Improving What Is Published

A Model in Search of an Editor

Joel Kupfersmid *PSI Associates, Inc.*

ABSTRACT: Researchers and practitioners of psychology have expressed considerable dissatisfaction with much of what is published in professional journals. Three major areas of discontent are as follows: (a) Many articles focus on irrelevant topics; (b) the use of statistical significance testing often results in meaningless or unusable findings; and (c) the decision-making process for manuscript acceptance/rejection may be biased. Each of these issues is discussed, and an alternative model for manuscript submission is proposed. The advantages and limitations of this model are presented as related to the three areas of current dissatisfaction.

Much dissatisfaction has been expressed regarding the quality of published articles and the decision-making process in manuscript acceptance and rejection. The purpose of this article is to review three major areas of discontent and to propose a model of manuscript submission and review that may greatly mitigate the current malaise.[1] The three major concerns regarding publication practice to be discussed are relevancy, meaningfulness, and bias.

The Issue of Relevancy

Lindsey (1977) asked, "How is it that so much triviality, illiteracy, and dullness is yearly entered into the scientific publication stream?" (p. 579). Richard Nisbett's (1978) recommendations to psychologists provide a partial answer to Lindsey's question. Nisbett advised researchers interested in increasing the chances of publishing their empirical investigations to avoid creative or innovative experimental designs and to concentrate efforts on areas that are easy to test and are noncontroversial.

Unobtrusive, circumstantial data suggest that disenchantment exists throughout the profession of psychology with what appears in journal articles. A survey conducted by Garvey and Griffith (1971) indicated that for any given study, only about 200 psychologists will read its contents within the first 60 days it appears in print. In the absence of comparable data from other scientific disciplines, it is difficult to determine if this finding indicates that psychologists' interest in reading research is representative of most scientists, or if psychologists find few studies worthy of attention. Garfield's (1972) data relate to this issue. He formulated an "impact factor" (i.e., average citation noted per published article, after correction for the number of articles a journal publishes yearly) for the 152 most frequently cited journals in science and technology. The two most often cited journals in psychology are *Psychological Review* and *Psychological Bulletin* (ranked 35th and 50th, respectively). However, the *Psychological Review* mainly contains theoretical articles, and *Psychological Bulletin* consists primarily of reviews of research (Markle & Rinn, 1977). The most frequently cited psychological journal of an experimental nature (ranked 117th), is the *Journal of Experimental Analysis of Behavior* (Markle & Rinn, 1977). This journal's impact factor is 2.3, suggesting that even articles from the most often cited experimental journal in psychology are often viewed as inconsequential and unlikely to be referenced in articles published in other journals. Garfield's (1972) results have led one reviewer (of this article) to conclude, "One would like to think that if scientists were content with what is currently being published, they would pay more attention to it (cite it) when they write their own papers." (Anonymous personal communication, September 11, 1987).

The above data suggest that much of psychology does not advance by an accumulative progression of empirically verifiable facts, but rather, that many investigators conduct isolated studies that rarely aid colleagues in this effort to understand psychological phenomena.

Results of surveys exploring practitioners' concerns with publication practices have yielded uniform results over the years. Practitioners report that psychotherapy research has little value for clinical application. When psychologists are requested to rank order the usefulness of informational sources to their practice, research articles and books of empirical research are consistently rated at the bottom on the scale (Cohen, 1979; Cohen, Sargent, & Sechrest, 1986; Morrow-Bradley & Elliot, 1986). Morrow-Bradley and Elliot's (1986) review of 18 references concludes, "With virtual unanimity, psychotherapy researchers have argued that (a) psychotherapy research should yield information useful to practicing therapists, (b) such research to date has not done so, and (c) this problem should be remedied" (p. 188). The end result is that many forms of therapy are adopted before data dem-

Correspondence concerning this article should be addressed to Joel Kupfersmid, 1560 Callander, Hudson, OH 44236.

[1] The three areas of dissatisfaction are not intended to be an exhaustive list. Rather, these issues represent areas where dissatisfaction can be reduced if the proposed model is instituted.

