: Objections and obligations

Douglas P. Peters[a] and Stephen J. Ceci[b]

[a] *Program in Social Ecology, University of California, Irvine, Calif. 92717 (on leave from Department of Psychology, University of North Dakota, Grand Forks, N.D. 58202) and* [b] *Department of Human Development and Family Studies, Cornell University, Ithaca, N.Y. 14853*

Peer review is an important and complex topic which is best treated in the open peer commentary format of BBS. In our opinion, the wide range of ideas that our commentators have shared with us makes this a genuinely archetypal BBS treatment. We are delighted that our paper could stimulate this process. We are also encouraged by the very high level of participation of our commentators - out of the approximately 130 individuals invited to contribute, about half responded, with several still to appear in a subsequent issue - and by the strong cross-disciplinary representation.

In preparing this response we tried to address all the substantive critical issues that were raised, as well as more general issues such as recommendations for reforming peer review. In studying the commentaries we were able to identify four rather global areas of discussion (methodology, data analysis, interpretation, reform). We address each of these in turn.

I. Methodological issues

a. Failure to test and control for a variety of possible factors. Fourteen of our commentators addressed the issue of control groups. They argued that since we resubmitted papers that were published by authors from high-status institutions only, it is not possible to rule out plausible alternative interpretations. We agree that it would have been very desirable to have included resubmissions of previously published papers of both high and low status, as well as some new papers that had not been previously submitted and some previously submitted but rejected papers. We would like to have been able to resubmit half of the high-status papers (published as well as rejected) under a low-status guise and half of the low-status papers under a high-status guise. The reader will recall that some of these were among our own suggestions (pp. 191-192). However, to test all of the special factors suggested by our commentators would have necessitated a grand design that would have been impossible to carry out without the cooperation of journal editors, which, unfortunately, was not forthcoming.[1] Such a design might look something like this (taking into account every control group that at least one commentator felt was needed):

Author status (high vs. low) × *author institution* (prestigious, nonprestigious but real, fictitious) × *reviewer status* (high vs. low) × *reviewer institution* (high vs. low) × *journal type* (blind vs. nonblind) × *journal status* (high vs. low) × *scientific discipline* (social science vs. physical science) × *paper's history* (published, rejected, new submission).

We might add, for those interested in affirmative action, that one could add race and sex to the grand design (see **Horrobin**'s commentary).

The fact that we were not able to collect such multifactorial data and controls, despite our interest in doing so, does not necessarily preclude interpreting certain aspects of our results, such as the bias and randomness hypotheses. It is often necessary when working outside the laboratory (in the methodologically hazardous real world) to settle for a quasi-experimental approach to a problem. Many social scientists would accept the argument that the quasi-experimental design we used is, in general, insufficient to permit *strong* tests of causal hypotheses because it fails to rule out alternative explanations unequivocally. However, as Cook and Campbell (1979, p. 96) point out,

> It should not be forgotten that experimental design is only one way to rule out alternative interpretations and that sometimes threats can be ruled out in nondesign ways. This is especially the case when particular threats seem implausible in light of accepted

theory or common sense, or when the threats are validly measured and it is shown in the statistical analysis that they are not operating.

These authors go on to caution readers against concluding that quasi-experimental designs of the type we used are invariably uninterpretable. Judging from the large number of commentators who accepted our interpretation, it appears that most would agree that quasi-experimental design *can* provide causal inferences when used along with convergent evidence and cogent reasoning and analysis.

On statistical, theoretical, and intuitive grounds we argued that two of the findings in our study (eight out of nine undetected manuscripts being rejected and all sets of reviewers being in perfect agreement) justified our suggested interpretations. We remain convinced of this justification and return to the details below.

b. Confounding factors. Seven commentaries addressed the problem of confounding factors in our study. The most frequently mentioned one had to do with our potential creation of a negative halo effect by using nonacademic, suspicious-sounding institutional affiliations (**Beyer, Cone, Crandall, DeBakey, Tax & Rubinstein**). According to these commentators, research reports emanating from unknown, "soft-sounding" institutes may be reacted to negatively. If this were true, our results would be better characterized as demonstrating bias against researchers at low-status institutions than bias in favor of those at high-status institutions. There may well be some validity to this interpretation; however, even if it were the whole story it would hardly imply that the peer-review process was any more objective or dispassionate.[2] To whatever extent factors other than the merit of one's ideas, design, methodology, or interpretation (e.g. institutional reputation) are used to judge the scientific quality of one's research reports, dissatisfaction and criticism would still be required and justified. While we do not know for certain which of the two forms of bias is more likely, neither is desirable.

Another frequently mentioned confounding factor concerned the modifications we made in the original papers (**Perlman, Witt & Hannafin**). We would certainly agree with these commentators that if those expository changes substantially altered the meaning or readability of the articles, that could account for our findings. We in fact took great care to make sure the modifications preserved meaning and readability; as we stated in the target article, they were purely cosmetic changes, and one would not reasonably expect them to be sources of variance. To these claims we can add that none of the editors who participated in the study ever gave any indications to the contrary. They were in the best position to judge whether the articles' publishability had been affected by the changes we made. Their stated reasons for rejecting these articles ran far deeper than any such minor alterations. In summary, we find no evidence – logical or empirical – for the salience of the cosmetic factor.

c. Ethics of deception. Seven of the commentaries explicitly criticized our methods on the grounds that they entailed unwarranted deception (**Beyer, Chubin, Fleiss, Honig, Lazarus, Tax & Rubinstein, Wilson**); four others expressed the opposite opinion, indicating that they found the deception justified under the circumstances (**Crandall, Howe, Mindick, Rosenthal**); and at least eleven others actively encouraged us to expand our study to other disciplines, and in so doing appeared to have no quarrel at all with the ethics of our methods (**Armstrong, Cone, Beaver, M. Gordon, Mahoney, Manwell & Baker, Moravcsik, Over, Perloff & Perloff, Whitehurst, Ziman**). Nor did any of the reviewers and editors of *Science* (N = 3) or *American Psychologist* (N = 6) express in their reports and correspondence any ethical objections concerning our treatment of editors and reviewers. (Prior to publication in BBS we had tried unsuccessfully to publish these results in both *Science* and *American Psychologist.*)

The ethics of deception will not be resolved by simple vote counts of the above sort, however, or even by argument. There will probably always be moral relativists, like ourselves, who believe that a code of ethics cannot be violated in the abstract but must be seen in context; who believe that anyone intending to practice deception must do a cost-benefit analysis. If the result of such an analysis is the belief that the deception results in no significant physical or psychological harm to the subject, that the knowledge to be gained is potentially of social or economic importance, and that there is no other feasible way to collect these data, then the deception is, or at least might be, warranted. Conversely, there will be moral absolutists or nonconsequentialists who will always view any deception, regardless of relative costs and benefits, as morally reprehensible (Holden 1979). We believe the nonconsequentialist's position to be untenable for us as researchers and citizens, but because we do not expect to be able to bridge this chasm in this response, we will concentrate instead on those who accept the relativist position but disagree with our cost-benefit analysis.

On the cost side one might argue, as have **Beyer** and **Fleiss**, that the study may have resulted in "abuse" and "distress" to the editors and reviewers, and that "it is hard to see what benefits accrue to society... that outweigh the distress" (Beyer). It is possible that editors were humiliated by the results of the study. However, it was never our intention to embarrass individuals, and we have not, nor do we ever intend to disclose the identities of the editors involved, or their journals (those editors who are known to have been participants in the study unilaterally disclosed this information themselves, without any form of prompting by us). If editors feel humiliated, they can speak for themselves. All the editors involved were given the opportunity by BBS to comment on our study with or without disclosing their role. Only three have chosen to do so thus far. What are we to infer from this? Out of 60 individuals commenting in their professional capacities as editors, reviewers, and practicing scientists, only seven criticized our study on ethical grounds. While this certainly indicates some cause for concern and reflection, it hardly represents a massive protest from those passing professional judgment on our work.

