The Journal of Mind and Behavior
Autumn 1991, Volume 12, Number 4
Pages 429–468
ISSN 0271–0137

Manuscript Review in Psychology: Psychometrics, Demand Characteristics, and an Alternative Model

Robert F. Bornstein

Gettysburg College

Manuscript reviews are intended to be objective, empirical assessments of the scientific worth of papers submitted for publication. However, critics have charged that manuscript reviews are unreliable, unconstructive, and biased in a number of ways (e.g., biased against new or unpopular ideas, against unknown or obscure researchers, and against researchers from less prestigious institutions). A review of the empirical literature in this area indicates: (1) that inter-reviewer reliability in manuscript assessments is clearly inadequate, (2) that reviewer bias can sometimes influence manuscript assessments, and (3) that there is a dearth of empirical data supporting the predictive and discriminant validity of manuscript assessment procedures. Based on the available evidence it seems that manuscript reviews are more strongly influenced by chance factors than by systematic reviewer or editorial bias. Nonetheless, our desire to conceptualize manuscript reviews in psychology as objective, empirical assessments has produced a number of undesired results. An alternative approach to manuscript review based on an adversary (i.e., legal) model rather than a scientific model is presented. Advantages of an adversary model as a method for identifying sound research are discussed. Changes in current publication policies that would allow research findings to be disseminated more efficiently are also described.

The role of the scientific journal has changed considerably since the first journals were established over 300 years ago. Originally, scientific journals were intended only to disseminate new findings efficiently and to establish publicly the source and "ownership" of new ideas and discoveries (Mahoney, 1987). However, as the number of professional scientists increased during the 19th and 20th centuries, the role of the scientific journal shifted from simply recording new information to evaluating scientific research and performing a "gatekeeping" function (Crane, 1967). At least two forces aside from the ex-

I would like to thank Drs. Rick Crandall, Paul R. D'Agostino, Mary A. Languirand, Ray Millimet, Janet M. Riggs, Raymond C. Russ, Thane S. Pittman and two anonymous reviewers for their helpful comments on earlier versions of this paper. Requests for reprints should be sent to Robert F. Bornstein, Ph.D., Department of Psychology, Gettysburg College, Gettysburg, Pennsylvania 17325.

ponential growth of the scientific community served to propagate and encourage the gatekeeping function of the scientific journal. First, because publication in an established journal gradually became associated with explicit approval and "legitimization" by the scientific community of new ideas and findings (Zuckerman and Merton, 1971), the need for journal editors to carefully screen manuscript submissions increased. Second, because research productivity became a primary criterion by which professional scientists are evaluated, motivation for scientists to publish simply for the sake of publishing increased (Mahoney, 1985), and safeguards against the dissemination of flawed or trivial research became even more important.

As a result of such professional and scientific concerns, the manuscript review system presently used by most scientific journals came into being. The general form of the review process today is straightforward: manuscripts submitted for publication are perused by a journal editor, who then solicits formal evaluations of the manuscript from reviewers who are knowledgeable regarding the topic addressed in the paper. Typically, two or more reviewers assess the methodological and conceptual soundness, potential importance and overall quality of a manuscript, and then forward their written critiques, along with a recommendation regarding publishability of the paper, to the journal editor. Based on the reviewers' recommendations and his or her own reading of the paper, the journal editor reaches a decision regarding disposition of the manuscript (e.g., publish as is, invite resubmission of the paper for re-review following revision, etc.).

Manuscript reviews are intended to serve several functions. First, reviews help to identify those manuscripts which report methodologically and conceptually sound research so that such research may be published in top-quality, highly visible journals. Second, reviews are intended to call to the authors' attention problems or flaws in the design and interpretation of their research, so that the final research product may be improved as a result of feedback from knowledgeable reviewers and journal editors. In addition, because manuscript reviews play a key role in determining which research gets published, reviews help to identify productive, talented researchers who may then be recognized and rewarded for their efforts by the scientific community.

In recent years, many aspects of the manuscript review system have been criticized on conceptual, methodological and even political grounds. Critics have charged that manuscript reviews are unreliable (Ciccetti, 1980; Munley, Sharkin, and Gelso, 1988), unconstructive (Glenn, 1976), illogical (Bradley, 1981; Walster and Cleary, 1970), and nasty (Garcia, 1981). Critics contend that reviewers are reluctant to recommend publication of nonsignificant findings (Kupfersmid, 1988) and replications of previous research (Mahoney, 1985), and are biased against new, innovative and unpopular ideas (Mahoney,

1987), and against unknown authors and less prestigious institutions (Peters and Ceci, 1982).

While no discipline which uses manuscript reviews to assess the quality of research papers has been immune from criticism, there is some evidence that the review process as it is presently conducted may be more effective and efficient in the "hard" sciences (e.g., biochemistry, physics) than in psychology (Bornstein, 1990c; Garvey, Lin, and Nelson, 1970; Pfeffer, Leong, and Strehl, 1977; Zuckerman and Merton, 1971). Inter-reviewer agreement for manuscript assessments is better in the hard sciences than in psychology and other social sciences (see Peters and Ceci, 1982 and commentaries). Reviewers of hard science papers also show greater consensus regarding suggestions for manuscript revision than do reviewers of psychological research (Beyer, 1978; Pfeffer et al., 1977). A smaller proportion of hard science than social science manuscripts require multiple submissions before being accepted for publication (Zuckerman and Merton, 1971). Review turnaround times and pre-publication lag times are shorter in the hard sciences than in psychology (Beyer, 1978; Garvey et al., 1970), and far fewer complaints regarding particularism, reviewer bias and publication politics are heard in the hard sciences than in psychology and other social sciences.

Although the review process appears to function better in the hard sciences than in psychology, it is not clear why this is so. To begin to investigate this issue, a brief discussion of potential sources of variability in manuscript assessments is needed.

Sources of Variability in Manuscript Assessments

In general, there are three potential sources of variability in manuscript assessments. First, manuscript reviews may be determined, in whole or in part, by the quality of research reported in the paper. Second, chance factors (i.e., unreliability) might account for some or all of the variability in manuscript reviews. Third, reviewer and/or editorial bias might account for some or all of the variability in manuscript reviews. If the available evidence suggests that manuscript quality is the primary determinant of the outcome of manuscript reviews, the validity of the review process as a mechanism for identifying sound research would be supported. However, to the extent that the second or third factors demonstrably influence the outcome of manuscript reviews, the validity of this procedure as a tool for the assessment of research papers would be called into question.

In this context, there are two plausible explanations for the diminished effectiveness of manuscript reviews in psychology. One possibility is that reviewers simply cannot reach a consensus regarding the quality of psychological research. Pfeffer et al. (1977) and others (e.g., Beyer, 1978; Yoels,

1974) suggest that preparadigmatic fields such as psychology will necessarily show less consensus regarding the value of a particular piece of research than will the more "mature" sciences (see also Kuhn, 1977). Other writers (e.g., Kupfersmid, 1988) have offered similar explanations for the perceived ineffectiveness of manuscript reviews in psychology. Whatever the underlying cause of reviewer unreliability may be, those who espouse this position assert that manuscript acceptance or rejection is unduly influenced by chance factors, but not by systematic bias against particular topics, findings, authors or institutions. If this position is correct, then the available data should suggest that reviews are unreliable, but that the internal validity of manuscript reviews in psychology is adequate (i.e., that factors such as the prestige of an author and her institution, or the consistency of the findings reported in a study with the a priori beliefs of the reviewer are unrelated to the likelihood of manuscript acceptance).

An alternative explanation is that chance factors alone do not account for the diminished effectiveness of manuscript reviews in psychology, but rather that systematic reviewer and editorial bias plays a significant role in determining which research gets published. A number of possible biases have been discussed in this context. Most critiques have focused on perceived biases against unknown or obscure authors (Mahoney, 1985), authors who work at less prestigious institutions (Peters and Ceci, 1982), nonsignificant findings (Kupfersmid, 1988), replications of previous research (Bornstein, 1990c), and findings that contradict a reviewer or editor's own beliefs (Abramowitz, Gomes, and Abramowitz, 1975). If this position is correct, then it should be possible to demonstrate empirically that at least some of these variables reliably predict how a manuscript under review will be evaluated—indicating that factors unrelated to the quality of the research being assessed compromise the internal validity of the manuscript review process in psychology.

Of course, the relationship between various aspects of reliability and validity of any assessment procedure is complex, and it would be misleading to imply that these dimensions can somehow be evaluated completely independently. As Anastasi (1988) and others (e.g., Berk, 1984) have pointed out, the reliability of an assessment instrument can have a substantial influence on various indices of validity (and vice versa; see Campbell and Stanley, 1963; Meehl, 1973). In particular, the reliability of an assessment instrument may serve to limit the predictive validity of that instrument, particularly when the reliability of the criterion measure is less than perfect. Thus, an analysis of the manuscript review process in which reviews are conceptualized as a kind of "test" or assessment instrument must take into account the nonindependence of relability and validity in manuscript reviews, and the implications of this nonindependence for the psychometrics of the manuscript review process.

The purpose of this paper is to assess the utility of the manuscript review process as a method for evaluating the quality of psychological research, and to suggest some ways that the review process may be improved. First, I will review research on reliability and reviewer bias in manuscript assessments to attempt to determine whether reliability problems, reviewer bias problems, or both are responsible for the perceived ineffectiveness of manuscript reviews in our field. Second, I will review research on the predictive and discriminant validity of reviews as assessments of the scientific worth of manuscripts, and I will describe some of the ways that psychologists have attempted to deal with the paucity of empirical evidence in this area. Third, I will discuss the dynamics of manuscript reviews, and how the demand characteristics of this process make objective, reliable assessments of psychological research papers impossible. Finally, I will offer an alternative model for manuscript review in psychology which will better meet the needs of our discipline.

The Psychometrics of Manuscript Review

While some subjectivity in the evaluation of psychological research is correctly regarded as a necessary aspect of the review process (see Peters and Ceci, 1982 and commentaries), psychology's goal of making manuscript assessments as objective and empirical as possible is reflected in the American Psychological Association's statements on publication policy (APA, 1983; Eichorn and VandenBos, 1985); in the frequent use of blind reviews to reduce examiner (i.e., reviewer) bias (see Ceci and Peters, 1984; Evans and Woolridge, 1987); and in the numerous studies assessing inter-reviewer reliability (e.g., Cicchetti, 1980, 1985; Marsh and Ball, 1981; Munley et al., 1988; Scarr and Weber, 1978; Watkins, 1979; Whitehurst, 1983, 1984, 1985), and threats to the internal validity of reviews (e.g., Abramowitz et al., 1975; Goodstein and Brazis, 1970; Mahoney, 1977; Peters and Ceci, 1982).

However, while we continue to conceptualize the manuscript review process as at least quasi-empirical, it fails even the most lenient, minimal psychometric criteria for controlled scientific assessment. We should be able to demonstrate that different expert reviewers come to similar conclusions regarding a manuscript (inter-rater reliability), and that manuscripts are assessed consistently if repeated measures are taken (test-retest reliability). We should also be able to demonstrate that reviews are not affected by variables unrelated to the quality of scientific research (e.g., the concordance of the findings reported in a study with the a priori beliefs of the reviewer), thereby providing evidence for the internal validity of manuscript reviews. Finally, we should be able to demonstrate that the manuscript review process has predictive and discriminant validity—that it can and does identify those manuscripts which will make the greatest contribu-

tion to the field, while detecting those that are methodologically flawed or make a less important contribution. To date, we have demonstrated none of these things.