Copyright 1988 by the American Psychological Association, Inc. 0003-066X/88/$00.75
Vol. 43, No. 8, 635-642

onstrating effectiveness are available (Barlow, Hayes, & Nelson, 1984).

When survey respondents are asked to list the specific aspects of psychotherapy research that contribute most to their dissatisfaction, the relevancy of topics addressed, the manner in which hypotheses are tested, and the inappropriateness of statistical significance testing emerge as prime areas of discontent.

No attitude survey was found that directly asked researchers and academicians their opinion concerning the relevancy of published studies in psychology. Given the above information regarding the opinion and behaviors of psychologists in general and the attitudes expressed by practitioners, it would not be surprising if experimentalists shared many of the same concerns.

The Concern for Meaningfulness

The paradigm for most experimental and correlational studies consists of postulating no difference (null hypothesis) between groups or variables, a hypothesis that the researcher then attempts to refute. Statistical significance testing is the method most experimenters adopt to determine whether the null hypothesis may be confidently rejected or retained. Significance testing involves selecting a level of probability (p value) to determine "how improbable an event could be under the null hypothesis" (Bakan, 1966, p. 429). Usually a probability of 5% or less ($p < .05$) is selected. "Thus the p value may be used to make a decision about accepting or rejecting the idea that chance caused the results. This is what statistical significance testing is—nothing more, nothing less" (Carver, 1978, p. 387).

The first problem is that the no difference (null) hypothesis is never capable of being retained; that is, the null hypothesis is always false to begin with (Bakan, 1966; Greenwald, 1975; Meehl, 1978). "If by the null hypothesis one refers to the hypothesis of *exactly* no difference or *exactly* no correlation, and so forth, then the initial probability of the null hypothesis being true must be regarded effectively as zero" (Greenwald, 1975, p. 6). Lykken (1968/1970) added that it is "foolish" even to suppose that the difference between two groups, or the correlation between two variables, is ever zero. Researchers can confidently make this claim because no two groups of subjects or variables are ever effectively equal; rather, the null hypothesis will always be rejected *if* the experimenter has a large enough sample N.

Meehl (1978) concluded that "reliance on merely refuting the null hypothesis . . . is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (p. 817).

A second problem related to the meaningfulness of "statistically significant" findings is that what is significant statistically and what is "significant" in a meaningful sense may be contradictory. The question is *never* whether the result is statistically significant, but rather, at what N the data would reach statistical significance.

Clinicians have expressed dissatisfaction with the use of statistical significance testing because the procedure bears no relationship to the size effect (i.e., degree of difference between two samples and/or degree of correlation; Smith, Glass, & Miller, 1980). The word "significance" is used even for trivial findings. Furthermore, statistical significance is a measure of group effects, whereas the practitioner is concerned with individuals (Barlow et al., 1984).

Fifty-two percent of psychotherapists surveyed (Morrow-Bradley & Elliot, 1986) criticized the use of statistical significance testing because the information does not address the important question of how many subjects changed or to what degree. Quoting Bergin and Strupp (1972, p. 440), Barlow et al. (1984) noted,

> In the area of psychotherapy, the kinds of effects we need to demonstrate . . . should be significant enough so that they are readily observable by inspection or descriptive statistics. If this cannot be done, no fixation upon statistical and mathematical niceties will generate fruitful insights. (p. 28)

Furthermore, in all of experimentation, it is critical to collect data that are relevant to psychological inquiry such that, regardless of outcome, via statistical manipulation, meaningful results are generated.

Jacobson, Follette, and Revenstorf (1984) directly addressed this issue by eschewing traditional statistical significance tests for psychotherapy outcome research and argued in favor of a "clinical significance" test. The authors provided a summary of five such measures as well as presenting their own approach: "Therefore, we propose that a change in therapy is clinically significant when the client moves from the dysfunctional to the functional range during the course of therapy on whatever variable is being used to measure the clinical problem" (p. 340).