From an ethical standpoint, *intention* is important. It may be illuminating to excerpt a quotation from an earlier review our study received from *American Psychologist* (AP). The reviewer to whom the study had

been sent (there were five reviewers in all) had been on record as opposed to research involving deception:

> Because of my public commitment to nonpublication of evidence unethically obtained [self-citation deleted] and because I find deception in research difficult to justify, I requested permission from Dr. Kiesler to consult with Kennedy Institute ethicists [X] and [Y]. In [X's] opinion there is some question about viewing the editor/reviewers as subjects. [Y] does see them as unknowing subjects. He does not, however, believe that status should interfere with publication since (a) there was no other way to gather this information; (b) the extreme importance of the data compels their publication; and (c) the deception results more in a shock to the faulty system than in a trauma to any one individual.

As our colleague and ethicist Burton **Mindick** argues, deception is never an ethically attractive alternative. However, like the above two ethicists, he too concludes that the deception was warranted and the data are important. (If it comes down to an assessment of the data's importance, it should also be noted that the majority of commentators did not share **Beyer**'s low opinion of the study.)

A further comment on the issue of ethics concerns the argument that the data, granted their importance, could have been obtained without the use of deception. Three commentators expressed the belief that the study could have been accomplished by asking the editors to participate (**Fleiss, Witt & Hannafin**). We had given serious thought to this idea in the planning stage of our study, but we eventually decided that it would be unsound because editors and reviewers might behave differently if they knew their journals' performances were being scrutinized. We felt that the literature on observational reactivity and experimenter expectancy effects (Rosenthal & Rubin 1978) compelled us to use unobtrusive methods if we were to assess accurately such things as editors' or reviewers' awareness that a particular manuscript had already been published in their journal.

We are somewhat confused with respect to **Fleiss**'s endorsement of Weiss's (1980) remark: "I know of no research involving deception in which the results could not have been obtained without [its] use." What are we to make of the suggestion that editors themselves could adequately examine their own behavior? Would it not be ignoring nearly two decades of social-psychological research to expect individuals who are part of a system suspected of bias, ignorance of relevant literature, and poor reliability to be objective? Can they be objective? We feel informed and affirmed in this regard by the keen insights of some editors themselves, including two of our commentators. Both **Whitehurst** and **Mahoney** described the subtle, sometimes unconscious, ways in which an editor can control a manuscript's disposition by the assignment of certain reviewers. Many other commentators also implicated the editorial role in the problems associated with peer review (**Chubin, Glenn, Goodstein, Manwell & Baker, Over, Scarr, Wilson, Yalow, Ziman**). How then can we collect such data without unobtrusive (deceptive) methods?

We should state that in 12 out of 13 cases we did obtain permission from the original authors to disguise their manuscripts and use them as stimulus materials for a study of possible bias in manuscript reviewing. The sole exception arose because of a clerical oversight. Furthermore, we also obtained permission from all non-APA publishers to use these articles in our study. We explained that our intention was to change the names and institutional affiliations and resubmit the papers to editors.

We asked the publishers not to disclose the nature of our study to their editors. Finally, if consciences are so finely attuned to ethical subtleties that they are shocked by the deception involved in our research, how much more deeply must they then be disturbed by the bearing of our findings on the likelihood that submissions to professional journals will receive objective and unbiased review.

d. Status of journals and articles. Three commentators found fault with our use of citation indices to establish the quality of the journals or papers we selected. **DeBakey** argued that citations cannot be used uncritically to assess quality, and both **White** and **Wilson** felt that the mean citation figures were not that impressive. We agree with DeBakey about the need to qualify citation indices. Many complications need to be considered, such as the number of active researchers in an area, the number of published journal pages, the tone of the citations (positive or negative), the age of the journal (correlated with impact), and the like. For these and other reasons we did not rely on citations alone, but used multiple criteria in our selection. We were especially persuaded by author ratings (e.g. Koulack & Keselman 1975). That is, these were among the journals in which authors most wanted to publish and to which they looked for important findings. White's queries about how prestigious the journals were cannot be answered without disclosing the identities of certain journals. The journals were, and still are, highly regarded by researchers. Naturally, we used only journals that employed nonblind peer review, which eliminates certain highly cited journals from our study.

II. Data analysis

a. Small N. Eight commentators remarked on our small sample size (**DeBakey, M. Gordon, Griffith, Perlman, Presser, Rosenthal, Rubin, Scott**). Thirteen journals, 12 of which provided analyzable data, and 38 editors and reviewers do not make a large sample in absolute terms. It was, however, as large a sample as we were able to get. It was also large enough to permit a few univariate comparisons and to calculate a few probabilities. We hope that these raw data will someday be assimilated into meta (integrative) analyses and in so doing serve an even more useful function. As scientists operating in times of particularly scarce resources, it becomes increasingly urgent for us to share our data with those doing related work. Perhaps studies like ours that examine complex issues outside the laboratory should be viewed developmentally. We have provided interesting data that can supplement the data of subsequent researchers who, like ourselves, may not be able to provide all the control groups desired.

b. Preference for random-error over bias explanation. Eight commentators preferred to account for our results in terms of randomness (**Beyer, Colman, Glenn, Goodstein, Hogan, Manwell & Baker, Presser, Ziman**). These commentators argued that our findings could be explained by assuming that peer review was so (random-) error prone that anywhere from approximately 44% of all high-quality papers (Presser) to 80% of them (Colman) get rejected. These arguments were based on a single aspect of our findings - that eight out of nine undetected manuscripts were rejected. However, if we add to this outcome an additional one - that all reviewers were in perfect agreement - then a completely random-error explanation is unsatisfactory. We believe that there is ample theoretical and statistical ground to reject the complete randomness hypothesis, the one that posits no selection at all in peer review (Colman). According to our conditional probability analysis, the fact that 8 out of 9 previously published articles were rejected implies that publishable articles in these journals had a less than 43% probability of acceptance; otherwise, the probability of our observed outcome (on a regression/random-error hypothesis) would be less than 5% (4.6%). We agree with Colman that ours was not a highly implausible outcome if the system is completely random (actually it need not even be completely random to handle this outcome). We argued in our target article, however, that if one considers the perfect agreement between reviewers in our study, the random hypothesis is less tenable. We know that the level of chance agreement among reviewers (assuming an 80% rejection rate for each) is 68%:

$$(.2)(.2) + (.8)(.8) = .04 + .64 = .68$$

Sixty-eight percent of eight pairs of reviewers = 5.44, which is significantly less agreement than what we actually observed ($p = .045$). As **Cicchetti** and **Whitehurst** point out, this very high level of reviewer agreement is nearly unheard of in the empirical literature, with kappas of .55 (and unadjusted proportions of .655) being the highest reported. This is a significantly improbable level of agreement: $.55^8 = .003$; $.655^8 = .03$; however, it is actually congruent with a bias explanation, since the original (prestige) bias was presumably in the positive direction and the resubmitted manuscripts removed this source of bias. In other words, reviewers can agree with one another on "invalid" components such as institutional prestige. This is not unlike the distinction between trait and method reliability (Campbell & Fiske 1959); to the degree that the latter intrudes into judges' assessment of the former, the reliability of the former would be high, even though invalid.