Reliability in the Manuscript Review Process

There have been a number of archival studies assessing inter-reviewer reliabilities for manuscripts submitted to selected journals during predetermined time periods. Results of these studies have been mixed, with most studies reporting very low (even chance) inter-reviewer reliabilities in manuscript assessments (e.g., Cicchetti, 1980, 1985; Kunda and Nisbett, 1986; Marsh and Ball, 1981; Munley et al., 1988; Scott, 1974; Watkins, 1979), and some studies reporting higher inter-reviewer agreement when different procedures and statistical techniques are used (e.g., Crandall, 1978; Scarr and Weber, 1978; Whitehurst, 1984). There is some disagreement regarding appropriate statistical procedures to assess inter-reviewer reliability (see, e.g., Cicchetti, 1985; Whitehurst, 1985), with different researchers arguing for (and against) the use of Pearson correlation coefficients (Kunda and Nisbett, 1986), percentage of agreement (Scarr and Weber, 1978), coefficient Kappa (Watkins, 1979), the intraclass correlation coefficient (Marsh and Ball, 1981), or Finn's *r* (Whitehurst, 1984).

Regardless of disagreements regarding methodology and despite some variability in the outcome of these studies, one thing is clear from archival research on inter-reviewer reliability for manuscript assessments in psychology: even in those studies reporting relatively high reliabilities (e.g., Whitehurst, 1984), reliability coefficients do not reach minimal levels required for a psychometrically-sound assessment instrument. Intraclass correlation coefficients greater than .40 have rarely been reported in archival studies of the reliability of manuscript reviews in our field, and most studies report much lower reliabilities (see Kupfersmid, 1988; Marsh and Ball, 1981; Munley et al., 1988). These archival studies suggest that well over 50% of the variance in manuscript assessments is attributable to chance factors and/or reviewer bias (although these studies do not provide any information regarding the specific source of unreliability in reviews).

In addition to archival studies, there have been a few experimental studies that provide data regarding inter-reviewer reliabilities in manuscript assessments. When Mahoney (1977) asked 75 *Journal of Applied Behavior Analysis* reviewers to assess the publishability of a bogus manuscript describing a study of the effects of external reward on intrinsic interest, these reviewers could not agree in their assessments of the manuscript at better-than-chance level. Neither global recommendations regarding publishability nor more specific ratings of manuscript qualities (e.g., methodological rigor, potential impor-

tance) reached minimal levels for acceptable inter-reviewer reliability. Intraclass correlation coefficients ranged from –.07 (for ratings of topical relevance) to .30 (for ratings of data presentation, importance and publishability). Furthermore, the reviewers in Mahoney's study estimated that they would agree regarding manuscript evaluations about 70% of the time, while overall, actual agreement in ratings was well below 20%. Kunda and Nisbett (1986) similarly found that both professional psychologists and undergraduate students vastly overestimated inter-reviewer reliabilities for *Journal of Personality and Social Psychology* (*JPSP*) manuscript assessments and assessments of National Science Foundation (NSF) grant proposals. Not only cannot reviewers agree in their evaluations of manuscripts, but they are apparently unaware of how poor the consensus in manuscript assessments actually is.

The now well-known study by Peters and Ceci (1982) also assessed reliability in manuscript reviews. Peters and Ceci selected 12 recently-published psychology articles, altered their titles, authorships, authors' affiliations and some material in the papers' introductory sections, and then resubmitted the papers to the same journals that had published them within the previous 18-32 months. Three of the 12 deceptions were detected, and of the remaining nine manuscripts, eight were rejected due to serious methodological flaws detected by one or more reviewers. This study produced considerable controversy (see Peters and Ceci, 1982 and commentaries), was criticized on conceptual (Wilson, 1982), methodological (Beyer, 1982), and ethical grounds (Fleiss, 1982), and produced a tremendous personal and professional backlash directed at the authors of the study (Mahoney, 1987). The results of Peters and Ceci's study would seem to suggest that reliability in manuscript reviews is inadequate, in that eight of nine previously-accepted papers (89%) were now deemed unacceptable by the same journals that had recently published them.

However, there are some problems with this study as an evaluation of reliability in manuscript reviews. For example, it is difficult to determine whether Peters and Ceci's (1982) data bear more closely on the issue of inter-rater or test-retest reliability, since both between-reviewer and time-sampling variability potentially contributed to the unreliability in manuscript assessments found in this study. In addition, because a number of changes were made in the resubmitted papers (e.g., changes in the introduction sections), the second round of reviews was not based on the same material as the initial reviews. Furthermore, it is not at all clear that the same reviewers assessed these manuscripts during the second round of reviews (although this point is moot if journal—rather than reviewer—judgments are used as the unit of analysis in the Peters and Ceci study). Finally, it is possible that findings which were novel and interesting 1–3 years earlier had become "old news" by the time that the papers were resubmitted, or had been contradicted by more recent findings (Beyer, 1982).

Overall, while there are some inconsistencies in inter-reviewer reliabilities obtained in archival studies, substantial disagreement regarding how inter-reviewer reliability should be assessed, and significant methodological limitations in some of the experimental research on this topic, it is clear that the reliability of manuscript reviews in psychology is at best only moderate, and unacceptable by even the most lenient scientific standards for the development of an assessment instrument. Experimental and archival studies in this area produce consistent results, despite using very different methodologies and procedures. If one attempted to publish research involving an assessment instrument whose reliability data were as weak as that of manuscript reviews in psychology, there is no question that studies involving this psychometrically-flawed instrument would be deemed unacceptable for publication.

Threats to the Internal Validity of Manuscript Reviews

Studies investigating threats to the internal validity of manuscript reviews can be classified into four categories: (1) studies assessing "confirmatory bias" in manuscript assessments (i.e., the effects of concordance of the findings reported in a paper with reviewers' a priori beliefs; Abramowitz et al., 1975; Goodstein and Brazis, 1970; Mahoney, 1977); (2) studies assessing the effects of author and/or institutional status on manuscript evaluations[1] (e.g., Cole et al., 1981; Peters and Ceci, 1982); (3) studies assessing possible biases against publication of nonsignificant results (e.g., Atkinson, Furlong, and Wampold, 1982; Greenwald, 1975); and (4) studies assessing possible biases against publication of replication studies (e.g., Bozarth and Roberts, 1972; Sterling, 1970).

Concordance of Findings with the A Priori Beliefs of the Reviewer

There is fairly strong evidence that manuscript assessments can be affected by the degree of concordance of a study's findings with the beliefs and biases of the reviewer (Abramowitz et al., 1975; Goodstein and Brazis, 1970; Mahoney, 1977). Both Abramowitz et al. and Mahoney found that manipu-

[1] Ideally, blind reviews should minimize or even eliminate problems in this area. However, evidence collected to date suggests that—despite precautions taken to disguise the identity of authors prior to obtaining manuscript reviews—reviewers are often able to identify one or more authors of a paper from information contained in the manuscript (e.g., citations of "in press" papers, use of paradigms or stimuli unique to a researcher or laboratory). Ceci and Peters (1984) found that reviewers could correctly identify at least one of the authors of a manuscript over 35% of the time during a typical blind review. Similar findings were reported by Rosenblatt and Kirk (1980). Furthermore, neither reviewers (Evans and Woolridge, 1987) nor authors (Bradley, 1981) believe that the procedures used to ensure blind reviews are at all effective. Apparently, current blind review procedures are inadequate to disguise the identity (and hence, the status) of many authors.

lating the direction (not the strength) of a study's results strongly influenced reviewers' assessments of the paper. When the outcome of a study was consistent with the beliefs of the reviewer regarding political activism (Abramowitz et al.) or the effectiveness of a behavioral reinforcement procedure (Mahoney), reviewers were far more likely to recommend publication than when the opposite findings were reported. Mahoney also found that reviewers' justifications for rejecting manuscripts inconsistent with their own beliefs focused primarily on the methodology and scientific contribution of the study. Ratings of methodological soundness and scientific worth were significantly more positive for manuscripts whose results confirmed the reviewers' beliefs than for manuscripts reporting results that contradicted the reviewers' beliefs (although, of course, the topics investigated and methodologies employed were identical in both sets of manuscripts).

Similarly, Goodstein and Brazis (1970) found that assessments of manuscripts reporting a bogus study of astrological predictions of vocational choice were strongly affected by the outcome of the study: when the manuscript reported results favoring the predictive value of astrological signs, reviewers' assessments were significantly more negative than when results did not support the astrological position. Like Mahoney (1977), Goodstein and Brazis found that ratings of methodological soundness and scientific worth were affected by the direction of results, with manuscripts whose findings supported the astrological position rated as less methodologically sound and less important than manuscripts reporting the opposite findings. Goodstein and Brazis also found that manuscript assessments were related to the strength of reviewers' beliefs regarding astrology. Reviewers who described themselves as "strong disbelievers" were influenced by the direction of the study's findings to a greater degree than reviewers who were less skeptical regarding astrology.

Results of experimental studies indicate that confirmatory bias can, in some situations, influence reviewers' assessments of manuscripts. Furthermore, these studies suggest that when confirmatory bias is exhibited, reviewers invoke methodological criteria to justify rejection of manuscripts which report counterattitudinal findings. Although these results are highly suggestive, they must be interpreted with caution for two reasons. First, very few research topics have been employed in experimental studies of confirmatory bias, and the extent to which such biases may operate for other topics is unclear. The topics used in these studies were selected in part because they involved emotionally-charged issues; use of such topics increases the likelihood that confirmatory bias will be exhibited. Second, it is not clear to what extent demand characteristics of these experiments may be responsible for findings in this area. In the absence of archival research demonstrating that similar biases are present in vivo, the generalizability of these findings to actual manuscript reviews remains open to question.

Effects of Author and/or Institutional Prestige

With respect to the effects of author and/or institutional prestige on manuscript reviews, results of archival studies have been somewhat inconsistent. While Prescott and Csikszentmihalyi (1977) found that authors from prestigious institutions were overrepresented in 11 APA journals during a seven-year period, Zuckerman and Merton (1971) and Lindzey (1977) found author status—but not institutional prestige—to be strongly associated with publication frequency. In contrast, Cole, Rubin, and Cole (1978) found that while neither institutional prestige nor author status reliably predicted the likelihood of a positive review in NSF grant proposals, prestige of the department from which the proposal originated did predict the outcome of grant proposal assessments.

These correlational data are, of course, open to more than one interpretation. While it is possible that these results reflect systematic bias in manuscript and grant proposal assessments against low-prestige authors, departments or institutions, it is also possible that these data simply reflect the fact that more prestigious institutions (or departments or authors) produce higher-quality work. To assess rigorously the effects of author or institutional prestige on manuscript reviews, experimental evidence is needed to confirm and extend these archival findings.

Peters and Ceci's (1982) study would seem to be an ideal vehicle to test empirically the relationship of author and institutional status to the outcome of manuscript reviews. Peters and Ceci do in fact attribute the almost unanimous rejection of previously-published papers in their study to the lowered institutional status of the authors on the resubmitted manuscripts relative to the original ones. However, certain limitations of this study raise questions regarding Peters and Ceci's interpretation of their data. Most important, so many changes were made in the manuscripts between the original (pre-publication) and subsequent submissions that it is impossible to determine which—if any—of these changes were responsible for the markedly different ratings obtained in the second set of reviews. It is possible that author or institutional prestige was largely responsible for Peters and Ceci's findings, but that conclusion cannot be stated definitively based on the data they provide.