The problem associated with generalization of findings to clients seen in a clinician's office has also received critical comment. Statistical inference is contingent on the criterion of random sampling techniques, and rarely, if ever, is this criterion satisfied in psychotherapy studies (Bakan, 1966; Barlow et al., 1984). Statistical inference is tied to sampling theory. A sample employed in research must represent, or closely approximate, individuals seen in clinical practice. Information that cannot be inferred or generalized beyond the sample of a particular study is not of any value to the practitioner. Thus, even if random samples are obtainable, the question usually remains whether the sample would be too heterogeneous or too homogeneous to generalize to specific clients (Barlow et al., 1984).

Another area of concern relates to the decision theory of acceptance/rejection of the null hypothesis based on statistical significance testing. The current use of statistical significance testing limits the outcome of hypothesis testing to two choices: Retain the null ($p > .05$) or reject the null ($p < .05$).

> What scientist in his [sic] right mind would ever feel there to be an appreciable difference between the interpretative significance of data, say, for which one-tailed $p = .04$ and that of data for which $p = .06$, even though the point of "significance" has been set at $p = .05$? (Rozeboom, 1960, p. 424)

Carver (1978) summarized the current dissatisfaction with hypothesis testing and the test of statistical significance as follows:

> If we can control statistical significance simply by changing sample size, if statistical significance is not equivalent to scientific significance, if statistical significance testing corrupts the scientific method, and if it has only questionable relevance to one out of fifteen threats to research validity, then I believe we should eliminate statistical significance testing in our research. (p. 392)

If this were not enough, the use of statistical significance testing and associated p values are often misinterpreted. The most noticeable misinterpretations are (a) the p value reflects the probability that the results are due to chance, (b) p values represent the probability of obtaining the same results upon experimental replication, and (c) the p value reflects the probability that the research hypothesis is true (Bakan, 1966; Carver, 1978).

Problems with misinterpretation of p values are reflected in the results of four studies involving psychologists, journal editors and reviewers, professors, and graduate students. The paradigm of all studies involved asking participants to rate their degree of belief or confidence in the results of hypothetical studies given varying p values (from .001 through .90) with either a small sample size (10 subjects) or a larger sample size (100–200 subjects). Across all conditions, participants placed greater confidence in hypothetical studies having larger numbers of subjects even if the p values for both samples were *exactly* the same. Psychologists ignored the fact that the mathematics of significance testing takes into account the size of the sample. They failed to realize that small samples require a greater disparity between groups in order to reach the same p value as studies utilizing a larger number of subjects (Bakan, 1966; Carver, 1978).[2]

In spite of the plethora of rational arguments against its use, the current method of hypothesis testing by statistical significance test continues. In spite of the profuse dissatisfaction with the end product of such research—irrelevant studies and meaningless results—the system flourishes. Why? There are three factors that seem to interact to maintain the current practice.

First, as previously discussed, many psychologists continue to misinterpret statistical significance tests and associated p values. Too often, experimenters believe that p values express confirmation of the experimental hypothesis and that p values represent a measure of confidence in the repeatability of the experiment. Additionally, many psychologists seem unaware (or deny) the ease with which sample size affects the rejection of the null hypothesis and the establishment of statistical significance. Many researchers believe they are "discovering" something when, in fact, they are not.

Second, there is a bias among editors and reviewers for publishing *almost exclusively* studies that reject the null hypothesis via statistical significance testing. Because careers and reputations are often associated with publication, there will be no change in the nature of what is published (regardless of relevancy or meaningfulness) until the decision-making practice of journal editors is altered. In the following section of this article, the bias that editors and reviewers exhibit with respect to statistical significance testing will be addressed.

Third, Kuhn's (1970) position on the nature of change in the history of scientific revolutions seems to be operative. Kuhn noted that current paradigms often persist, regardless of inadequacy, until there are alternative paradigms that can take their place. It is my intention in this article to propose an alternative to manuscript submission that, if employed, may offer a substantial improvement over the current system.

The Editorial Bias Controversy

The three areas of editorial bias discussed below have received considerable attention in the literature. If any of the three charges have merit, there is an additional argument that the current method of manuscript decision making be reconsidered in favor of an alternative model.