It should be noted that the complete randomness hypothesis, though superficially similar to **Presser**'s randomness hypothesis, is really quite different. The Stinchcombe and Ofshe (1969) model that is the basis of Presser's argument predicts that high-quality manuscripts fare substantially better than chance (about four times better) while low-quality manuscripts (those defined by Stinchcombe & Ofshe as less than one standard deviation above average) fare substantially worse than chance. So, while this model is a nonconspiratorial one, it is not a completely random one. It assumes a .5 reliability coefficient, all covariation between measures being due to the variable of interest, validity of reviewer judgments of manuscript quality = .7, and a rejection rate of 84%. An important point is that it also assumes a perfectly normal distribution of manuscript quality. Some of these assumptions have been altered by subsequent investigators of peer review who used this model (e.g., Lindsey, 1978) to produce outcomes similar to those **Colman** advocates. While we find that basic model interesting, it is possible to demonstrate that it will not encompass our rejection data if we alter the assumptions even slightly to conform to our study (e.g., a mean 22% acceptance rate, and a slight departure from normality). But to argue along these lines is to miss an important point: The randomness and bias hypotheses need not be mutually exclusive. Our total data (including reliability), along with what is known from correlational studies like Gordon's (1980), have persuaded us to include at least some degree of bias in our explanation. We are not averse to the view that the peer-review process is unreliable (random error), but we think that any model of our data will have to involve a more complex explanation than that of randomness alone.

c. Chi square and independence. **Colman** and **Beyer** both take us to task for including editors in our chi square calculation, because the dependence of editors and reviewers violates the stochastic independence of the chi square model. However, even if one recalculates the chi square under various conservative assumptions, the result still supports our conclusion. We know from the most recent literature on peer review (Cicchetti & Eron 1979) that approximately 67% of all published articles in psychology journals are recommended for publication by both reviewers, that is, at most one third of published articles have split reviews. If we accept this figure as our estimate, we would assume that of the nine manuscripts we used, six had originally had unanimous positive reviews (N = 12) and three had had split reviews (three positive, three negative). Thus, originally, there were probably at least 15 positive reviews and only three negative ones. These numbers are obviously very different from the observed outcome of only two positive reviews and 16 negative ones.

III. Interpreting the results

a. Results unique to social science. Four commentators stated their belief that similar outcomes could not occur in their fields (**Adair, Beyer, Moravcsik, Rubin**), while five others felt that the problems were general and could be detected in most fields (**Beaver, Chubin, Glenn, Horrobin, Nelson**), and two were undecided (**Belshaw, Ziman**). It seems clear that reliability (or lack thereof) is a universal problem wherever qualitative judgments are made (e.g., see the recent reports of poor reliability of National Science Foundation grant-reviewing decision in physics, chemistry, and economics; Cole, Cole & Simon 1981). Of course, in the specific arena of journal reviewing, differences might exist between social and physical sciences. It is hard to judge how real these differences are (Moravscik) and how much they may be the results of superficial variations. For example, in physics the accep-

tance rates for the best journals are more than twice as high as those found in psychology and sociology (Adair, **Lazarus**).

b. Surprise at the results. Twenty-two colleagues who had no knowledge of our study were polled by **Armstrong** to find out whether they could predict the outcome. Over 60% of them were editors or associate editors. He found that most were genuinely surprised by the results, having predicted much higher detection rates and far lower rejection rates. Fourteen BBS commentators also expressed opinions on this matter. Concerning the poor detection rates, eight out of 14 commentators maintained that they were not surprised by the findings (**Adair, Beaver, Chubin, Glenn, Hogan, Lazarus, Manwell & Baker, Scarr**) with the remainder expressing surprise (**Cone, Goodstein, Palermo, Rubin, Yalow, Ziman**). Concerning the rejection of eight of the nine manuscripts, ten of the 12 commentators who addressed this finding were not surprised, the exceptions being **Perloff & Perloff** and Ziman. One possible reason why so many of our commentators professed to be unsurprised by our findings might be that, as a group, they are keenly interested in peer review, and they may have been familiar with a literature that is increasingly critical of it.

c. Other interpretations. Many interpretations other than randomness or bias were put forward. Four of the commentators felt that the papers we selected may have been marginal to begin with, or perhaps ordinary papers with faults (**Blissett, Nelson, Presser, Ziman**), and hence that the high rejection rate might be seen as a corrective. In line with this interpretation, four other commentators hypothesized differences between the initial set of reviewers and those used the second time (**DeBakey, Geen, Over, Perloff & Perloff**). Perhaps, it was argued, the second set of reviewers was younger or more critical than the original one. This makes sense if one agrees with Over that editors selectively assign reviewers on the basis of the author's status (see also **Mahoney, Whitehurst**). Nine commentators focused on the issue of "methodological flaws," offering explanations of why reviewers and editors frequently use this criticism to reject a paper even when it may not be their underlying reason (**Beaver, Crandall, Honig, Manwell & Baker, Palermo, Thomas, Wilson, Yalow, Zeaman**). If correct, this still does not help us answer the question of why the rejection rates were so high. Granted that there are social pressures on editors and reviewers to avoid going into tenuous subjective or theoretical rationales for recommending rejection of a manuscript, leading them to substitute the "universal criticism of choice," that is, methodological flaws. The question still remains: Why did the papers get published in the first place if these unmentionable problems existed? Why was the "universal criticism of choice" not invoked to preclude the earlier publication? The answers to these questions take us back to the original hypotheses, bias and randomness. (We do not find plausible the rather far-fetched hypothesis that the true reason for rejection was a [correct] subliminal sense on the part of the reviewer that the findings in each study had already appeared somewhere, sometime.)

d. Misrepresented findings in the literature. So far, we have dealt with internal considerations, such as design problems, alternative interpretations, and the like. In this section we deal with external considerations. Two commentaries (**White, Witt & Hannafin**) chided us in rather strident tones for supposedly selectively reviewing the literature and using references we did not read. It is difficult not to appear defensive in addressing these criticisms, though not to refute them solidly could cast a cloud of doubt over our scholarship.

Witt & Hannafin complained that our literature review was not just selective but misleading. They stated that we were unjustified in using Scarr and Weber's (1978) and Watkins's (1979) data to document our claim of "low to moderate . . . intraclass correlation coefficients of 0.55 at best." They point out that in the former study of reviews for the *American Psychologist*, 79% of reviewers agreed; and Watkins, though finding very low agreement, mentions a kappa of .49 for another journal. They go on to say, "This game of reporting statements taken out of context could go on and on, but such instances of bias call into question the degree to which authors such as P & C remain objective, open minded, and unbiased in the pursuit of 'truth' about review bias." To add to this charge, Witt & Hannafin suggest that since we were clearly biased in our behavior, perhaps we altered the manuscripts in ways that were not really "cosmetic," such as changing a title like *Gone with the Wind* to *Off with the Breeze*.

Even before reading the following rejoinder it should be apparent to the reader that our statement of low to moderate intraclass correlation coefficients, of .55 at best, is correct. First of all, Scarr and Weber did not present their findings in terms of kappa; that is, there was no adjustment for chance agreement. Both **Cicchetti** (1980; as well as his accompanying commentary) and **Whitehurst** report that the .55 value is indeed the highest observed: "Consistent with this finding was a study of peer-review practices for the *American Psychologist*, which reported the highest interreferee agreement levels to date (R(I) = .55)" (**Cicchetti**). "Scarr and Weber (1978; also see Cicchetti 1980) have reported an R_I of .54 and a weighted kappa of .52 for the reliability of reviews of manuscripts submitted to the *American Psychologist*. These are among the highest reliabilities on record, with intraclass correlation coefficients in the .20s being the rule in other reports" (Whitehurst).

What **Witt & Hannafin** mistake for selective reporting on our part is in reality accurate reporting. Their own use of the 79% figure is misleading; since only 65% of Scarr and Weber's reviewers were in *precise* agreement, the 79% figure reflected reindexing. But, most important, we expect a certain level of agreement because of chance, and when this is considered one finds R_I = .55, .54, .52, depending on indexing, for the highest level on record.

We hope we have made it clear that our review was not selective or biased and that we were not guilty of distorting anyone's data. Were this not true we would hardly have recommended many of the commentators we did, given their well-known opposite viewpoints. (We recommended Sandra **Scarr** as a commentor and it should be telling that she, unlike **Witt & Hannafin**, has not complained that we distorted her findings.) The issue

of impeding readability by distorting titles of articles has already been dealt with. Let us leave it to the editors who received the altered manuscripts to judge whether we are wrong in our disavowal of superposing such a flagrant bias.