Stronger experimental evidence regarding this issue is provided by Mahoney, Kazdin, and Kenigsberg (1978). Mahoney et al. solicited assessments of a behavioral manuscript from 68 behaviorally-oriented reviewers. In half the manuscripts the fictitious author was from a prestigious university, and in the other half the author was from an obscure institution. Mahoney et al. found no effect of institutional status on manuscript assessments. Neither global ratings of publishability, nor ratings of specific manuscript qualities (e.g., innovativeness) were related to institutional status in this study.

Additional data are needed before firm conclusions are drawn regarding the effects of author and institutional status on manuscript evaluations. Neither archival nor experimental studies have produced very consistent results, and these studies do not provide strong support for the hypothesis that author or institutional status is a significant source of bias in manuscript reviews. Experimental studies of this issue typically utilize extremely obscure (or fictitious) institutions to represent low-prestige academic settings (e.g., "The Tri-Valley Center for Human Potential"; Peters and Ceci, 1982), and it is not clear that observed biases against such institutions would operate for less obscure (but still low-prestige) research settings (e.g., Gettysburg College). While it is certainly possible that author and/or institutional status influence manuscript reviews in some situations, it is not at all clear from the available data that such biases systematically influence manuscript assessments on a wide scale, that institutional status is a more (or less) significant factor than departmental or individual prestige, or that such biases operate anywhere but at the extremes of the status hierarchy (see Crandall, 1982). The relationship between prestige and manuscript assessments—if it exists at all—is not a simple one.

Bias Against Nonsignificant Results

Two lines of evidence suggest that reviewers and editors in psychology might be reluctant to recommend publication of studies that report nonsignificant results. First, the historical traditions, methodological procedures and statistical assumptions which characterize our field have led to a widespread mistrust of nonsignificant research findings (see Dar, 1987; Kupfersmid, 1988). This general mistrust of studies that retain the null hypothesis has been codified by the APA, which describes reporting of nonsignificant results as an important "defect" in submitted manuscripts, and goes on to explicitly caution researchers against submitting *any* nonsignificant results for publication unless ". . . repeated studies contradict a strong theoretical or empirical base" (APA, 1983, p. 19).

Second, archival research confirms that nonsignificant findings are rarely published in psychology journals. Sterling (1970) found that only 3% of the articles published in four psychology journals during a one year period, and 5% of a random sample of *Psychological Abstracts* citations described studies that retained the null hypothesis. Greenwald (1975) found that 12% of a sample of 199 *JPSP* articles reported nonsignificant results, while Bozarth and Roberts (1972) found that 6% of a sample of 1334 papers from three psychology journals reported nonsignificant results. Smart's (1964) data regarding this issue are consistent with these findings, and further suggest that a kind of progressive screening out of nonsignificant results may occur as psychological

research findings are disseminated. Smart found that while about 70% of a random sample of doctoral dissertations in psychology rejected the null hypothesis, 80% of papers presented at APA conventions and 90% of psychology journal articles rejected the null hypothesis.

Given the APA's explicit warning regarding publication of nonsignificant findings and the results of archival studies in this area, it would not be surprising to find that reviewers and editors set higher criterion levels in evaluating studies which accept the null hypothesis. An experiment by Atkinson et al. (1982) suggests that reviewers do in fact use more stringent criteria to evaluate manuscripts which report nonsignificant findings. Atkinson et al. asked 101 consulting editors from two APA journals to assess the publishability of a bogus manuscript where results were reported as statistically significant, marginally significant or nonsignificant. Reviewers were over three times more likely to recommend publication of manuscripts reporting statistically significant results than manuscripts which reported either nonsignificant or marginally significant results. Consistent with the findings of Goodstein and Brazis (1970) and Mahoney (1977) regarding confirmatory bias in manuscript assessments, Atkinson et al. found that reviewers' ratings of methodological rigor were significantly less positive when a study reported nonsignificant or marginally significant results than when the study reported significant results (although the methodologies in all three sets of manuscripts were identical).

The converging results of archival and experimental studies provide fairly strong evidence that reviewer bias can hinder the publication of studies reporting nonsignificant results. Clearly, additional experimental studies are needed to confirm and extend Atkinson et al.'s (1982) findings. Future research on this topic should address the question of whether bias against nonsignificant results is present for all research topics in psychology, or whether nonsignificant results are published more frequently in certain subfields than in others. Research is also needed to determine whether the plausibility of an hypothesis and the statistical significance of a study's results interact to determine the outcome of manuscript evaluations: while nonsignificant results might work against publication of studies investigating "mainstream" research topics, it seems likely that in a study of ESP (or some other counterattitudinal phenomenon), nonsignificant findings might actually *increase* the likelihood of receiving a positive review (see, e.g., Bornstein, 1990c; Goodstein and Brazis, 1970).

In addition, archival research is needed to determine whether the base rate for submission of studies reporting nonsignificant findings actually differs from the base rate of published papers which report nonsignificant results. Given Greenwald's (1975) finding that psychological researchers are nearly ten times less likely to submit a paper for publication if the null hypothesis is

retained than if it is rejected, the underrepresentation of studies reporting nonsignificant results in psychology journals probably reflects both systematic reviewer bias *and* researchers' reluctance to submit for publication studies that obtain nonsignificant findings.

Bias Against Replications of Previous Research

Evidence supporting the contention that reviewers and editors are systematically biased against publishing replications of previous research is weak. It is clear that direct replications are published much less frequently in psychology than in the hard sciences (see Bornstein, 1990c; Bozarth and Roberts, 1972; Greenwald, 1975; Mahoney, 1985). Of the 362 papers included in Greenwald's archival study of *JPSP* articles (described earlier), none reported a direct replication of previous research. Similarly, Bozarth and Roberts found that less than 1% of a sample of published papers in three psychology journals reported replication studies.

While this evidence is certainly suggestive (and not very encouraging to the author who hopes to submit a replication study for publication), it suffers from the same limitations as archival research assessing other forms of reviewer bias. In addition to gathering experimental data to confirm the results of these archival studies, researchers investigating this issue must take into account the base rate of manuscript submissions from different categories (in this case, the base rate of manuscript submissions reporting replication vs nonreplication studies), in order to discover exactly where the cause of the underrepresentation of replication studies in psychology journals resides. It is possible that reviewers are reluctant to recommend publication of studies which report replications of psychological research. However, an equally plausible interpretation of these data is that reviewers and editors are not at all biased with respect to such studies, but that researchers are simply unwilling to conduct (or submit) them.

Manuscript Reviews: "Random" or Biased?

Both archival and experimental studies confirm that inter-reviewer reliability in manuscript reviews is inadequate. Overall, findings regarding confirmatory bias and bias against nonsignificant results are somewhat stronger than findings regarding other forms of reviewer bias in manuscript assessments. The evidence supplied by studies of reviewer bias is interesting and compelling, in part because it appears to confirm what many of us have suspected all along (i.e., that we don't always get a fair shake during the review process). Nonetheless, much of the evidence regarding bias in manuscript assessments is open to more than one interpretation. Studies in this area have

not addressed the question of whether those biases that appear to influence manuscript reviews in certain situations represent deliberate, conscious reviewer and editorial bias or instead reflect unintentional, unconscious bias on the part of reviewers and editors. Studies to date have also failed to investigate the limits of reviewer and editorial bias (e.g., whether confirmatory bias is exhibited for all research topics or only controversial ones, whether replication studies are more likely to be published in certain subfields than in others, etc.).

The charge of reviewer or editorial bias is a very strong one, and one which (depending on the particular form of bias being charged) implies censorship, cronyism and other forms of unethical, unprofessional behavior. In the absence of stronger evidence to the contrary we must, for now, draw a more conservative (and benign) conclusion: on the basis of the available data, it appears that manuscript reviews are more strongly compromised by reviewer unreliability than by systematic reviewer and editorial bias (cf, Mahoney, 1985). The proportion of variance in manuscript assessments attributable to specific reviewer biases is far smaller than the proportion of variance attributable to unreliability in general (compare, e.g., the results of Goodstein and Brazis [1970], Mahoney [1977], Abramowitz et al. [1975] and Atkinson et al. [1982] regarding reviewer bias with those of Watkins [1979], Whitehurst [1984], Cicchetti [1985] and Munley et al. [1988] regarding inter-reviewer reliability). Furthermore, in most studies of reviewer bias, the proportion of variance attributable to specific forms of bias is far smaller than the within-manuscript variance (i.e., unreliability) in ratings.

It may well be that reviewer and editorial bias *are* important factors underlying the perceived ineffectiveness of manuscript reviews in psychology. Relatively few studies of this issue have been conducted to date, and future research might delineate other sources of bias or demonstrate that the variance attributable to reviewer bias is greater than these early studies suggest. However, the burden of proof in demonstrating the validity of this very serious charge must fall upon those who make such assertions, and stronger empirical evidence—from both archival and experimental research—is needed to support these claims.

To make a compelling case in support of the hypothesis that reviewer or editorial bias systematically influences manuscript assessments in psychology, researchers must provide converging evidence from experimental and archival studies, demonstrating: (1) that bias is present in both controlled laboratory (i.e., analog) studies and in actual manuscript reviews; and (2) that a significant proportion of the variance in manuscript assessments can be attributed to specific forms of bias. Researchers investigating this issue must also assess the generalizability of their findings by assessing the effects of bias on manuscript evaluations across a range of topics. In addition, to rigorously

assess the effects of reviewer bias on manuscript evaluations, future researchers should examine interactions among variables which potentially contribute to reviewer bias, rather than investigating only one source of bias in each experiment. The present approach only allows us to assess the main effects of different sources of bias on manuscript assessments. However, different forms of bias probably interact to determine the outcome of manuscript assessments in many cases. For example, it is very unlikely that nonsignificant findings would produce similar effects on reviewer assessments for studies which differ in a priori predictability (e.g., for studies of ESP or astrology vs studies of more mainstream research topics).

The Predictive and Discriminant Validity of Manuscript Reviews

The recent attention to reliability and reviewer bias issues in manuscript assessments has been valuable and constructive, calling our attention to some fundamental flaws in the review process in our field. However, by focusing mainly on reliability and reviewer bias issues in manuscript assessments, we may have inadvertently obscured a more fundamental and serious problem: there is no evidence supporting the predictive and discriminant validity of manuscript reviews as assessments of the quality of psychological research. Put simply, manuscript reviews may be "random" or they may be biased, but regardless, there is no empirical evidence which suggests that they can discriminate between high- and low-quality research.

Difficulties in selecting an appropriate criterion measure with which to assess research quality have hindered efforts to investigate this issue empirically. A thorough search of the literature revealed only one study that addresses this topic directly: Gottfredson (1978) utilized citation frequency as a criterion measure in assessing the validity of manuscript reviews, testing the hypothesis that if manuscript reviews have predictive and discriminant validity, then studies which receive highly positive reviews should be the most important, well-designed studies, and should therefore be cited more frequently than studies which receive less positive reviews. Unfortunately, there are numerous conceptual and methodological problems with Gottfredson's investigation (see Bornstein, in press, for a detailed discussion of these problems). In any case, Gottfredson's results are not reassuring. He obtained only low to moderate correlations between reviewers' estimates of manuscript quality and impact, and the number of citations received by a manuscript during the first nine years following publication. Reviewers' ratings of research impact were most strongly predictive of subsequent citation frequency ($r = .37$). Ratings of research quality did not fare as well ($r = .24$). Gottfredson's results suggest that reviewer assessments account for less than 15% of the variance in subsequent citation frequencies.