The first contention of bias is that only those manuscripts that report rejection of the null hypothesis by use of statistical significance testing get published. This bias is the most dangerous of all because it would mean that the data bank of psychological knowledge is filled with Type I error (rejection of a true null hypothesis). Type I error is more serious than Type II error (rejection of a true research hypothesis) because when a Type I error appears in print it often stops researchers from studying the phenomena and/or reporting nonsignificant results (Bakan, 1966). Rosenthal (1979) termed this the "file drawer problem" in that "the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with 95% of the studies that show nonsignificant (e.g., $p > .05$) results" (p. 638).

However, do findings of nonsignificant difference actually lead researchers to "file" their studies? Do the findings of significant differences encourage researchers to submit their results? Do journals tend to publish only experiments in which statistical significance establishes rejection of the null hypothesis? The current data suggest the answers are all affirmative.

Greenwald (1975) presented evidence from a survey in which the authors noted that there was a 50% chance that they would submit a manuscript if the null hypothesis was rejected and a 6% chance of submission if the null hypothesis was retained. Similarly, Sterling (1959/1970) reviewed the number of articles published in four psychology journals in which the null hypothesis was retained. Of the 362 articles, 8 retained the null hypothesis, and *none* of the studies replicated previous experiments. Sterling also selected 100 research titles at random from *Psychological Abstracts* and found that 95 articles rejected the null hypothesis, 5 failed to reject the null hypothesis,

[2] Critics of statistical significance testing advocate a variety of statistical alternatives, including Bayes' theorem (Bakan, 1966; Greenwald, 1975); decision-theory of Neyman, Pearson, and Wold (Bakan, 1966); omega squared (Carver, 1978); eta squared (Carver, 1978); interval estimation (Greenwald, 1975); and greater use of descriptive statistics (Carver, 1978).

and 1 was a replication study. A more recent study conducted by Greenwald (1975) on all articles published (N = 199) by the *Journal of Personality and Social Psychology* revealed that only 12% of the articles retained the null hypothesis.[3]

The outcry regarding editorial bias in favor of null hypothesis testing via statistical significance testing has been heated. A few representative quotes will demonstrate the intensity of concern:

> The use of statistical tests of significance are not likely to decline until one or more journal editors speak against statistical significance testing. . . . (Carver, 1978, p. 397)

> The stranglehold that conventional null-hypothesis significance testing has clamped on publication standards must be broken. (Rozeboom, 1960, p. 428)

> If one could no longer use statistical significance to determine the "significance" of a difference, researchers would be forced to use designs that more clearly reveal the scientific importance of a difference. (Carver, 1978, p. 397)

> When passing null hypothesis tests becomes the criterion . . . for journal publications, there is no pressure on the psychology researcher to build a solid, accurate theory; all he or she is required to do, it seems, is produce "statistically significant" results. (Dar, 1987, p. 149)

> The moral of this story is that the finding of statistical significance is perhaps the least important attribute of a good experiment: It is *never* a sufficient condition . . . that an experimental report ought to be published. (Lykken, 1968/1970, p. 278)

> *Support for the null hypothesis must be regarded as a research outcome that is as acceptable as any other.* (Greenwald, 1975, p. 16)

There is little evidence that these concerns have resulted in a change of publication decision making. The American Psychological Association (APA)'s *Publication Manual* (1983) lists "reporting of negative results" as a major "defect" editors find in papers submitted.

The second form of alleged bias suggests that manuscripts are published on the basis of the submitter's status in the field and/or prestige of author's institutional affiliation. Implicit in this form of bias is the belief that editors and reviewers are either incapable of discriminating among manuscripts in order to choose those that are truly exceptional in relevance and methodological vigor or there are so few exceptional studies that professional status becomes the informal, unspoken criterion for manuscript decision making.

Those who charge such bias maintain that removing the title page from a manuscript before review is inadequate because (a) authors frequently refer to their previous work in the text, (b) many experimenters have a unique style of conducting research that others in the field can easily recognize, and (c) there is often a small network of researchers in a given field and manuscripts are shared among these individuals. Thus, it is highly probable that a reviewer would be familiar with a submitter's work (Ceci & Peters, 1984).