The other external criticism (**White**) focused on a reference we made to White and White (1977) in the section of our target article entitled "Procedure." The sentence reads: "Each article had at least one author from the top 25 institutions in terms of citations of faculty research (White & White 1977), with 50% in the first six (Endler, Rushton & Roediger 1978)." The White and White (1977) citation *is* regrettably out of place in the middle of this sentence. It should be clear that the Endler et al. (1978) reference is the source we intended. How on earth the White and White reference crept into this paragraph is anyone's guess (it probably got transposed by clerical error), for we intended it for the preceding paragraph, where it does fit in; it certainly was not used here intentionally, nor is it a reflection of our failure to have read White's interesting work. (Again, **White** was a commentator we ourselves recommended.)

IV. Improving peer review

a. Blind reviewing. Nearly all of our commentators believed that the role of a responsible editor was critical to the peer-review process. It is hardly new or controversial to endorse such concepts as editorial accountability and responsibility. What *is* interesting is that almost all of our group of 56 commentators are or have been editors, associate editors, or on editorial boards of scientific journals; and despite differences in interpretation, no one attempted to defend peer review or hide its faults. Many believed that moving to blind reviewing would help remedy numerous concerns or would at least be "a step in the right direction" (**Armstrong, Crandall, Nelson, Palermo, Perlman, Presser, Rosenthal, Scarr, Tax & Rubinstein, Wilson**). However, a substantial group differed with the view that blind reviewing was effective (**Adair, Howe, Lazarus, Over, Thomas**). Two of these commentators (Howe and Over) felt that the concept of blind review was laudable, but that in practice it is really quite difficult to prevent authors from revealing their identities in subtle ways, if they so desire. However, recent findings of Rosenblatt and Kirk (1980) and some of our own data (partial return from study in progress) indicate that blind reviewing *is* fairly blind; only between 15 and 30% of reviewers have been found to be accurate in guessing an author's identity, and even then they exhibit very little confidence in the accuracy of author identification. A related concern is that editors are not blind, and in cases of doubt they are the ones who make the determination. Two of the physicist commentators felt that the author's name and institutional identity were valid and important pieces of information for reviewers to have (Adair, Lazarus). It would be tempting to speculate about the differences between physical and psychological research that yielded this disparity, but that would exceed the scope of this response. What we find interesting is the view of some that authors' institutional affiliations are a reflection of their ability as scientists (e.g., Lazarus), and the opposite view of others that because of shifting employment opportunities many very qualified scientists have taken up positions at low-prestige institutions and, hence that institutional affiliation is not a valid reflection of ability (e.g., Palermo, Tax & Rubinstein).

Some commentators stated that blind reviewing was not appropriate for grant-proposal reviews (**Louttit, Nelson**). According to this view it is important for reviewers to know the principal investigators' (PI's) backgrounds to assess their abilities meaningfully. We deliberately did not enter this fray in our target article - we were having enough difficulty with journal reviews. On the one hand, we recognize the value of knowing a PI's identity. A grant proposal is essentially a "promise." One prefers maximum assurances that the promise can be kept. The track record of the PI, institutional resources that are available, and the like, figure into this assurance. On the other hand, what evidence do we have that these variables are truly critical? Were it only a matter of personal taste on this question, we would unhesitatingly advocate "erring" on the conservative side and keeping grant reviews nonblind. However, recent studies by Stark-Adamec & Adamec (in press) cause us discomfort with this recommendation. The same grant that had been previously rejected with only her name (Stark-Adamec's) as PI was resubmitted with a more prestigious co-PI, and it was funded. The recent Cole, Cole, and Simon (1981) report does not suggest that personal or institutional bias exists in grant peer review (NSF), but the authors do report quite unsatisfactory reliabilities (also see **Manwell & Baker**'s commentary on NSF grant reviewing), and their design did not permit them to test the bias hypothesis as strongly as they would have liked. All of this should give Louttit pause in his enthusiastic belief that grant reviewing is more reliable than journal reviewing.

b. Reviewer accountability. Numerous commentators felt that the suggestion of an open review, wherein reviewers' identities are made known to authors, was not a good idea (**Cicchetti, R. Gordon, Scarr, Thomas, Wilson**). Most agreed with Scarr that anonymity protects younger, less powerful reviewers from possible retribution on the part of rejected authors. Also, it might compromise the effectiveness of one's review. However, Freese (1979) has argued persuasively against the present system of anonymous reviewers and has addressed most of the same concerns mentioned by our commentators. He concluded his rejoinder as follows:

> But I want one of my own questions answered - one to which these comments were not responsive and one for which survey data would be useless. Forget the fact that anonymous referees sometimes behave like gorillas. Remember that they are accountable to editors, not authors; that they have one-sided power over authors nonetheless; that a very small number of them (often just one) can deny an author access to significant professional rewards; and that, in such an event, an author has no right *in principle* to know who they are. What other voluntary role relationships quite like that can you name, and why can't you? (Freese, 1979, p. 245)

In order to improve accountability, reduce the incidence of ad hominem, insulting reviews, and the like, several commentators proposed either paying reviewers

(**Perloff & Perloff**), or acknowledging their names by adding them to articles they recommended (**Chubin**). Several commentators found merit in our description of an "author's review" (**Yalow**), but several found problems with it (e.g., **Howe, Wilson**).

V. Conclusion

If anything seems clear after reading all of the commentaries, it is that scientists differ widely in their interpretation of our study and in their suggestions for improving the peer-review system. A number of commentators felt that our study was a valuable demonstration of several serious ills in modern peer-review practices (e.g., unreliability, author-institutional bias, referee inadequacy), while others believed that nothing could be learned until a more complete experimental design was employed. As several of our commentators have pointed out, when one is dealing with reviewers, one must face the human side of science - differences of opinion.

The diversity of opinion among our commentators on every important issue etches in bold relief the difficulties of peer review. The only real consensus among the commentators is the belief that the present peer-review process needs reform. One might be surprised to discover lack of reviewer agreement, poor reliability, and possible bias, but having made such discoveries one might be even more surprised to note professional indifference in response to them. Will the peer-review system of science mimic that of the literary world and ignore the evidence, and with it the calls for reform (see **Ross**'s commentary)? In the three years that we have been struggling with some of these issues it has been a frequent observation of ours that a great professional inertia needs to be surmounted before we can improve our peer-review system. Frequently, we have seen suggestions for improvements countered with "it won't work." Unfortunately, this criticism has no empirical base, only opinion. We need to test suggestions for improvements to assess the impact they have on peer review. With one or two exceptions, all of the commentators who offered suggestions for improving peer review view indicated a need for continued and more extensive research to test empirically the utility and effectiveness of reform suggestions. Obviously, there is much work ahead. The study we conducted and the ensuing open peer commentary have been steps in the right direction, we believe. It is important for scientists to discuss openly the topic of peer review and to identify extant problems. Only then will solutions be forthcoming. Studies like **Whitehurst**'s, Cicchetti and Eron's (1979), and Rosenblatt and Kirk's (1980), which have examined reviewer training, the use of standard rating forms, and blind reviewing, have provided us with information that will be valuable in our efforts to shape a better reasoned, more workable review process. The area that seems to be most promising - that of cross-disciplinary comparisons - is still relatively unresearched. If **Moravscik** and **Adair** are right in stating that our outcomes could not obtain in physics, then a better understanding of the way peer review operates in other disciplines may be helpful to us, if not in remedying our problems, then at least in understanding them.

Finally, regardless of all other issues and questions raised by commentators, everyone would agree that science must uphold a fairness doctrine. This, to us, means that everyone should have fair access to journal space and federal funds. Fair is defined here as being judged on the merit of one's ideas, not on the basis of academic rank, sex, place of work, publication record, and so on. Peer review in science is a tribunal of sorts. It is instrumental in the great "sorting" process which ultimately is linked to the dissemination of professional rewards. It is necessary that we all insist on equal justice in this system. More important, peer review should be used in such fashion that it enhances rather than impedes the progress of science.