While difficulties in devising outcome criteria that can be used to assess the validity of manuscript reviews have hindered research efforts in this area, it is possible to design an experiment to assess rigorously the validity of manuscript reviews as a measure of the scientific worth of a piece of research. Such an experiment would not be terribly difficult to conduct. However, a rigorous assessment of the predictive and discriminant validity of manuscript reviews involves so many serious and insurmountable ethical problems that it will never be conducted. The reaction to such a study would make the many strong, negative responses to Peters and Ceci's (1982) research (e.g., Fleiss, 1982) seem mild by comparison. Nonetheless, a rigorous study of the validity of the manuscript review process as a method for identifying high-quality research might proceed as follows.

First, reviews of all manuscripts submitted to a particular journal (or group of journals) during a predetermined time period would be obtained. All the reviews would be classified according to their assessment of the publishability of the manuscript in question (e.g., clearly publishable, marginally publishable, marginally unpublishable, clearly unpublishable). Reviews could also be scored for their assessments of more specific evaluative dimensions (e.g., methodological rigor, potential importance). Then, regardless of the reviewers' recommendations, subsamples of manuscripts from each group would be published without revision. For each paper, assessments of methodological soundness, importance, impact on the field, and any other relevant criterion variables could be collected at some later date from various groups of judges (e.g., journal subscribers, experts in each field, random or stratified samples of academic psychologists), all of whom are blind to the content of the initial reviews. Repeated follow-up measures could be taken. Presumably, those manuscripts that had received positive reviews would later be rated as methodologically and conceptually tighter, and as more important and influential than manuscripts that received negative reviews initially. If so, then stronger evidence than has been collected to date would support the predictive and discriminant validity of the review process. If not, then the validity of this process as a mechanism for assessing psychological research would be called into question.

On the Paucity of Reliability and Validity Data Supporting the Use of Manuscript Reviews, and How We Deal With It

Although we regard manuscript reviews as a "test" or measure of the scientific worth of manuscripts in psychology, even a cursory reading of the APA's (1985) *Standards for Educational and Psychological Testing* reveals that this "test" fails miserably with respect to *every* technical criterion for establishing the reliability and validity of an assessment instrument (see APA,

1985, pp. 9–44).[2] How can we so readily accept the fact that the final link in the chain of scientific research in psychology is unable or unwilling to demonstrate empirically its reliability and validity? Although there have been some calls for increased attention to, and research investigating, this topic (e.g., Crandall, 1982; Mahoney, 1985), psychologists and other scientists have typically taken one of two approaches to this issue. Some have argued that this is not an issue at all; that the dynamics of the review process with its multiple checks and assessments will ensure that the best and most significant research is published. Others have argued that while the manuscript review system is imperfect, there is a self-correcting component to science, so that the best research will be recognized and applauded, and poor research will eventually reveal itself to be just that. In the following sections, I will illustrate why each of these positions is untenable.

Reliability and Validity of Manuscript Reviews as a "Non-Issue"

Because the manuscript review process has face validity (i.e., the system seems logical and there is no "hard" evidence that it is not valid), and is already in place and is perceived by many to be working reasonably well (see Scarr and Weber, 1978; Whitehurst, 1984), some psychologists suggest that we should simply leave well enough alone. This line of reasoning asserts that there is a natural subjectivity in the assessment of scientific research which is both necessary and constructive, and that to try to enumerate specific criteria for good vs bad science is not possible or desirable (see Peters and Ceci, 1982 and commentaries).

In light of evidence indicating that reviews are not reliable and that certain extraneous variables can influence the likelihood of receiving a positive review, this position seems difficult to defend. Face validity is hardly equivalent to predictive validity. While it is true that the review system is already in place, whether or not it is working well is debatable. Given the critical importance of manuscript reviews in the identification and dissemination of new ideas and findings, resistance to rigorously examining and improving this process is, to say the least, unconstructive.

---

[2] The manuscript review process also violates Principle 8 of the APA's (1981) *Ethical Principles of Psychologists*, which states that: "Psychologists responsible for the development and standardization of psychological tests and other assessment techniques utilize established scientific procedures . . . . In reporting assessment results, psychologists indicate any reservations that exist regarding validity or reliability" (p. 637). Given that: (1) evidence supporting the reliability of manuscript reviews is—at best—weak; (2) research on the predictive and discriminant validity of manuscript reviews is scanty, and clearly does not meet minimal criteria for "established scientific procedures"; and (3) recent studies have raised serious questions regarding the internal validity of manuscript reviews, each of us who is involved in the manuscript review process is in violation of Ethical Principle 8.

Science as Self-Correcting

Researchers have often argued that because science builds on earlier findings via replications and extensions of previous work, poorly done research and invalid findings will eventually be discovered and corrected (Koshland, 1987). There are no data which indicate that this view is not correct, although it seems likely that this self-correction process functions better in the physical than in the social sciences (Meehl, 1978).

However, evidence suggests that only a tiny minority of journal articles are actually read by more than a few researchers (Merton, 1973; Zuckerman and Merton, 1971). Mahoney (1987) estimates that a randomly selected article from any scientific journal is actually read by less than 1% of the journal's readers. Other researchers have reached similar conclusions regarding psychologists' surprising inattention to their own research findings (e.g., Garvey and Griffith, 1971; Kupfersmid, 1988). Apparently, much published research is never scrutinized, replicated, or integrated into the mainstream of psychological science.

In any case, the self-correcting process in science—even under optimal conditions—is only able to recognize a "false positive" finding (i.e., poor-quality research that should never have been published). There is no mechanism to determine the number or frequency of "false negative" findings (i.e., those manuscripts that should have been published but weren't). In the end, we actually have no idea how many potentially valuable findings go unpublished, or are published in obscure journals where they are never scrutinized (see Rosenthal, 1979 for a related discussion of the "file drawer problem" in social science research). The self-correcting properties of science are useful in correcting false positive manuscript reviews, but not all that useful in detecting false negatives.

Dynamics and Demand Characteristics of Manuscript Reviews

Occasional abuses of the manuscript review system do occur. Examples of biased, unconstructive—even *ad hominem*—reviews are not hard to come by (see Bradley, 1981; Garcia, 1981; Mahoney, 1987). However, the conclusion that in general, reviews are more "random" than biased is not all that surprising when viewed in the context of the dynamics and demand characteristics of the review process. The central obstacles to reliability and validity in manuscript reviews are unrelated to the effort, skill or intentions of manuscript reviewers and editors, nor to any flaws or limitations in the science of psychology. Rather, reviewers and editors are working within the constraints of a system that, in and of itself, literally *precludes* any possibility that manuscript reviews could ever be reliable or valid in an empirical sense.

Two Obstacles to Valid and Reliable Manuscript Reviews

Two aspects of the review system are primarily responsible for reliability and validity problems in this area, and are also the primary cause of authors' complaints regarding this system. The first problem is that reviewers and editors do not have at their disposal a useful operational definition of good research. The second problem is that journal page limitations do not permit us to publish all methodologically and conceptually sound manuscripts which are submitted for publication in psychology journals.

Problem #1. Because reviewers are never supplied with a precise operational definition of what characteristics and qualities must be present for a manuscript to be acceptable for publication, each reviewer is forced to employ his or her own subjective, idiosyncratic criteria in assessing manuscripts. Attempts to enumerate and quantify the dimensions on which manuscripts are evaluated have proved largely unsuccessful (see Chase, 1970; Gottfredson, 1978; Lindzey, 1978; Mahoney, 1977; Munley et al., 1988; Scott, 1974; Wolff, 1970). Many researchers correctly argue that certain of the criteria for "good science" differ, to some degree, as a function of the topic or issue (or sample) being studied, and furthermore cannot be broken down into a finite number of discrete categories. Nonetheless, without an operational definition of good research, we have no way to tell to what extent different reviewers are using the same criteria and criterion levels to evaluate manuscripts.

In fact, Chase (1970), Gottfredson (1978), Lindzey (1978), Mahoney (1977), Munley et al. (1988), Scott (1974) and Wolff (1970) all found that while reviewers can reach moderate agreement regarding what general criteria are useful in evaluating manuscripts, they cannot reach anything resembling a consensus regarding the relative importance of different criteria for manuscript evaluations. If reviewers are unable to reach a consensus regarding the relative importance of different criteria for good research under optimal conditions, it is difficult to believe that different researchers are using the same criteria and criterion levels to evaluate manuscripts in vivo, where ratings are made under much less controlled conditions.

Problem #2. Differences in the number of journal pages available for publication in the hard sciences vs psychology are such that the base rates for manuscript acceptance are typically around 75-80% in the hard sciences, and 20-25% in psychology (Bornstein, 1990a; Eichorn and VandenBos, 1985; Gordon, 1978; Zuckerman and Merton, 1971). Although some have argued that the different base rates for manuscript acceptance in psychology vs the hard sciences reflect the fact that psychological research is less robust and rigorous than hard science research (Skinner, 1987), there is no empirical evidence supporting this contention (Bornstein, 1988), while there is substantial evidence to the contrary (Hedges, 1987).

In fact, the primary reason that there are very different base rates for manuscript acceptance in psychology vs the hard sciences is economic: historically, federal and institutional funding for hard science research has been far greater than that for research in psychology. Authors of papers which appear in hard science journals typically pay all or part of the publication costs for their article. Grants and funding requests for hard science research nearly always include substantial provisions for publication costs in addition to funds to cover the costs of the research itself. Thus, in contrast to the economics of hard science publication, the economics of social science publication are such that the base rate for acceptance of submitted manuscripts is both low and relatively inflexible (see Bornstein, 1990a for a detailed discussion of this issue). In the following sections, I will offer some hypotheses regarding how these two factors affect the dynamics and demand characteristics of manuscript reviews in psychology, and I will then discuss why our attempts to treat this process as an objective, empirical assessment procedure have produced some undesired and destructive results.

Manuscript Review as a Clinical Decision

Manuscript review has more in common with a clinical decision than an objective, empirical assessment. Rather than being based on normative data and some actuarial assessment of the extent to which a given piece of research is methodologically sound and worthy of publication, manuscript evaluations are made by different reviewers based on idiosyncratic criteria which need never be fully elaborated or subjected to empirical testing. Thus, reviewers assign different weights to different aspects of the paper (e.g., methodological soundness, timeliness, potential importance, size of experimental effects; see Chase, 1970; Lindzey, 1978). A flaw that seems trivial to one reviewer is sometimes regarded as "fatal" by another (e.g., small sample size, failure to control for a particular extraneous variable). Like the clinician interpreting a projective test on the basis of his or her intuition and (necessarily limited) past experience with similar test protocols, the reviewer must rely on subjective criteria and personal knowledge of previous publications in the area to render a decision regarding the scientific worth of a paper. Like the clinical decision, the reviewer's judgment will undoubtedly have face validity, and because it will never be tested against some external criterion of validity, the reviewer never actually knows whether their (or any) manuscript reviews are valid in an empirical sense.

The conditions characterizing peer review are optimal for producing a classic illusory correlation (Chapman and Chapman, 1969). Most reviews will seem (at least to the reviewer) to be valid, and little motivation or opportunity to scrutinize and rigorously evaluate one's reasoning and judgment in

this arena is present. Thus, each new review may be perceived as additional confirmation of one's expertise. Just as illusory correlations serve to propagate social stereotypes in vivo (Nisbett and Wilson, 1977b), they may propagate stereotypes and biases regarding research topics and methodologies in manuscript reviews. Of course, there is no need to describe in detail the many reliability and validity problems that characterize clinical decisions in psychological assessments. These have been discussed at length by Meehl (1973) and others. The same flaws and problems inherent in all clinical decisions characterize those which make up the manuscript review process.