The belief that reviewers are able to identify the author of a manuscript is widespread. Two studies report that over 70% of authors believe that reviewers are able to correctly identify authorship even though the title page is removed prior to manuscript review (Ceci & Peters, 1984).

Two studies, employing manuscripts from seven psychology journals, focus on the issue of authorship detection. In both studies reviewers correctly identified at least one author in approximately 25% of the manuscripts submitted (Ceci & Peters, 1984).

The key question, however, is whether manuscripts are differentially accepted or rejected on the basis of author detection (Surwillo, 1986). Two studies address this issue. Mahoney, Kazdin, and Kenigsberg (cited in Presser, 1982) reported that institutional prestige of authors had no effect on reviewer's judgment of manuscripts submitted to "behavioristic journals." Peters and Ceci (1982) submitted for editorial review 12 *published* articles to the same journals that had published them, but changed the author's name and prestige of institutional affiliation. Three of the articles were detected. Of the remaining nine, eight manuscripts were rejected for publication on methodological grounds. This study has led to a heated debate on blind review. The entire issue of *The Behavior and Brain Sciences* (Harnad, 1982) is devoted to a discussion of this topic.

The third charge of bias is, perhaps, the most insidious. This form of bias suggests that chance factors determine which articles are accepted or rejected for publication. The data on this issue involve studies that assess agreement between two or more reviewers regarding manuscript decision making.

Many journals have investigated the agreement be-

[3] Rosenthal (1979) proposed a formula for estimating the number of studies in the file drawers (or those that needed to be published in the future) that retain the null hypothesis, based on current numbers of studies in print that report statistically significant findings. Essentially, Rosenthal's formula involves (a) transforming into Z scores the p values reported in each study, (b) computing the mean Z score for these studies, and (c) multiplying this product by the number of studies gathered. Rosenthal noted, however, that only six studies in the file drawer that support the null hypothesis are necessary when there are as many as 15 studies published that report statistically significant findings. Rosenthal's formula is best suited for areas where large numbers of experiments have been conducted.

Another caveat regarding this formula is that the p value is partially increased or decreased contingent on the sample N. This formula does not solve the problem of knowing that statistical significance is a likely outcome before the study is initiated, when large numbers of subjects are employed. Additionally, methodologically sound studies with smaller Ns are more likely to have larger p values than methodologically questionable studies involving hundreds of subjects. Rosenthal's formula would potentially give greater weight to the latter type of studies in spite of the fact that the former type of studies require a greater difference between means to produce statistically significant results. In an earlier work, Rosenthal (1978) provided a summary of nine other methods that may be used to combine results of independent studies.

Rosenthal's formula has some merit in reducing the concern about Type I error when large numbers of studies are available in a given area. Unfortunately, this procedure does not reduce concerns about issues related to research relevancy, the meaningfulness of the experimental results, misinterpretations of the meaning of statistical significance, or editorial bias.

tween judges on manuscript submissions. The data on interrater agreement have indicated that there is little agreement among reviewers regarding the worthiness of a manuscript for publication. Much debate has centered on what statistic to employ to measure interrater agreement. The intraclass coefficient has been the most often utilized measure (Marsh & Ball, 1981). Arguments have been advanced for using Kappa (Watkins, 1979) or Finn's *r* (Whitehurst, 1984) because these statistics measure the *proportion* of agreement after removal of chance.

Interclass correlations have been reported for 13 studies involving eight journals (Marsh & Ball, 1981; Whitehurst, 1982). Interclass correlations usually range between .20 and .40. Reanalyzing interrater agreement by the use of a Kappa or Finn *r* for some of the same journals in which intraclass correlations were performed has not produced dramatic changes in outcome (Watkins, 1979; Whitehurst, 1984). With the exception of the *American Psychologist,* most studies report an unimpressive degree of agreement between reviewers regarding the value of a manuscript, regardless of the statistic employed.