NOTES

1. Several editors refused our request for additional information (e.g., to furnish original reviewers' comments, to describe how detections came about) and in one case, the editor did not honor our request for confidentiality, attempting instead to alert other editors about our study.

2. While it is possible that our data show a "known" versus "unknown" institutional bias, the fact that out of three manuscripts initially submitted with a University of North Dakota affiliation only one was accepted does not lend strong support for this view. (In a letter to the editor one week after submission, this affiliation was changed to that of one of the fictitious institutions.)

References

Abelson, P. H. (1979) Problems of science faculties (editorial). *Science* 204:133. [CM]

Adair, R. K. (1981) Anonymous refereeing. *Physics Today* 34:13–15. [RAG]

American Psychological Association (1963–1969) *Reports of Project on Scientific Information Exchange in Psychology.* Vols. 1–3. Washington, D.C.: APA, [BCG]

(1972) Eight APA journals initiate controversial blind reviewing. *APA Monitor* 3:5. [taDPP]

(1973) Ethical principles in the conduct of research with human participants. *American Psychologist* 28:79–80 [JLF, BM]

(1980) Summary report of journal operations for 1979. *American Psychologist* 35:575. [taDPP]

Anastasi, A. (1958) *Differential psychology.* 3rd ed. New York: Macmillan. [GJW]

Armstrong, J. S. (1979) Advocacy and objectivity in science. *Management Science* 25:423–28. [JSA]

(1980a) Advocacy as a scientific strategy: The mitroff myth. *Academy of Management Review* 5:509–11. [JSA]

(1980b) Unintelligible management research and academic prestige. *Interfaces* 10:80–86. [JSA]

(1982, forthcoming) Research on scientific journals: Implications for editors and authors. *Journal of Forecasting* and *Medical Hypotheses.* [JSA]

Bartko, J. J. (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19:3–11. [GJW]

Bartley, W. W. (1962) *The retreat to commitment.* New York: Alfred A. Knopf. [MJM]

Benwell, R. (1979) Authors anonymous? *Physics Bulletin* 30:288. [RAG]

Bernard, H. R. (1980) CARS: Computer assisted referee selection. *Journal of Research Communication Studies* 2:149–57. [HRB]

Beyer, J. (1978) Editorial policies and practices among leading journals in four scientific fields. *Sociological Quarterly* 19:68–88. [JMB]

Blissett, M. (1972) *Politics in science.* Boston: Little, Brown. [MB]

Boffey, P. M. (1975) *The brain bank of America: An inquiry into the politics of science.* New York: McGraw-Hill. [CM]

Bowen, D. D.; Perloff, R. & Jacoby, J. (1972) Improving manuscript evaluation procedures. *American Psychologist* 27:221–25. [taDPP]

Brackbill, Y. & Korton, F. (1970) Journal reviewing practices: Authors' and APA members' suggestions for revision. *American Psychologist* 25:937–40. [taDPP]

Broad, W. J. (1980a) Imbroglio at Yale. 1. Emergence of a fraud. *Science* 210:38–41. [DdB]

(1980b) Would-be academician pirates papers. *Science* 208:1438–40. [DdB]

(1981a) Congress told fraud issue "exaggerated." *Science* 212:421. [CM]

(1981b) Fraud and the structure of science. *Science* 212:137–41. [DdB, CM]

(1981c) The publishing game: Getting more for less. *Science* 211:1137–39. [DdB]

Bronfenbrenner, U. (1977) Toward an experimental ecology of human development. *American Psychologist* 32:513–31. [taDPP]

Bruner, J. S. (1962) *On knowing: Essays for the left hand.* Cambridge, Mass.: Harvard University Press. [HM]

California may be sued on secret files. (1978) *Times Higher Education Supplement*, November 3, p. 5. [CM]

Campbell, D. T. & Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56:81–105. [rDPP]

Caplow, T. & McGee, R. J. (1958) *The academic marketplace.* New York: Basic Books. [CM]

Carey, W. D. (1975) Peer review revisited. *Science* 189:331. [MB]

Carta, D. G. (1978) Forum for rejected papers. *IEEE Spectrum* 15:13. [RAG]

Cherfas, J. (1980) Only the names have been changed to protect . . . whom? *New Scientist*, 20 March, p. 950. [*Ed.*]

Chubin, D. E. (1980) Competence is not enough. *Contemporary Sociology* 9:204–7. [DEC, CM]

Chubin, D. E. & Connolly, T. (1982) Research trails and science policies: Local and extra-local negotiation of scientific work. In: *Scientific establishments and hierarchies*, ed. N. Elias, H. Martins, & R. Whitley. Sociology of the Sciences, vol. 6, pp. 293–311. [ALP]

Cicchetti, D. V. (1980) Reliability of reviews for the *American Psychologist*: A biostatistical assessment of the data. *American Psychologist* 35:300–303. [DVC, tarDPP, GJW]

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* 49:373–83. [taDPP]

Cicchetti, D. V. & Eron, L. D. (1979) The reliability of manuscript reviewing for the *Journal of Abnormal Psychology*. *1979 Proceedings of the Social Statistics Section*, pp. 596–600. Washington, D.C.: American Statistics Association. [DVC, tarDPP]

Coe, R. K. & Weinstock, I. (1967) Editorial policies of major economic journals. *Quarterly Review of Economics and Business* 7:37–43. [JSA]

Cohen, J. (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70:213–20. [GJW]

Cole, J. R. & Cole, S. (1972) The Ortega hypothesis: Citation analysis suggests that only a few scientists contribute to scientific progress. *Science* 178:368–74. [CM, taDPP]

(1973) *Social stratification in science.* Chicago: University of Chicago Press. [CM]

(1981) *Peer review in the National Science Foundation: Phase II of a study.* Washington, D.C.: National Academy of Sciences. [*Ed.*]

Cole, S.; Cole, J. R. & Simon, G. A. (1981) Chance and consensus in peer review. *Science* 214:881–86. [*Ed.*, rDPP]

Cole, S.; Rubin, L. & Cole, J. R. (1977) Peer review and the support of science. *Scientific American* 237:34–41. [RTL, CM]

(1978) *Peer review in the National Science Foundation: Phase I of a study.* Washington, D.C.: National Academy of Sciences. [DEC, CM]

Collins, H. M. (1981) Stages in the empirical programme of relativism. *Social Studies of Science* 11:3–10. [MDG]

Colman, A. M. (1979) Editorial role in author-referee disagreements. *Bulletin of the British Psychological Society* 32:390–91. [AMC, taDPP]

(1981) *What is psychology?* London: Kogan Page. [AMC]

Cone, J. D. & Foster, S. L. (in press) Direct observation in clinical psychology. In: *Handbook of research methods in clinical psychology*, ed. P. C. Kendall & J. N. Butcher. New York: John Wiley & Sons. [JDC]

Cook, T. D. & Campbell, D. T. (1979) *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally. [rDPP]

Cox, W. M. & Catt, V. (1977) Productivity ratings of graduate programs in psychology based on publication in the journals of the American Psychological Association. *American Psychologist* 32:793–813. [taDPP]

Crandall, R. (1977) *How* qualified are editors? *American Psychologist* 32:578–79. [RC]

(1978a) Interrater agreement on manuscripts is not so bad! *American Psychologist* 33:623–24. [RC, taDPP]

(1978b) The relationship between quantity and quality of publications. *Personality and Social Psychology Bulletin* 4:379–80. [RC]

Crandall, R. & Diener, E. (1978) Determining authorships of scientific papers. *Drug Intelligence and Clinical Pharmacy* 12:375. [RC]

Crane, D. (1967) The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *American Sociologist* 32:195–201. [MDG, MJM, taDPP]

(1972) *Invisible colleges.* Chicago: University of *Chicago Press.* [DLE]