Manuscript Review as Signal Detection

In addition to sharing a number of unfortunate characteristics with the dynamics of a clinical decision, manuscript review is in many respects similar to a signal detection task. Here, the task of reviewer and editor is to detect those manuscripts (signals) whose perceived quality and contribution (i.e., whose intensity—d' is great enough that they can be reliably distinguished from the sample of flawed manuscripts (i.e., noise) in the system. As Peters and Ceci (1982) note, the response of a reviewer or editor to a given manuscript reflects a subjective criterion (called c or n in signal detection theory) which attempts simultaneously to minimize the number of "false positives" (poor manuscripts that should not have been accepted) and "false negatives" (publishable manuscripts that were erroneously rejected).

The fact that reviewers and editors must necessarily engage in a kind of signal detection task when assessing research is not, in and of itself, problematic. However, because there are a limited number of journal pages available for publication of psychological research, both reviewer and editor know ahead of time that the vast majority of submitted manuscripts must be rejected, regardless of the quality of the manuscript pool. Furthermore, the perceived cost of accepting (or recommending acceptance of) a flawed manuscript is greater than the cost of rejecting a manuscript about which one is unsure. Neither reviewer nor editor wants to get a reputation of being insufficiently rigorous in evaluating research papers. These two considerations, combined, will cause both reviewer and editor to set artificially stringent criterion levels in attempting to detect publishable manuscripts (see also Crane, 1967; Gordon, 1978; Gustin, 1973; Snizek and Fuhrman, 1979, for detailed discussions of reviewers' and editors' motivations for using overly stringent criteria to evaluate manuscripts).

It is not difficult to distinguish poor- from high-quality research when one is not forced to make a priori decisions regarding what percentage of manuscripts must fall into each category. However, the reviewer or editor attempting to distinguish poor- from high-quality research while simultaneously

attempting to stay close to a relatively inflexible base rate for journal acceptance which may not coincide with the base rate of methodologically and conceptually sound manuscripts submitted for publication, is in an impossible situation. The dilemma is similar to that of a subject in a signal detection experiment who thinks she perceives signals about 50% of the time (for example), but who knows that she is only permitted to call 20% of the presentations signals. It will still be easy to distinguish very strong from very weak signals (i.e., to distinguish excellent from terrible research), but attempting to dichotomize the numerous signals that are of middle-range intensity into "present" or "absent" categories (i.e., acceptable or unacceptable for publication) will necessarily be extremely unreliable, and the reasoning used to dichotomize this sample of signals will be idiosyncratic. In order to stay within the confines of a 20% ceiling for signals, our subject must find some way to decide which of the signals that she *thinks* she perceives should instead be classified as "noise."

The reviewer or editor faced with this dilemma, like the subject in the signal detection experiment, must come up with some criteria—no matter how idiosyncratic or arbitrary—to dichotomize these middle-intensity signals. Like the subject in a signal detection experiment who is asked to explain how he or she distinguished two signals whose intensities are fairly close (i.e., where the difference in d' between the two signals is barely noticeable), reviewer and editor must somehow explain their reasoning in making this distinction. However, unlike the subject in the signal detection experiment, the reviewer does not have the luxury of responding: "I don't know . . . the first one just seemed stronger." The reviewer must explain, using scientific language and reasoning, how she made a distinction which is extremely difficult if not impossible to make, using criteria which in many cases she cannot fully verbalize. The reviewer is forced to somehow rationalize what is, in many cases, a "gut" decision (see Cone, 1982; Mitroff, 1982; Palermo, 1982). As Nisbett and Wilson (1977a) and others have demonstrated, we cannot report accurately on the reasoning behind such decisions and judgments, so we are forced to reconstruct our reasoning after the fact.

Thus, the constraints of the manuscript review process in psychology are such that reviewers are frequently forced to make idiosyncratic, unreliable "gut" decisions and then weave a quasi-scientific, cognitive web of rationalization (the review and editorial response) around those decisions, justifying them to the author and to themselves. The stilted, unscientific reasoning often found in reviews of that large pool of methodologically sound manuscripts which must be rejected due to lack of journal space does not reflect an absence of skill, talent, effort or good intentions on the part of reviewers and editors. It is a consequence of their being assigned two incompatible tasks simultaneously (identify sound research, but not too much of it) and then

being forced to verbalize their reasoning using scientific logic and language, focusing exclusively on the paper under review rather than the demand characteristics of the process.

Most of the reliability problems in manuscript review, and many of authors' complaints regarding the review process stem from reviewers' and editors' earnest attempts to simultaneously perform these two incompatible tasks. While editors are sometimes willing to present journal space limitations as one of many reasons that a paper was not accepted, few if any journal editors are willing to reject a paper solely on that basis (i.e., to return to the author a letter which states, in effect: "The reviewers and I could find nothing seriously wrong with your paper, but there were others that we liked better, so we're rejecting yours because there just wasn't room for it"). To maintain the illusion that manuscript acceptance or rejection is based primarily on scientific considerations (rather than the overriding practical constraint of limited journal space), reviewers and editors are forced to provide scientific justifications for their decisions. However, because many of these decisions are in fact based on journal space limitations rather than the delineation of serious, uncorrectible flaws in a manuscript, the reasoning behind many manuscript rejections is illogical and unscientific. It is not surprising that authors are frequently dissatisfied and even angered by such reviews and editorial responses.

The Role of Rebuttal in the Manuscript Review Process

There are a number of pseudo-criteria and straw-man arguments invoked to justify rejection of methodologically and conceptually sound manuscripts that—for whatever reason—are not as compelling, interesting or attractive to reviewers and editor as other, equally sound manuscripts. Each of us who publishes papers in scientific journals has encountered such reviews at one time or another, and the arguments characterizing them are far too numerous to catalogue here. Garcia (1981)—who takes a somewhat less charitable view than I do of the manuscript review process—has provided an eloquent summary of this unfortunate, almost universal experience. He writes: "The author's confrontation with the editors often begins not with paranoid delusions, but with great hope and expectation. The author submits the final product of an arduous writing and rewriting process and receives a warm note of thanks from the editor. Then, after many months, the second editorial response finally arrives. It is apt to be a supercilious sophistry bearing so tenuous a relationship to the manuscript that the author concludes the consultants must have been out to lunch when the paper was being reviewed. Often, the critique is embellished with gratuitous personal insults . . . . The dissonance produced by the first courteous response and the second caustic one leads many authors to believe journals are governed by Janus-faced demons" (p. 149).

Negative reviews of methodologically and conceptually sound studies are typically illogical and unscientific, and do not present very compelling arguments regarding flaws and limitations in the paper. Ironically, such arguments would be easy to rebut, were the author given a chance to do so. Obviously, authors are always free to challenge a reviewer or editor's judgment, but many authors—especially those early in their careers—may not feel confident enough to challenge the authority of the review system directly. Furthermore, there is no established mechanism in place to handle authors' questions or complaints regarding reviews and editorial decisions. The end result is that reviewers almost invariably get the "last word" in the manuscript review process (Bornstein, 1990b; Glenn, 1976; Palermo, 1982).

The following examples illustrate the kinds of unscientific thinking and illogical reasoning which are invoked to justify rejection of methodologically and conceptually sound manuscripts. It is easy to imagine how one could summarily rebut such facile arguments, if only there were a mechanism readily available to do so.

Failing to Distinguish Correctible from Uncorrectible Flaws

Occasionally, a reviewer or editor will describe a number of trivial or correctible flaws in a manuscript (e.g., failure to control for the effects of some extraneous variable that there is no reason to suspect has anything to do with the findings; failure to utilize what reviewer or editor believes is the most appropriate statistical test for a particular analysis), and then recommend that the paper be rejected based on the presence of one or more of these flaws. This reasoning reflects one of two illogical (but unstated) assumptions: (1) that trivial or correctible flaws are as serious as more fundamental conceptual or methodological flaws; or (2) that some number of trivial, correctible flaws, when "combined," render a manuscript unpublishable. However, all flaws are not created equal. One serious conceptual or methodological flaw is enough to render a piece of scientific research unpublishable, but no number of trivial, correctible flaws are fatal to a manuscript.

Rejection Based on Unpredicted or Puzzling Findings

Many manuscripts are rejected on the grounds that the results were not predicted, that only certain of the results were predicted, or that the findings are "puzzling." The manuscript review system permits—and may even encourage—reviewers to utilize a "floating criterion level" to evaluate studies which differ in a priori predictability (see APA, 1983). However, to publish only those results that were predicted ahead of time or to set an unreasonably stringent criterion level for acceptance of puzzling findings is stultifying and

conservatizing, and violates a key tenet underlying progress in science (namely, that puzzling findings and disconfirmations of hypotheses are far more informative than predicted findings; Popper, 1972). Nonetheless, papers are rejected on this basis, reflecting a kind of superstitious belief that unpredicted findings in psychology are less reliable or "real" than predicted ones. The robustness of an experimental effect is unaffected by whether or not it was predicted.

Rejection Based on Timeliness or Importance

The "timeliness" or importance of a study is occasionally invoked as a criterion on which to base manuscript rejection. Reviewer and/or editor will sometimes find no serious methodological or conceptual problems with a manuscript, but will nonetheless reject it on the basis that it is not timely, interesting or valuable enough for publication. This reasoning can easily become tautological. Papers that are published in highly visible journals generate new ideas and influence the direction of a field; papers that are not published, or are published in obscure journals, do not (Mahoney, 1985). When a manuscript is rejected on this basis, a self-fulfilling prophecy is created, as the paper can never become timely (because it was not given the chance to do so)—reviewer and editor—in the vast majority of cases—will never be proved wrong (because the paper was not permitted to enter the mainstream of scientific research), and future research on the topic can then be rejected using the same reasoning (because the initial submission was never published). Questions regarding the timeliness or importance of well-designed, clearly-described research are better left to scientists working in the field, and ought not to be decided by reviewers and journal editors.

An Alternative Approach to Manuscript Review

Given the widespread dissatisfaction with the manuscript review process in psychology, it is not surprising that numerous suggestions for modifications of this system have been proposed (e.g., Armstrong, 1982; Bowen, Perloff, and Jacoby, 1972; Brackbill and Korten, 1970; Glenn, 1976; Kupfersmid, 1988; Mahoney, 1985, 1987; Millimet, 1981; Peters and Ceci, 1982; Walster and Cleary, 1970). Many interesting and valuable insights are contained in these proposals, and they deserve attention and further debate. However, because none of the proposals made to date contain components that can correct the two central obstacles to valid and reliable manuscript reviews in psychology, the proposed changes by themselves will not solve completely the problems that plague manuscript assessments in our field. Virtually all proposals to date have suggested revising the review process while still employing an

"empirical assessment" approach to manuscript reviews. It is time to confront our shared illusion that manuscript reviews in psychology—given the absence of an operational definition of good research and a skewed ratio of sound manuscripts to available journal space—are or could ever be valid and reliable assessments of the scientific worth of manuscripts. To improve the manuscript review process in psychology, we must stop treating manuscript evaluations as objective, empirical assessments of research products.