Gottfredson (1978) investigated qualitative agreement between editors and reviewers for nine psychology journals on characteristics that make for a publishable article. Factor analytic results suggest that there is agreement on what characteristics lead to rejecting a manuscript for publication (i.e., a "list of don'ts"). However, agreement between reviewers on what characteristics make for a quality article range from .35 to .41. Similarly, the correlation between the number of citations a published article received over a nine-year period with that of a reviewer's judgment of manuscript quality or impact ranged from .24 to .37. Gottfredson's data suggest that editors and reviewers can agree qualitatively on characteristics that make a manuscript unacceptable (list of don'ts), but there is only a modest consensus regarding the quality or importance of submissions.

The preceding section addressed agreement between reviewers regarding the recommendations to accept or reject a manuscript for publication. The final decision to publish is with the editor-in-chief. Lindsey (1977) explored characteristics of reviewers' influence, or "editorial power," on an editor's decision to publish manuscripts. Editorial power is defined as the percentage of manuscripts a reviewer recommends for publication that are, in fact, published. The number of articles a reviewer has published correlated positively with his or her editorial power (beta weight = .406), but a reviewer's editorial power is negatively associated with the number of citations a reviewer's articles receive (beta weight = −.456). Lindsey suggested that these paradoxical results may be attributed to overcriticalness and likelihood of manuscript rejection by judges deemed eminent (i.e., those who have published articles receiving many citations).

What can be concluded about the allegations of editorial bias and the current data bearing on the issue?

With respect to the "file drawer problem," available data suggest that this bias exists. Researchers are more likely to submit manuscripts in which the null hypothesis is rejected via statistical significance testing. Likewise, editors are much more willing to publish articles rejecting the null hypothesis by means of statistical significance testing. In spite of the questionable relevance of this form of "discovery" and the questionable logic of the approach, manuscript decision makers seem unable to escape this response set.

The second contention of bias relates to the "blindness" of anonymous reviews. Available data suggest that reviewers are able to correctly identify authorship of approximately one fourth of the manuscripts submitted. However, there are little consistent data on whether identification of authorship, status of the author, or prestige of the author's institutional affiliation affects publication decision making.

The third area of bias involves agreement between reviewers regarding quality of manuscript submission versus the influence of chance in publication practice. Measurement of interjudge agreement seems to be as controversial as the charge of bias itself. Although there seems to be satisfactory qualitative agreement as to what constitutes a manuscript of poor quality (a list of don'ts), there is little consistency among reviewers as to what constitutes good quality.

Marsh and Ball's (1981) conclusion of the literature on interrater reliability seems appropriate as the conclusion for the three areas of bias reviewed: "It seems ironic that [the] scientific method has scarcely been used to determine how best to evaluate the products of scientific research" (p. 880).

An Alternative Model

The growing dissatisfaction expressed by practitioners and researchers with respect to psychology's data bank of published knowledge has been summarized. Many contend the situation will not change *unless* editors and reviewers alter the manner in which manuscripts are judged to be acceptable for print. Strangely, editors and reviewers have published many articles critical of publication decision making, as evidenced by the sheer number of references in this work as well as the publication of this article. Yet, these same individuals seem unwilling to change their practice.

As previously noted, practices entrenched in the sciences do not perish on the basis of convincing rational arguments or empirical evidence. Rather, a practice is most likely to be displaced only if there is an alternative method to take its place. It is the purpose of this article to present an alternative model for manuscript submission that may mitigate some of the current dissatisfaction. No claim is made that this model will dissipate all areas of dissatisfaction, but I believe that this proposal is (a) a "significant" improvement over the current system, (b) simple in its design, (c) easy to implement, and (d) requires only a small modification in the current publication practice.

The proposed model is designed primarily for manuscript submission of experimental and quasi-ex-

perimental studies. Manuscripts discussing special issues, theoretical controversies, and reviews of literature would not fit into the model's structure.

It is proposed that experimental and quasi-experimental manuscripts be written in the same form as currently outlined in the *APA Publication Manual* (third edition, 1983) with the exception that the Results and Discussion sections are not submitted. Once the manuscript is accepted for publication, then the Results and Discussion sections are forwarded to the editor. Instead of submitting the results, it is suggested that authors submit a Results section that outlines what statistical procedures will be employed to analyze each hypothesis.