Davidson, J. M. & Davidson, R. J., eds. (1980) *The psychobiology of consciousness.* New York: Plenum. [MJM]

DeBakey, L. (1976) Reviewing. In: *The scientific journal: Editorial policies and practices*, pp. 1–23. St. Louis: C. V. Mosby Company. [LD]

(1978) Communication, biomedical: II. Scientific publishing. In: *Encyclopedia of bioethics*, ed. W. T. Reich, vol. 1, pp. 188–94. New York: Free Press. [LD]

DeBakey, L. & DeBakey, S. (1976) Impartial, signed reviews. *New England Journal of Medicine* 294:564. [LD]

Diener, E. & Crandall, R. (1978) *Ethics in social and behavioral research.* Chicago: University of Chicago Press. [RC]

Eckberg, D. & Hill, L. (1979) The paradigm concept and sociology. *American Sociological Review* 44:925–37. [DLE]

Editorial note. (1965) *Helgolander Wissenschaftliche Meeresuntersuchen* 12 (July): 218. [CM]

Editoral. (1979) *Physical Review Letters* 43. [RKA]

Endler, N. S.; Rushton, J. P. & Roediger, H. L. (1978) Productivity and scholarly impact (citations) of British, Canadian, and U.S. departments of psychology. *American Psychologist* 33:1064–82. [taDPP]

Freese, L. (1979) On changing some role relationships in the editorial review process. *American Sociologist* 14:231–38. [DEC, rDPP]

Garfield, E. (1979a) *Citation indexing: Its theory and application in science, technology, and humanities.* New York: John Wiley & Sons. [taDPP, MJW]

(1979b) *Journal citation report; A bibliometric analysis of social science journals in the ISI data base.* Philadelphia: Institute for Scientific Information. [taDPP, MJW]

Garvey, W. D. & Griffith, B. C. (1964) Scientific information exchange in psychology. *Science* 146:1655–59. [BCG]

(1971) Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist* 26:349–62. [DdB, BCG, taDPP]

Garvey, W.; Lin, N. & Nelson, C. (1970) Communication in the physical and social sciences. *Science* 170:1166–73. [BCG]

Gibbs, J. C. (1979) The meaning of ecologically oriented inquiry in contemporary psychology. *American Psychologist* 34:127–40. [taDPP]

Glenn, N. (1976) The journal article review process: Some proposals for change. *American Sociologist* 11:179–85. [DEC]

Goodstein, L. D. & Brazis, K. L. (1970) Credibility of psychologists: An empirical study. *Psychological Reports* 27:835–38. [JSA, taDPP]

Gordon, M. D. (1980) The role of referees in scientific communication. In: *The psychology of written communication*, ed. J. Hartley, pp. 263–75. London: Kogan Page. [MDG, JH, RMP, tarDPP, SP]

Gordon, R. A. (1980) The advantages of a simple system of optional published refereeing. *Speculations in Science and Technology* 3:607–9. [RAG]

Gottfredson, S. D. (1978) Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist* 33:920–34. [RO, taDPP]

Gove, W. R. (1979) The review process and its consequences in the major sociology journals. *Contemporary Sociology* 8:799–804. [taDPP]

Greenberg, D. S. (1980) Scams and sleaze in science. *Clinical Chemistry News* 6:5. [DEC]

Griffith, B. & Small, H. (1976) A Philadelphia study of the structure of science: The structure of the social and behavioral sciences' literature. In: *Proceedings, First International Conference on Social Studies of Science.* Ithaca, N.Y.: Society for the Social Studies of Science. [BCG]

Hagstrom, W. O. (1974) Competition in science. *American Sociological Review* 39:1–18. [CM]

Hall, J. (1979) Author review of reviewers. *American Psychologist* 34:798. [taDPP, GJW]

Hargens, L. (1975) *Patterns of scientific research: A comparative analysis of research in three scientific fields.* Washington, D.C.: American Sociological Association. [JMB]

Harnad, S. (1979) Creative disagreement. *Sciences* 19:18–20. [DVC, *Ed.*, WMH, taDPP]

(1982) Rational disagreement in peer review. (Submitted for publication.) [*Ed.*]

Hartmann, D. P. & Wood, D. D. (1981) Observational methods. In: *International handbook of behavior modification and therapy*, ed. A. E. Kazdin. New York: Plenum. [JDC]

Hawkins, R. G.; Ritter, L. S. & Walter, I. (1973) What economists think of their journals. *Journal of Political Economy* 81:1017–32. [JSA]

Hendrick, C. (1976) Editorial comment. *Personality and Social Psychology Bulletin* 2:207–8. [GJW]

(1977) Editorial comment. *Personality and Social Psychology Bulletin* 3:1–2. [taDPP]

Hensler, D. R. (1976) Perceptions of the National Science Foundation peer review process: A report on a survey of NSF reviewers and applicants. NSF publication #77-33. [RTL, CM]

Herrnstein, R. J. (1977) Doing what comes naturally: A reply to Professor Skinner. *American Psychologist* 32:1013-16. [taDPP]

Holden, C. (1979) Ethics and social science research. *Science* 206:357-340. [rDPP]

(1980) Not what you know, but where you're from. *Science* 209:1097. [SP]

Honig, W. (1980a) Concluding anti-relativity. *Speculations in Science and Technology* 3:361-63. [WMH]

(1980b) Editorials on our evolving policy: Opening statements; Further statements on speculation; "They laughed at Columbus" and other author syndromes; Our first year, Modesty and age "paradigms"; Einstein centennial issue - Alternates to special relativity; Mathematics in physical science, or why the tail wags the dog; Comment on submissions; Some additional thoughts on speculation; Anti-relativity. *Speculations in Science and Technology* 3:233-43. [WMH]

(1980c) The review process: Before, after and during. *Speculations in Science and Technology* 3:513-16. [WMH]

Horn, R. E. (1980) Results with structured writing using the information mapping writing service standards. Paper available from the author, Information Resources, Inc., 133 Massachusetts Avenue, Lexington, MA 02173. [JH]

Horrobin, D. F. (1974) Referees and research administrators: Barriers to scientific research? *British Medical Journal* 2:216-18. [CM]

Huxley, L. (1900) *Life and letters of Thomas Henry Huxley.* London: Macmillan and Co. [LD]

Ingelfinger, F. J. (1974) Peer review in biomedical publication. *American Journal of Medicine* 56:686-92. [taDPP]

Jones, R. (1974) Rights, wrongs and referees. *New Scientist* 61:758-59. [taDPP]

Kelly, W. (1972) *Pogo: We have met the enemy and he is us.* New York: Simon & Schuster. [DVC]

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. [JSA]

Koestler, A. (1971) *The case of the midwife toad.* London: Hutchinson. [CM]

Korten, F. & Griffith, B. (ca. 1970) Editorial review in psychological journals: A survey of authors, editors, and readers. Washington, D.C.: American Psychological Association. [BCG]

Kosinski, J. (1968) *Steps.* New York: Random House. [CR]

Koulack, D. & Keselman, H. J. (1975) Ratings of psychology journals by members of the American Psychological Association. *American Psychologist* 30:1049-53. [taDPP]

Kuhn, T. S. (1970) *The structure of scientific revolutions.* 2d ed., enl. Chicago: University of Chicago Press. [DLE, BM]

(1974) Second thoughts on paradigms. In: *The structure of scientific theories,* ed. F. Suppe, pp. 459-99. Urbana: University of Illinois Press. [DLE]

Kumar, K. (1979) Optional published refereeing. *Physics Today* 32:13-14. [RAG]

Lakatos, I. (1970) Falsification and the methodology of scientific research programmes. In: *Criticism and the growth of knowledge,* ed. I. Lakatos & A. Musgrave, pp. 91-196. Cambridge: Cambridge University Press. [MJM]

Lakatos, I. & Musgrave, A., eds. (1970) *Criticism and the growth of knowledge.* Cambridge: Cambridge University Press. [MJM]