In the following sections, I describe an alternative approach to manuscript review. A preliminary version of this alternative approach was described in Bornstein (1990a, 1990b). The version described here is an elaboration, extension and refinement of those preliminary ideas. For the sake of clarity and completeness, I recapitulate certain points made in the preliminary discussions as I describe the procedural and conceptual underpinnings of the alternative model, paraphrasing some earlier ideas and quoting others directly.

The alternative approach to manuscript review that I will describe consists of two components. The first component involves a change in the way that manuscript reviews are conceptualized and structured. It is derived from an adversary (i.e., legal) approach rather than a scientific model, and is designed: (1) to minimize the potential for reviewer bias to influence the manuscript review process; and (2) to improve the overall quality of manuscript reviews (i.e., to make reviews more constructive). Some preliminary issues regarding the use of an adversary model in science were described by Levine (1974), who contrasted principles and assumptions of the adversary model with those of the empirical approach. More recently, Latham, Erez, and Locke (1988) applied principles of an adversary model to the problem of testing competing theories in psychology. I will apply principles of an adversary model to the manuscript review process.

The second component of this proposal involves a number of changes in journal publication policy that are intended.to make the dissemination of psychological research findings more efficient.

Manuscript Review: An Adversary Model

There are a number of differences between the scientific and legal conceptualizations of proof, argument and evidence. These are described in detail by Levine (1974). The most important difference between the scientific and legal models—at least in the present context—has to do with objectivity. While a scientist attempts to consider all sides of an issue as fully and objectively as possible (including those arguments and data that contradict the scientist's particular viewpoint), an attorney makes no attempt to present any evidence except that which supports his or her position. This fundamental difference between science and law reflects the conflicting goals of the

scientist and the attorney. Ideally (and perhaps idealistically), the primary motivation of the scientist has historically been conceptualized as the "search for truth." In this view, the scientist is presumed to be driven by a desire to discover and record new knowledge and information, regardless of whether that information does or does not confirm the scientist's own predictions, ideas and beliefs.

Unlike the scientist, a primary objective of the attorney is to argue against evidence that contradicts his or her position, regardless of whether the evidence has some validity. Simply put, the goal of the attorney in an adversarial courtroom situation is to win the case being argued, not necessarily to present all sides of the issue fully and objectively. Ironically, each of us who engages in scientific research is well aware of the fact that—despite the traditional view of the scientist as a pure and unbiased seeker of knowledge—the behavior of many researchers is somewhat closer to that of the attorney than to that of the "ideal" scientist.

There are, of course, many other important differences between the scientific and legal conceptualizations of proof, argument and evidence. For example, in contrast to the scientific model, which maintains a rather rigid and arbitrary (but clearly defined) criterion for evidence to be accepted as significant (the .05 probability level), an adversary approach acknowledges that "proof" is necessarily subjective and that no universal criterion can be applied across situations and circumstances. Along different lines, the nature and goals of rebuttal and "cross-examination" differ in the scientific and legal arenas (see, e.g., Brackbill and Korten, 1970; Glenn, 1976). The rules and norms regarding disclosure of evidence and data are also quite different in the laboratory and the courtroom (Levine, 1974).

I believe that the manuscript review system in psychology would function better within the parameters of a legal model than within the confines of the traditional "empirical assessment" approach. Because the manuscript review process in psychology cannot fulfill even the most lenient psychometric criteria for the reliability and validity of an assessment tool, and because our inability or unwillingness to acknowledge and correct this has produced a number of undesired results, I propose that we adopt an adversary model of manuscript review. It does not matter whether reviewer bias, reviewer unreliability, or both are responsible for the problems which plague manuscript reviews in psychology. This model will address and correct problems in both areas.

The basic tenets of an adversary model of manuscript review are simple. Rather than instructing reviewers to assess the methodological soundness, strength of findings and overall contribution of a manuscript and then make a recommendation regarding publishability of the paper, the reviewer should be asked to assume the role of adversary (i.e., prosecuting attorney). In other words, instead of attempting to assess the value of a manuscript and then

render a judgment of its worth, the reviewer should make every effort to challenge and rebut it. The reviewer's job would be to enumerate every conceptual and methodological flaw in the paper, rigorously critiqueing the author's ideas, methods, findings and conclusions.

At first glance, such a policy might seem to produce manuscript reviews that are invariably negative and unnecessarily critical. While it is true that when an adversary model is employed, manuscript reviews will focus exclusively on flaws and problems in a paper, such reviews need not necessarily be nasty, nor unconstructive. In critiqueing a manuscript, reviewers would still be free to make suggestions and point out important findings, additional analyses, and alternative interpretations of the data that the author failed to address. However, when an adversary model is employed, reviewers will no longer waste time describing the strengths and contributions of a study. After all, authors generally do a good job of that by themselves. Of course, if the reviewer believes that an author's assertions regarding the strengths and contributions of a study are incorrect or unwarranted, the reviewer can point that out in his or her critique of the paper. The reviewer will simply critique, correct and challenge the author's ideas, findings and assertions, point out ways that the study could be improved, and leave it at that.

Once the reviewer has completed his or her critique of the manuscript, the critique would be forwarded to an associate editor of the journal to which the manuscript was submitted. The associate editor—without assessing the reviewer's critique in any way—would record its arrival and forward a copy of it directly to the author. The author then would have an opportunity to rebut the reviewer's assertions, assuming the role of "defense attorney," but again making no pretense regarding objectivity. Objectivity is not the author's concern here; defense of his or her work is. When the author has completed the rebuttal of the review, this response would be forwarded to the associate editor, who will evaluate the critique and rebuttal, and arrive at a decision regarding publication.

Although asking the author of a paper to assume the role of "defense attorney" might seem awkward and somewhat problematic, this approach has some distinct advantages over the traditional legal model in which, in most cases, a third party acts as defense attorney. Because the author of a paper is already familiar with the logic and arguments put forth in a manuscript, and is very invested in defending his or her reasoning, the author will be the most competent, motivated "defense attorney" available to rebut the arguments of critics as a paper is assessed. While it may be true that the person who defends him-or herself in court has a fool for a lawyer, I nonetheless suggest that in an adversary model of manuscript review, the author of a manuscript is the most knowledgable, motivated "attorney" available to argue in defense of a submitted paper.

Needless to say, the adversary model involves a fundamental change in what an author can and should expect to hear in a typical manuscript review. Because the role of reviewer is shifted to that of "prosecuting attorney" in the adversary model, authors can no longer expect an objective assessment of their work, but merely a rigorous and thorough critique. This redefinition of the role of the reviewer has several advantages over the traditional empirical approach to manuscript review. For example, "[when] the reviewer's role is to critique the work rather than evaluate it, potential problems associated with nonblind reviews would be eliminated. Reviewers would be forced to stay close to the concepts and methods used in the study in formulating their critique, because they will be held accountable for their assertions and are aware that the author will soon be rebutting their criticisms. Blind reviews would no longer be necessary, nor desirable" (Bornstein, 1990b, p. 672). Given the numerous insurmountable problems associated with blind reviews of psychological research (see Bradley, 1981; Ceci and Peters, 1984), elimination of blind reviews would be a welcome change.

Furthermore, when the adversary model is employed, the quality of manuscript reviews would improve substantially. Weak, straw-man arguments could no longer be used to justify rejection of a methodologically and conceptually sound manuscript. Because reviews would routinely be rebutted by authors, reviewers would be forced to delineate compelling counterarguments challenging the ideas put forth in a paper. Consequently, "the frequency of ill-spirited, *ad hominem* reviews would diminish. Reviewers' fears that an unfavorable review might produce a backlash (e.g., a retaliatory negative assessment of one of the reviewer's own papers) would no longer be warranted: all reviews will be 'negative,' inasmuch as that is inherent in the prosecuting attorney role that the reviewer has been asked to adopt. Anonymous views would no longer be necessary, nor desirable" (Bornstein, 1990b, p. 672).

In order that reviewers would be able to make as thorough, informed and rigorous a critique as possible, copies of all "in press" and unpublished papers cited in the manuscript would be provided by the author (as is now done regularly in certain journals, e.g., *Science*). The same courtesy would be extended by reviewer to author for any unpublished or "in press" papers cited in the critique. Following the legal model, the "burden of proof" (i.e., the burden of demonstrating that a manuscript was seriously flawed) would rest on the reviewer, and a study would be considered "innocent until proven guilty" (i.e., publishable until shown to be significantly flawed).

If reviewers were obligated to return their critiques within the same time period that we now use for manuscript reviews, any significant lengthening of prepublication lag due to excessive time taken by the author in rebutting the initial critique would be the responsibility of the author, not the reviewer or editor. Rebuttals should take no longer than a few weeks—or at most, a

couple of months—to pull together. In additon, it seems likely that in some cases, the initial critique would uncover some serious, uncorrectible flaw in the design or execution of the study, in which case the author would have the option of choosing not to rebut the critique, and (temporarily, at least) withdrawing the paper from consideration to attempt to account for the issues raised by the reviewer. It would, of course, be possible to have two or more reviewers simultaneously critique the paper, as we now do, so that the author would receive copies of both critiques and have an opportunity to rebut the criticisms of both reviewers prior to editorial decision.

Limiting auctorial responses to some reasonable length would prevent editors from being confronted with excessively long rebuttals that present so many counterarguments in so much detail that a thorough, careful editorial assessment of the manuscript, critiques and rebuttal would be impossible. To prevent authors from using the rebuttal as a kind of "filibuster," rebuttals could be limited to two single-spaced typed pages, to the combined length of the reviewers' critiques, or to some fixed proportion of the length of the original manuscript (e.g., 10%). Similarly, reviewers' initial critiques could (and should) be limited to some reasonable length.

By having an associate editor always make the initial judgment regarding acceptance or rejection of a manuscript, the mechanism for "appeal" of a decision perceived as unfair would be clear. Such an appeal would be handled by the journal editor, and would consist of the initial critique and rebuttal, along with the associate editor's decision letter and a further (brief) follow-up rebuttal by the author (Bornstein, 1990b). Alternatively, both the editor and the associate editors of a journal could become involved in making initial editorial judgments. In this case, a third party could be appointed to serve as a kind of "appeals judge," before whom authors' appeals of an initial editorial decision could be brought.

Thus, when an adversary model is used, the potential for reviewer bias to influence manuscript assessments would be minimized. Even if a reviewer was highly motivated to prevent a particular manuscript from being published, the reviewer could not hinder publication of the paper simply by asserting that the manuscript is flawed or trivial. Rather, the reviewer would have to make a more compelling argument against acceptance of the paper than the author makes in favor of publication. Manuscript reviews would be held to the same high scientific standards as the manuscripts themselves.

In addition, the absence of a useful operational definition of "good research" would not hinder manuscript assessments when an adversary model is employed. Reviewers and editors would no longer be faced with an impossible signal detection task, nor an ill-defined clinical decision. Consequently, reviewers and editors would no longer be forced to justify rejection of methodologically and conceptually sound manuscripts using post hoc, quasi-

scientific reasoning. Instead, an editor would be free to evaluate papers with respect to how well they withstood the rigorous criticisms of expert reviewers, and select for publication those that fared best. Because the role of each party in the author–reviewer–editor interaction would be clearly-defined and unambiguous, and because the author is routinely given the opportunity to rebut the reviewer's assertions, we would likely find fewer complaints and fewer ill feelings on the part of authors, and an increase in the quality of manuscript reviews.