The argument being made is that if a study has focused on a relevant topic, and if the experimenter has provided a sound rationale, used satisfactory sample selection with methodological and procedural vigor, and proposed appropriate statistical analyses, then the results of the study are apt to be informative regardless of outcome. It is also suggested that anonymous review continue as is.

The implementation of this model offers several advantages over the current system as follows:

1. The number of pages that need to be reviewed in order for an editor to make an accept–reject decision would be reduced because the Results and Discussion sections are not submitted. The turnaround time from the point of submission to author's notification may also be reduced.

2. Authors would save time. Experimenters would not have to analyze data until informed of publication acceptance. Studies rejected could be corrected or discarded without time spent in data analysis or writing a Discussion section.

3. Experimenters could know the fate of the research before data analysis. If the manuscript is accepted, they would not feel pressured to analyze the data in every conceivable way to produce results that appear "significant."

4. Studies would be accepted on the basis of topic relevance and methodological soundness. This would reduce the number of irrelevant and procedurally flawed studies in print. Editors and reviewers would not be biased by the results of statistical significance testing: This learning set could be broken.

5. Because editors and reviewers would not know the results of the study, the chance of authors submitting results, and of articles being published, that retain the null hypothesis would be increased. The "file drawer"/Type I error problem would be reduced.

6. An experimenter, as well as editors and reviewers, would be less likely to be "locked in" to the statistical significance test of a null hypothesis. Editors and reviewers critiquing studies utilizing large numbers of subjects would know that finding statistical significance is likely and rejection of the null hypothesis is imminent. Editors may be more likely to accept for publication manuscripts that present more meaningful methods of data analysis. Likewise, reviewers may be more likely to recommend that acceptance for publication be contingent on the author's analyzing results in the manner the reviewer specifically requests.

7. Given the pressure in some circles to publish, there would be less likelihood of researchers "manufacturing" results (sometimes fraudulently).

8. The effects that publishing only statistically significant results have had in terms of stifling replication studies would be reduced.

9. Because the manuscript would be accepted prior to the writing of the Discussion section, there would be no need for authors to engage in lengthy prose about the relevancy of the research to all areas of psychology. The relevancy of the study would be made clear in the Introductory section. The pressure to "explain" negative findings may be less, thus reducing the length of the Discussion section.

10. Agreement among reviewers regarding acceptance or rejection might be increased because decision making would be based on the study's relevancy and methodological appropriateness. This would require that investigations compare interrater agreement between the current method and the proposed model.

There are potential limitations to this model. Two of these, with possible resolutions, are discussed here.

1. Accidental discoveries would not be apt to get published by the proposed model because the Results section would not be submitted.

If serendipity had occurred and was the major contribution of the study, then the experimenter could conduct a second study. As part of the rationale for the second study, reference would be made to the experimenter's first study. It is a waste of editorial time and journal space to publish the first study, which contains much irrelevant information and findings, when the real contribution is an accidental discovery. By reinvestigating the serendipitous finding directly, a much stronger case could be made for its importance to the field.[4]

[4] Occasionally, serendipitous discovery is made by the reviewer and not the author. It may be argued that without the initial submission of the Results section, the potential for such discovery would be compromised. However, under both the current system of manuscript submission and the proposed system, authors ordinarily only submit a manuscript if there is some potentially useful finding. In both systems, the author's finding (and any further informative results that may be detected by a reviewer) would only be considered meaningful if the study maintained methodological soundness. A reviewer's rejection of the manuscript, prior to the submission of the Results section, implies a questionable rationale and/or substantial weakness in experimental design. Any accidental "discovery" by the reviewer under this condition would be scientifically suspect.

In those rare circumstances in which a study is accepted for publication under the proposed model and a reviewer finds a serendipitous discovery on submission of the Results section, the reviewer can inform the author of this discovery. The manuscript can be returned to the author to either (a) comment on this discovery in the Discussion section, (b) re-analyze data in light of the discovery, or (c) conduct a new study explicitly focusing on this discovery. When a reviewer makes a serendipitous discovery, whether under the current method or the proposed model, a return of the entire manuscript to the author is necessary. It is recognized that the proposed model is more cumbersome than the

2. Because data are not required on submission, experimenters may submit "proposals" and not an experiment that has been implemented. Submitters may believe that *if* the manuscript is accepted, then they will conduct the study.