Lazarus, D. (1980) Changes in "The Physical Review" and "Physical Review Letters." *Physical Review Letters* 45:1605-6. [RAG]

Leopold, A. C. (1978) The act of creation: Creative processes in science. *BioScience* 28:436-40. [CM]

Levenson, H.; Burford, B.; Bonno, B. & Davis, L. (1975) Are women still prejudiced against women? A replica and extension of Goldberg's study. *Journal of Psychology* 89:67-71. [RO]

Lewis, L. S. (1972) Academic freedom cases and their disposition. *Change* 4:8, 77. [CM]

(1975) *Scaling the ivory tower: Merit and its limits in academic careers.* Baltimore: Johns Hopkins University Press. [CM]

Lindsey, D. (1978) *The scientific publication system in social science.* San Francisco: Jossey-Bass. [DEC, rDPP, RO]

Lovas, S. (1980) Higher degree examination procedures in Australian universities. *Vestes* (Federation of Australian University Staff Associations) 23:10-20. [CM]

McCall, R. B. (1977) Challenges to a science of developmental psychology. *Child Development* 48:333-44. [taDPP]

McCartney, J. L. (1973) Manuscript reviewing. *Sociological Quarterly* 14:290, 444-46. [RC, taDPP]

McCutchen, C. (1976) An evolved conspiracy. *New Scientist* 70:225. [taDPP]

McGuigan, F. J. (1968) *Experimental psychology: A methodological approach.* 2d ed. Englewood Cliffs, N.J.: Prentice-Hall. [BM]

McGuire, W. J. (1973) The yin and yang of social psychology: Seven koan. *Journal of Personality and Social Psychology* 26:446-56. [taDPP]

McReynolds, P. (1971) Reliability of ratings of research papers. *American Psychologist* 26:400-401. [DP, taDPP]

Mahoney, M. J. (1976) *Scientist as subject: The psychological imperative.* Cambridge, Mass.: Ballinger. [MJM, CM, RO]

(1977) Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research* 1:161-75. [JSA, MJM, RO, taDPP]

(1979) Psychology of the scientist: An evaluative review. *Social Studies of Science* 9:349-75. [JSA, DEC, CM]

(1982) Psychotherapy and human change processes. In: *Psychotherapy research and behavior change,* ed. American Psychological Association. Washington, D.C.: American Psychological Association. [MJM]

(in press) Clinical psychology and scientific inquiry. *International Journal of Psychology.* [MJM]

Mahoney, M. J.; Kazdin, A. E. & Kenigsberg, M. (1978) Getting published: The effects of self-citation and institutional affiliation. *Cognitive Therapy and Research* 2:69-70. [JSA, MJM, RO, SP]

Mahoney, M. J. & Kimper, T. P. (1976) From ethics to logic: A survey of scientists. In: *Scientist as subject,* M. J. Mahoney, ed. pp. 187-93. Cambridge, Mass.: Ballinger. [JSA]

Manwell, C. (1979) Peer review: A case history from the Australian Research Grants Committee. *Search* (ANZAAS) 10:81-86. [CM]

(1981) An open letter to the president of FAUSA. *Australian Higher Education Supplement,* ed. J. Bremer, May 27. [CM]

Manwell, C. & Baker, C. M. A. (1979) The double helix: Science and myth in the act of creation. *BioScience* 29:742-46. [CM]

(1981) Honesty in science: A partial test of a sociobiological model of the social structure of science. *Search* (ANZAAS) 12:151-60. [CM]

Margulis, L. (1977) Peer review attacked (letter). *The Sciences* 17:5, 31. [CM]

Markle, A. & Rinn, R. C. (1977) *Author's guide to journals in psychology, psychiatry, & social work.* New York: Haworth Press. [taDPP]

Martin, B. (1979) *The bias of science.* Society for Social Responsibility in Science, P. O. Box 48, O'Connor, A. C. T., Australia 2601. [CM]

(1981a) The dismissal of Dr. M. E. Spautz from the University of Newcastle (New South Wales). Unpublished manuscript, availabe from Dr. Brian Martin, Dept. of Applied Mathematics, School of General Studies, Australian National University, Canberra, A. C. T. 2600. [CM]

(1981b) The scientific straightjacket: The power structure of science and the suppression of environmental scholarship. *Ecologist* 11:33-43. [CM]

Merton, R. K. (1968a) The Matthew Effect in science. *Science* 159:56-63. [DdB, CM]

(1968b) *Social theory and social structure.* New York: Free Press. [MDG, taDPP]

(1973) *The sociology of science.* Chicago: University of Chicago Press. [DLE]

Michie, D. (1978) Peer review and the bureaucracy. *Times Higher Education Supplement* August 4, p. 11. [CM]

Mitroff, I. I. (1974) *The subjective side of science.* Amsterdam: Elsevier. [IIM]

Mitroff, I. I. & Chubin, D. E. (1979) Peer review at NSF: A dialectical policy analysis. *Social Studies of Science* 9:199-232. [DEC, CM]

Moore, M. (1978) Discrimination or favoritism? Sex bias in book reviews. *American Psychologist* 33:936-38. [RO]

Moravcsik, M. J. (1980) *How to grow science.* New York: Universe Books. [*Ed.*]

Moyal, A. (1980) The Australian Academy of Science: The anatomy of a scientific elite: parts 1 and 2. *Search* (ANZAAS) 11:231-39, 281-88. [CM]

Neisser, U. (1976) *Cognition and reality: Principles and implications of cognitive psychology.* San Francisco: Freeman. [taDPP]

NIH Grants Peer Review Study Team (1978) Grants peer review: Opinions on the NIH grants peer review system. Report to the Director, NIH. [RTL]

NIH Study Committee, Dean E. Wooldridge, chairman. (1965) Biomedical science and its administration: A study of the National Institutes of Health. Washington, D.C.: The White House. [RTL]

Nisbett, R. E. & Wilson, T. D. (1977) The halo effect: Evidence for unconscious alterations of judgments. *Journal of Personality and Social Psychology* 35:250-56. [RO]

Oromaner, M. (1977) Professional age and the reception of sociological publications: A test of the Zuckerman-Merton hypothesis. *Social Studies of Science* 7:381-88. [taDPP]

Orr, R. & Kassab, J. (1965) Peer group judgment on scientific merit: Editorial refereeing. Presented to the Congress of the International Federation for Documentation, Washington, D.C. [BCG]

Over, R. (1981) Representation of women on the editorial boards of psychology journals. *American Psychologist* 36:885-91. [RO]

Patterson, E. H. (1969) Evaluation of manuscripts submitted for publication. *American Psychologist* 24:73. [DVC]

Peters, D. P., & Ceci, S. J. (1980) A manuscript masquerade. How well does the review process work? *Sciences* 20:16-19. [JJB]

Pfeffer, J.; Leong, A. & Strehl, K. (1977) Paradigm development and particularism: Journal publication in three scientific disciplines. *Social Forces* 55:938–51. [JMB]

Pheterson, G. I.; Kiesler, S. B. & Goldberg, P. A. (1971) Evaluation of the performance of women as a function of their sex, achievement, and personal history. *Journal of Personality and Social Psychology* 19:114–18. [RO]

Price, D. (1970) Citation measures of hard science, soft science, technology and nonscience. In: *Communications among scientists and engineers*, ed. C. E. Nelson & D. Pollack. Lexington, Mass.: D. C. Heath and Company. [BCG]

Ravetz, J. R. (1981) Avoiding fraud (correspondence). *Nature* 291:7. [CM]

Revusky, S. (1977) Interference with progress by the scientific establishment: Examples from flavor aversion learning. In: *Food aversion learning*, ed. N. W. Milgram, L. Knames & T. M. Alloway. London: Plenum. [JH, taDPP, GJT]

Robertson, P. (1976) Towards open refereeing. *New Scientist* 71:410. [RAG]