Of course, an adversary model cannot remove *all* possible sources of bias and unreliability from the review process. Identifying sound research involves the judgments of individuals, which are necessarily subjective and occasionally fallible. Ultimately, an editor must assess the manuscript, reviewers' critiques and auctorial rebuttal and use his or her best judgment to arrive at a decision regarding publication. Clearly, the potential for editorial bias to influence the outcome of manuscript reviews exists in an adversary model, as it does under the present system. However, because authors will rebut the assertions of reviewers prior to editorial judgment, editors will have the opportunity to consider explicit arguments representing both sides of an issue as a manuscript is being evaluated. Thus, the possibility that a manuscript will be rejected for illogical, unscientific reasons is minimized. Although an adversary model is not completely objective or totally free of bias, it is much less susceptible to potential bias problems than is the present system.

The Problem of Journal Page Limitations

While an adversary model helps to solve problems related to unreliable, unconstructive, biased evaluation of manuscripts, it does not address the question of how research findings could be disseminated more efficiently (see Bornstein, 1990c; Mahoney, 1985, 1987, for detailed discussions of this issue). The problem that remains is this: once the adversary model has identified that research which is methodologically and conceptually sound, how can this research be disseminated without increasing the number of journal pages prohibitively, or placing unreasonable financial burdens on authors, publishers or journal subscribers. The solution to this problem requires several changes in the present journal publication system.

First, psychology journals must begin requiring authors to help defray publication costs. This has long been the norm in hard science journals, but in the social sciences those journals that impose "page charges" are often perceived as inferior, less prestigious outlets for research findings. If modest publication fees were charged to authors by the most prestigious journals in our field, much of the resistance to this idea would eventually diminish. As in

the hard sciences, psychological researchers would begin including requests for funds to defray publication costs in external and institutional grant proposals. In order to increase the number of pages available to publish psychological research in highly visible journals, publication costs need not be prohibitive. If a publication charge of $20 per journal page was instituted for all articles published in the 18 APA Primary Journals, these publication fees would yield over $200,000 in additional revenues each year (an average of over $11,000 per journal, based on 1989 APA journal page allocations; see Summary Report of Journal Operations, 1990). Since the average APA journal article is approximately nine pages long (Summary Report, 1990), this publication fee would amount to about $180 per published article.

Because these figures are only estimates based on the most recent data available regarding APA journal operations, there is no way to predict how the imposition of publication charges would affect manuscript submission rates, the average length of submitted manuscripts, or other variables that potentially alter these estimates. Nonetheless, it is likely that if such charges became the norm in social science journals, academic institutions and other funding sources would begin to allocate funds to help defray these expenses (Bornstein, 1990a). Furthermore, because those researchers who are most productive tend to receive higher salaries (Gottfredson, 1978) and more external funding (Cole et al., 1981) than less productive researchers, there will likely be a positive relationship between the amount that a researcher is asked to pay for publication costs, and the amount that the researcher can afford to pay. Regardless, exceptions could be made for authors who are genuinely unable to pay these fees, or some form of sliding scale could be used to assure that fees are reasonable and appropriate for different authors.

It would also be useful to impose a moderate submission fee on all papers submitted for publication. As I noted earlier, "Reviewing manuscripts is costly, primarily in terms of professional time (i.e., the hours spent by reviewers assessing manuscripts and the time required on the part of journal editors to coordinate the review process), but also financially. Oddly, we have come to *expect* that reviewers' and editors' time is available to us in unlimited quantities (or, more accurately, that the only limit on the amount of reviewer and editorial time available to us lies in the number of papers that we can funnel into the review process). It is a *privilege* to receive feedback on your work from experts in your field" (Bornstein, 1990a, p. 673).

The privilege of having your work reviewed and critiqued by experts is surely worth a $25 or $50 submission fee, especially in light of the more thoughtful, reasoned reviews that authors would receive were an adversary model to be implemented. Journal submission fees would also generate considerable revenue, helping to defray publication (and subscription) costs. If a $25 submission fee were charged by all APA Primary Journals, this would

raise over $120,000 in additional funds each year, based on 1989 APA journal submission data (Summary Report, 1990).

In addition to (or in lieu of) helping to defray journal publication costs, the additional funds made available by modest publication charges and submission fees might be used to allocate some additional pages to each APA journal. If the number of pages published by the APA Primary Journals were increased by only 10%, there would be enough journal pages available to publish more than 150 additional full-length papers each year (Summary Report, 1990).

Regardless of whether these modest publication and submission fees were used by journals to increase the number of pages published, increasing the number of Brief Reports published in APA journals would allow a greater proportion of submitted manuscripts to be accepted without expanding the size of these journals prohibitively. Papers that are methodologically and conceptually sound, but whose contribution may not be great enough to justify a large amount of journal space could be published as Brief Reports consisting of 2–3 journal pages, with a longer version of the paper available directly from the author if a fuller treatment of the topic is warranted (see also Bornstein, 1990c, for a detailed discussion of the advantages of increasing the proportion of Brief Reports in psychology journals). This use of Brief Reports has been implemented successfully by the *Journal of Consulting and Clinical Psychology*, and modified versions of this approach are used by most other APA journals, as well as by a number of other journals in the social and biomedical sciences. Increased use of Brief Reports would allow researchers to keep abreast of a greater number of findings and ideas in less time, more efficiently (since a greater number of studies are available in a few highly-visible journals), and without onerous financial burdens placed on journal publishers or subscribers (Bornstein, 1990a).

When an adversary model of manuscript review is coupled with these practical changes in publication policies, both internal problems (e.g., reliability and bias issues) and external limitations (i.e., financial constraints) that have diminished the effectiveness of the manuscript review process in psychology would be minimized. Because authors would routinely be given an opportunity to rebut reviewers' assertions, the quality of manuscript reviews would increase. Because researchers would have ready access to a greater number of findings (via increased use of Brief Reports and modest increases in journal page allocations), time and money that is presently being spent conducting a study that your colleague across town actually tried (unsuccessfully) to run last year could be saved and spent on new, untested ideas. As it stands, given reviewers' apparent reluctance to recommend publication of nonsignificant, counterattitudinal and "puzzling" findings, one can only guess at how often researchers have wasted time and money trying to reinvent a wheel that other researchers had already discovered didn't work in the first place.

Some Initial Reactions to the Proposed Model

Three *Journal of Mind and Behavior* reviewers offered their reactions to the adversary model during the initial review of this paper. Although I do not agree with all of their criticisms, the reviewers made many excellent points regarding the proposed model and its implications. Their reactions were thoughtful and constructive, and may well reflect the kinds of responses that many readers of this paper will have. Thus, it is worthwhile to mention briefly a few of the reviewers' most intriguing and challenging comments regarding perceived flaws and weaknesses in the adversary model.

Reviewer A was concerned with the problems that might result from utilizing a manuscript review system based on the legal model, and expressed these concerns eloquently. Thus, Reviewer A wrote: "In my opinion, the legal system in this country can be held responsible for many of today's evils . . . . I doubt that the pursuit of truth can be turned over to an institution or a social process in which the participants are not personally accountable and sworn to the quest for truth. So let me challenge the author . . . tell me how an adversarial system will help remove the vanity, the guile, the ambition—the evil—from the hearts of those who merely claim to be scientists?"

I believe—as does Reviewer A—that an adversary system will not turn overly ambitious or machiavellian researchers into "ideal" scientists whose sole motivation is the quest for truth and knowledge. Of course, the traditional manuscript review system also cannot prevent such individuals from allowing personal and political concerns to bias their reviews of others' work. However, it is worth noting that while no review system can eliminate completely reviewer (or author) bias from the research enterprise, the adversary model might well do a better job than the traditional model in minimizing these biases. Because reviewers' assertions will routinely be challenged and rebutted by authors when an adversary model is employed, reviewers will be forced to frame their criticisms more carefully and conservatively under this system than they do under the present system (where reviewers are aware that in the vast majority of cases, their assertions will go unchallenged). Perhaps this aspect of the adversary model will help to minimize reviewer bias and increase reviewer accountability. Although the adversary model might not be useful in altering reviewers' base motivations, it could prove to be a highly effective means of extinguishing certain undesirable and destructive reviewer behaviors.

Along somewhat similar lines, Reviewer B raised an important issue regarding the ability of the adversary model to identify methodologically and conceptually sound research. Reviewer B argued that "the author's suggestion that use of the adversary model will lead to the identification of research which is methodologically and conceptually sound may represent a 'leap of

faith.' . . . All one may be ble to say is that the manuscript 'survived' a round of rational criticism. In other words, a different set of reviewers and authors may have come to a different conclusion regarding the manuscript."

I agree completely with Reviewer B on this point. Particularism can influence reviewers' responses under both the traditional model and the adversary model of manuscript review. Of course, the question of whether the adversary model can minimize reviewer bias and particularism more effectively than does the traditional approach to manuscript review is ultimately an empirical one. Nonetheless, I would argue that the tension created by the reviewer–author "debate" that is the centerpiece of the adversary model will result in a mechanism for the identification of methodologically and conceptually sound research that—while far from perfect—is less susceptible to bias and unreliability problems than is the present review system.

Reviewer C began by pointing out the ironic, "Catch-22" aspect of reviewing a manuscript like this one, and wrote: "Dare I recommend to reject this manuscript lest I be accused of bias, unreliability and being an agent of a corrupt journal system? No, I dare not." Reviewer C then went on to make a number of cogent points regarding the strengths and weaknesses of the adversary model. He expressed concern that "the adversary model does not provide for manuscripts that under the present system would have been deemed 'quite acceptable' by a reviewer or reviewers . . . . It seems to be inefficient to demand that these same reviewers be forced to find all that they see wrong with a manuscript (and not identify their actual approval of it for publication), thereby requiring the editor to make the decision that would have been made by this time in the author's favor."

Reviewer C makes an excellent point. Several colleagues who commented on this paper before it was submitted for publication expressed similar concerns. My response to this point is twofold. First, the proportion of manuscripts that are deemed acceptable without revision following an initial submission is quite small. Eichorn and VandenBos (1985) estimated that approximately 2% of all initial manuscript submissions are accepted for publication "as is." Thus, this is a relatively rare event. Second, a positive response to a paper can be communicated clearly by a reviewer even within the confines of the adversary model, if the reviewer: (1) focuses mainly on minor, correctible flaws in a manuscript; and (2) explicitly identifies these flaws as minor and correctible ones. It is important to note, in this context, that if a reviewer erroneously identifies minor, correctible flaws as "fatal" problems in a manuscript (a situation which, unfortunately, occurs with some regularity under the present system; see Bradley, 1981), the author will have an opportunity to challenge this assertion before an editorial decision is made.

Conclusion

Clearly, adopting an adversary model of manuscript review involves some practical problems. However, the fact that altering our approach to the review process might be difficult and complicated is not reason enough to continue using a model which in its present form has many serious and insurmountable flaws. The present approach to manuscript reviews has hindered the science of psychology, wasted time and money, and alienated many talented researchers and writers. The proposed changes in editorial and publication policies would make the identification and dissemination of methodologically sound psychological research much more efficient. Nonetheless, even if an adversary model of manuscript review in psychology is not adopted wholesale, if the problems and limitations of the present system are scrutinized more fully in the context of this proposal, then this paper will have served its purpose.