Roose, K. D. & Anderson, C. J. (1970) *A rating of graduate programs*. Washington, D.C.: American Council on Education. [DP, taDPP]

Rosenblatt, A. & Kirk, S. A. (1980) Recognition of authors in blind review of manuscripts. *Journal of Social Service Research* 3:383–94. [tarDPP]

Rosenthal, R. (1966) *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts; rev. ed., New York: Irvington, 1976. [taDPP, RR]

Rosenthal, R. & Rubin, D. B. (1978) Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences* 3:377–86. [tarDPP]

Ross, C. (1979) Rejected. *New West* 4:39–41. [JSA, CR]

(1980) Editors choice: None of the above. *Los Angeles Times*, February 17, p. 3. [JSA, CR]

Ross, P. F. (1981) The sciences' self-management: Manuscript refereeing, peer review, and goals in science. Unpublished manuscript. [taDPP]

Rowney, J. A. & Zenisek, T. J. (1980) Manuscript characteristics influencing reviewers' decisions. *Canadian Psychology* 21:17–21. [DP]

Roy, R. (1981) An alternative funding mechanism. *Science* 211:1377. [DEC]

Ruderfer, M. (1980) The fallacy of peer review – Judgment without science and a case history. *Speculations in Science and Technology* 3:533–62. [RAG, WMH, taDPP]

Rushton, J. P. & Roediger, H. L. (1978) An evaluation of 80 psychology journals based on the Science Citation Index. *American Psychologist* 33:520–23. [taDPP]

Saracevic, T. (1975) Relevance: A review and framework for thinking on the notion in information science. *Journal of the American Society for Information Science* 26:321–43. [BCG]

Sayre, A. (1975) *Rosalind Franklin and DNA*. New York: Norton. [CM]

Scarr, S. & Weber, B. L. R. (1978) The reliability of reviews for the *American Psychologist*. *American Psychologist* 33:935. [tarDPP, GJW, JCW]

Schaeffer, D. L. (1970) Do APA journals play professional favorites? *American Psychologist* 25:362–65. [JSA]

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. [DVC, taDPP, WAS, GJW]

(1977) Methodological consensus among journal referees: A follow-up study of JPSP manuscripts. Unpublished manuscript. [WAS]

Shaw, R. & Bransford, J., eds. (1977) *Perceiving, acting, and knowing: Toward an ecological psychology*. Hillsdale, N.J.: Lawrence Erlbaum. [MJM]

Siegfried, J. J. (1970) A first lesson in econometrics. *Journal of Political Economy* 78:1378–79. [JSA]

Slovic, P. & Fischhoff, B. (1977) On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance* 3:544–51. [JSA]

Smigel, E. D. & Ross, H. L. (1970) Factors in the editorial decision. *American Sociologist* 5:19–21. [DVC]

Snizek, W. E.; Fuhrmam, E. R. & Wood, M. R. (1981) The effect of theory group association on the evaluative content of book reviews in sociology. *American Sociologist* 16:185–95. [DLE]

Social Science Citation Index. (1977) *Guide and journal lists*. Philadelphia: Institute for Scientific Information. [DLE]

Stark-Adamec, C. & Adamec, R. (in press) Breaking into the grant proposal market. *International Journal of Women's Studies*. [rDPP]

Stinchcombe, A. & Ofshe, R. (1969) On journal editing as a probabilistic process. *American Sociologist* 4:116–17. [SP, rDPP]

Stumpf, W. E. (1980) "Peer" review. *Science* 207:822–23. [taDPP]

Swindel, R. F. & Perry, T. O. (1975) A previously unannounced form of the Gaussian distribution: The golden rule of arts and sciences. *Journal of Irreproducible Results* 21:8–9. [DVC]

Symposium. (1979) Reviews of Lindsey's *The scientific publication system in social science*. *Contemporary Sociology* 8:814–24. [DEC]

Szasz, T. (1973) *The second sin*. London: Routledge and Kegan Paul. [JSA]

Tobach, E. (1980) "... that ye be judged." In An evaluation of the peer review system in psychological research, chairman J. Demarest, open forum presented at the American Psychological Convention, Montreal. [taDPP]

Trafford, A. (1981) Behind the scandals in science labs. *U.S. News and World Report*, March 2, p. 54. [JSA]

Tuckman, H. P. & Leaky, J. (1975) What is an article worth? *Journal of Political Economy* 83:951–67. [RC]

Virgo, J. (1974) A statistical procedure for evaluating the importance of scientific papers. Ph.D. thesis, University of Chicago. [BCG]

Walster, G. W. & Cleary, T. A. (1970) A proposal for a new editorial policy in the social sciences. *American Statistician* 24:16–19. [taDPP]

Watkins, M. W. (1979) Chance and interrater agreement on manuscripts. *American Psychologist* 34:796–97. [RMP, tarDPP, JCW]

Webb, W. B. (1979) Continuing education: Refereeing journal articles. *Teaching Psychology* 6:59–60. [taDPP]

Webster, E. C. (1964) *Decision making in the employment interview*. Montreal: Eagle. [JSA]

Weick, K. E. (in press) Systematic observational methods. In: *The handbook of social psychology*, ed. G. Lindzey & E. Aronson. 3d ed. [JDC]

Weimer, W. B. (1977) A conceptual framework for cognitive psychology: Motor theories of the mind. In: *Perceiving, acting, and knowing: Toward an ecological psychology*, eds. R. Shaw & J. Bransford, pp. 267–311. Hillsdale, N.J.: Lawrence Erlbaum. [MJM]

(1979) *Notes on the methodology of scientific research*. Hillsdale, N.J.: Lawrence Erlbaum. [MJM]

Weimer, W. B. & Palermo, D. S., eds. (1981) *Cognition and the symbolic processes*. Vol. 2. Hillsdale, N.J.: Lawrence Erlbaum. [MJM]

Weiss, R. J. (1980). The use and abuse of deception. *American Journal of Public Health* 70:1097–98. [JLF, rDPP]

White, M. J. & White, K. G. (1977) Citation analysis of psychology journals. *American Psychologist* 32:301–5. [tarDPP, MJW]

Winer, B. J. (1971). *Statistical principles in experimental design*. 2d ed. New York: McGraw-Hill. [JLF]

Wolff, W. M. (1973) Publication problems in psychology and an explicit evaluation schema for manuscripts. *American Psychologist* 28:257–61. [taDPP]

Wolin, L. (1962) Responsibility for raw data. *American Psychologist* 17:657–58. [JSA]

Wright, R. D. (1970) Truth and its keepers. *New Scientist* 45:402–4. [LD]

Yalow, R. S. (1978) Radioimmunoassay: A probe for the fine structure of biology systems. In: *Les prix nobel en 1977*, pp. 243–64. Nobel Foundation. Stockholm: Almqvist & Wiksell. [RSY]

Yoels, W. (1974) The structure of scientific fields and the allocation of editorship on scientific journals: Some observations on the politics of knowledge. *Sociological Quarterly* 15:264–76. [JMB]

Yotopoulos, P. A. (1961) Institutional affiliation of the contributors to three professional journals. *American Economic Review* 51:665–70. [taDPP]

Ziman, J. (1968) *Public knowledge: The social dimension of science*. Cambridge: Cambridge University Press. [MJM, CM, BM]

(1976) *The force of knowledge: The scientific dimension of society*. Cambridge: Cambridge University Press. [AMC]

Zinberg, D. S. (1976) Education through science: The early stages of career development in chemistry. *Social Studies of Science* 6:215–46. [CM]

Zuckerman, H. (1970) Stratification in American science. *Sociological Inquiry* 40:235–57. [taDPP]

Zuckerman, H. & Merton, R. (1973) Patterns of evaluation in science: Institutionalization, structure and functions of the referee system. In: *The scoiology of science*, ed. N. Storer. Chicago: University of Chicago Press. [BCG, taDPP, SP]