It is worth noting, though, that psychology is in a position to take the lead in this area. Just as psychologists have described many techniques in experimental design and data analysis that have since been adopted by other disciplines, we now have the opportunity to be at the forefront in delineating and implementing procedures for manuscript review that might better serve other scientific disciplines—in fact, every field that relies on refereed journals as a primary source for the dissemination of new ideas or findings. Although the luxury of a far greater number of journal pages available for publication has helped the hard sciences to circumvent some of the problems that hinder manuscript assessments in our field, peer review of hard science research is by no means perfect, and it is likely that hard science journals could and would benefit by innovations in this area. By rigorously scrutinizing the manuscript review system in psychology, and then acting to correct those aspects of the system that do not function as well as they could, we will not only benefit our field, but other disciplines as well.

References

Abramowitz, S.I., Gomes, B., and Abramowitz, C.V. (1975). Publish or politic: Referee bias in manuscript review. *Journal of Applied Social Psychology, 5*, 187-200.

American Psychological Association. (1981). Ethical principles of psychologists. *American Psychologist, 36*, 633–638.

American Psychological Association. (1983). *Publication manual of the American Psychological Association* (3rd edition). Washington, D.C.: Author.

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, D.C.: Author.

Anastasi, A. (1988). *Psychological testing* (sixth edition). New York: MacMillan.

Armstrong, J.S. (1982). Research on scientific journals: Implications for editors and authors. *Journal of Forecasting, 1*, 83–104.

Atkinson, D.R., Furlong, M.J., and Wampold, B.E. (1982). Statistical significance, reviewer evaluations and the scientific process. *Journal of Counseling Psychology, 29*, 189–194.

Berk, R.A. (1984). *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.
Beyer, J.M. (1978). Editorial policies and practices among leading journals in four scientific fields. *Sociological Quarterly, 19*, 68–88.
Beyer, J.M. (1982). Explaining an unsurprising demonstration: High rejection rates and scarcity of space. *Behavioral and Brain Science, 5*, 202–203.
Bornstein, R.F. (1988). Radical behaviorism, internal states and the science of psychology. *American Psychologist, 43*, 819–821.
Bornstein, R.F. (1990a). Epistemic progress and journal page limitations: A proposal for increasing the base rate of manuscript acceptance in psychology journals. *American Psychologist, 45*, 673–674.
Bornstein, R.F. (1990b). Manuscript review in psychology: An alternative model. *American Psychologist, 45*, 672–673.
Bornstein, R.F. (1990c). Publication politics, experimenter bias and the replication process in social science research. *Journal of Social Behavior and Personality, 5*, 71–81.
Bornstein, R.F. (in press). The predictive validity of manuscript reviews: A neglected issue. *Behavioral and Brain Sciences*.
Bowen, D.B., Perloff, R., and Jacoby, J. (1972). Improving manuscript evaluation procedures. *American Psychologist, 27*, 221–225.
Bozarth, J.D., and Roberts, R.R. (1972). Signifying significant significance. *American Psychologist, 27*, 774–775.
Brackbill, Y., and Korten, F. (1970). Journal reviewing practices: Authors' and APA members' suggestions for revision. *American Psychologist, 25*, 937–940.
Bradley, J.V. (1981). Pernicious publication practices. *Bulletin of the Psychonomic Society, 18*, 31–34.
Campbell, D.T., and Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton–Mifflin.
Ceci, S.J., and Peters, D. (1984). How blind is blind review? *American Psychologist, 39*, 1491–1494.
Chapman, L.J., and Chapman, J.P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271–280.
Chase, J.M. (1970). Normative criteria for scientific publication. *American Sociologist, 5*, 262–265.
Cicchetti, D.V. (1980). Reliability of reviews for the *American Psychologist*. *American Psychologist, 35*, 300–303.
Cicchetti, D.V. (1985). A critique of Whitehurst's "Interrater Agreement for Journal Manuscript Reviews." *American Psychologist, 40*, 563–568.
Cole, S., Rubin, L., and Cole, J.R. (1978). *Peer review in the NSF*. Washington, D.C.: National Academy of Sciences.
Cole, S., Cole, J.R., and Simon, G.A. (1981). Chance and consensus in peer review. *Science, 214*, 881–886.
Cone, J.D. (1982). Criterion problems in journal review practices. *Behavioral and Brain Sciences, 5*, 206–207.
Crandall, R. (1978). Interrater agreement on manuscripts is not so bad! *American Psychologist, 33*, 623–624.
Crandall, R. (1982). Editorial responsibilities in manuscript review. *Behavioral and Brain Sciences, 5*, 207–208.
Crane, D. (1967). The gatekeepers of science. *American Sociologist, 2*, 195–201.
Dar, D. (1987). Another look at Meehl, Lakatos and the scientific practices of psychologists. *American Psychologist, 42*, 145–151.
Eichorn, D.H., and VandenBos, G.R. (1985). Dissemination of scientific and professional knowledge. *American Psychologist, 40*, 1301–1316.
Evans, K., and Woolridge, B. (1987). Journal peer review: A comparison with employee peer performance appraisal. *Journal of Social Behavior and Personality, 2*, 385–396.
Fleiss, J.L. (1982). Deception in the study of the peer review process. *Behavioral and Brain Sciences, 5*, 210–211.
Garcia, J. (1981). Tilting at the papermills of academe. *American Psychologist, 36*, 149–158.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178,* 471–479.
Garvey, W.D., and Griffith, B.C. (1971). Scientific communication. *American Psychologist, 26,* 349–362.
Garvey, W.D., Lin, N., and Nelson, C.E. (1970). Communication in the physical and social sciences. *Science, 170,* 1166–1173.
Glenn, N.D. (1976). The journal article review process. *American Sociologist, 11,* 179–185.
Goodstein, L.D., and Brazis, K.L. (1970). Psychology of the scientist. *Psychological Reports, 27,* 835–838.
Gordon, M. (1978). *A study of the evaluation of research by primary journals in the UK.* London: Primary Communications Research Center.
Gottfredson, S.D. (1978). Evaluating psychology research reports: Dimensions, reliability and correlates of quality judgments. *American Psychologist, 33,* 920–934.
Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.
Gustin, B.H. (1973). Charisma, recognition and the motivation of scientists. *American Journal of Sociology, 78,* 1119–1134.
Hedges, L.V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42,* 443–455.
Koshland, D.E. (1987). Fraud in science. *Science, 235,* 141.
Kuhn, T.S. (1977). *The essential tension.* Chicago: University of Chicago Press.
Kunda, Z., and Nisbett, R.E. (1986). The psychometrics of everyday life. *Cognitive Psychology, 18,* 195–224.
Kupfersmid, J. (1988). Improving what is published. *American Psychologist, 43,* 635–642.
Latham, G.P., Erez, M., and Locke, E.A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists. *Journal of Applied Psychology, 73,* 753–772.
Levine, M. (1974). Scientific method and the adversary model. *American Psychologist, 29,* 661–677.
Lindzey, D. (1977). Participation and influence in publication review proceedings. *American Psychologist, 32,* 379–586.
Lindzey, D. (1978). *The scientific publication system in social science.* San Francisco: Jossey–Bass.
Mahoney, M.J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1,* 161–175.
Mahoney, M.J. (1985). Open exchange and the epistemic process. *American Psychologist, 40,* 29–39.
Mahoney, M.J. (1987). Scientific publication and knowledge politics. *Journal of Social Behavior and Personality, 2,* 165–176.
Mahoney, M.J., Kazdin, A.E., and Kenigsberg, M. (1978). Getting published. *Cognitive Therapy and Research, 2,* 69–70.
Marsh, H.W., and Ball, S. (1981). Interjudgmental reliability of reviews for the *Journal of Educational Psychology. Journal of Educational Psychology, 73,* 872–880.
Meehl, P.E. (1973). *Psychodiagnosis.* New York: Norton.
Meehl, P.E. (1978). Theoretical risks and tabular asterisks. *Journal of Consulting and Clinical Psychology, 46,* 806–834.
Merton, R.K. (1973). *The sociology of science.* Chicago: University of Chicago Press.
Millimet, C.R. (1981). Toward a reformulation of editorial policy. *Journal of Mind and Behavior, 2,* 57–64.
Mitroff, I.I. (1982). Designing peer review for the subjective as well as the objective side of science. *Behavioral and Brain Sciences, 5,* 227–228.
Munley, P.H., Sharkin, B., and Gelso, C.J. (1988). Reviewer ratings and agreement on manuscripts reviewed for the *Journal of Counseling Psychology. Journal of Counseling Psychology, 35,* 198–202.
Nisbett, R.E., and Wilson, T.D. (1977a). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 7,* 231–259.
Nisbett, R., and Wilson, T.D. (1977b). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35,* 250–256.

Palermo, D.S. (1982). Biases, decisions and auctorial rebuttal in the peer review process. *Behavioral and Brain Sciences, 5*, 230–231.

Peters, D.P., and Ceci, S.J. (1982). Peer review practices of psychology journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences, 5*, 187–205.

Pfeffer, J., Leong, A., and Strehl, K. (1977). Paradigm development and particularism. *Social Forces, 55*, 938–951.

Pondy, L.R. (1985). The reviewer as defense attorney. In L.L. Cummings and P.J. Frost (Eds.), *Publishing in the organizational sciences* (pp. 210–219). Homewood, Illinois: Irwin.

Popper, K.R. (1972). *Conjectures and refutations*. New York: Harper.

Prescott, S., and Csikszentmihalyi, M. (1977). Institutional status and publication rates in professional journals. *Quarterly Journal of Ideology, 1*, 30–37.

Rosenblatt, A., and Kirk, S.A. (1980). Recognition of authors in blind review of manuscripts. *Journal of Social Service Research, 3*, 383–394.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Scarr, S., and Weber, B.L.R. (1978). The reliability of reviews for the *American Psychologist*. *American Psychologist, 33*, 935.

Scott. W.A. (1974). Interreviewer agreement on some characteristics of manuscripts submitted to *Journal of Personality and Social Psychology*. *American Psychologist, 29*, 698–702.

Skinner, B. F. (1987). Whatever happened to psychology as the science of behavior? *American Psychologist, 42*, 780–786.

Smart, R. (1964). The importance of negative results in psychological research. *Canadian Psychologist, 5*, 225–232.

Snizek, W.E., and Fuhrman, E.R. (1979). Some factors affecting the evaluative content of book reviews in sociology. *American Sociologist, 14*, 108–114.

Sterling, T. (1970). Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. In D. Morrison and R. Henkel (Eds.), *The significance test controversy* (pp. 295–300). Chicago: Alding.

Summary Report of Journal Operations: 1989. (1990). *American Psychologist, 45*, 884.

Walster, G.W., and Cleary, T.A. (1970). A proposal for a new editorial policy in the social sciences. *American Statistician, 24*, 16–19.

Watkins, M.W. (1979). Chance and interrater agreement on manuscripts. *American Psychologist, 34*, 796–798.

Whitehurst, G.J. (1983). Interrater agreement for reviews for *Developmental Review*. *Developmental Review, 3*, 73–78.

Whitehurst, G.J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39*, 22–28.

Whitehurst, G.J. (1985). On lies, damned lies and statistics. *American Psychologist, 40*, 568–569.

Wilson, W.A. (1982). Research on peer review practices. *Behavioral and Brain Sciences, 5*, 242–243.

Wolff, W.M. (1970). A study of criteria for journal manuscripts. *American Psychologist, 25*, 636–639.

Yoels, W.L. (1974). The structure of scientific fields and the allocation of editorships in scientific journals. *Sociological Quarterly, 15*, 264–276.

Zuckerman, H., and Merton, R.K. (1971). Patterns of evaluation in science. *Minerva, 9*, 66–100.