If a manuscript were accepted, or accepted contingent on revision, then the acceptance would be operative for a specific time interval. If the author failed to submit the entire manuscript (with Results and Discussion section) by a specified date, manuscript acceptance would be revoked.

Four other limitations to the model are noted, with comments regarding the model's potential to address each concern:

1. Because acceptance for publication is not contingent on experimental outcome, there may be a flood of manuscript submissions as well as many studies which discover "nothing."

It is possible that the initial introduction of this model will result in an increase of submissions, but only those studies that provide a satisfactory rationale and design will be found acceptable for print. Thus, the feedback loop to authors will change. Instead of researchers' focusing on the statistical significance of the results, as is the current condition, the focus will be on topic relevance and methodology. It is anticipated that the initial flood of submissions would plateau and further manuscript submissions utilizing the proposed model would be similar to the frequency of manuscripts currently submitted. This speculation regarding submission frequency may be easily studied if the proposed model were to be adopted.

2. The proposed model is not applicable to manuscripts focusing on review of literature, survey studies, theoretical articles, and so on.

This model is certainly not a panacea, but it may potentially reduce much of the current dissatisfaction with experimental and quasi-experimental research. This, in turn, may result in generating more consistency across studies exploring the same area. The hackneyed conclusion "that more research is needed" in review of literature articles may be reduced.

3. The proposed model does not address alleged bias of manuscript acceptance on the basis of author's status, prestige of author's institutional affiliation, or reviewer's ability to correctly identify manuscript authorship.

The proposed model does not reduce this form of potential bias. However, current data suggest that correct identification of authorship occurs in approximately 25% of the submissions, and it is not clear whether identification of authorship, author status and/or institutional affiliation biases publication acceptance and rejection.

4. This proposal would not eliminate the use of statistical significance in hypothesis testing.

This criticism is valid, but the model has a potential for researchers, editors, and reviewers to explore other avenues of data analysis that appear more meaningful. Because editors and reviewers would not know the experimental outcome, they might be more likely to insist on analyses that are most apt to provide a meaningful interpretation.

current system in this situation, that is, a serendipitous discovery cannot be detected by a reviewer until the Results section is submitted. The issue is whether a reviewer's detection of an accidental discovery occurs frequently enough to offset the other advantages of the proposed model.

Summary and Conclusion

Many psychologists have argued that much of the published research focuses on irrelevant issues, that statistical significance testing provides meaningless data, and the publication decision-making process is fraught with bias. Journal editors, by nature of their gate-keeping function, strongly influence what appears in print and set (or reinforce) the standards of the scientific enterprise. When the decision-making practice of manuscript acceptance and rejection is changed, the dissatisfaction currently expressed may be rectified.

The history of science suggests that current practices are not replaced on the basis of rational arguments or empirical research. What is required is an alternative model that may readily and relatively painlessly replace traditional approaches, and it is toward this end that the present model is offered. The proposal suggests a model of manuscript submission that offers the possibility of greatly reducing the dissatisfaction psychologists have expressed toward the current publication decision-making system. This model has limitations, but the limitations appear outweighed by the model's advantages. It is hoped that this approach will find an editor willing to break tradition in an effort to improve psychology's method of analyzing and sharing scientific information.

REFERENCES

American Psychological Association. (1983). *Publication manual of the American Psychological Association* (3rd ed.). Washington, DC: Author.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Barlow, D., Hays, S., & Nelson, R. (1984). *The scientist practitioner: Research and accountability in clinical and educational settings.* New York: Pergamon Press.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Ceci, S., & Peters, D. (1984). How blind is blind review? *American Psychologist, 39,* 1491–1494.

Cohen, L. (1979). The research readership and information source reliance of clinical psychologists. *Professional Psychology, 10,* 780–785.

Cohen, L., Sargent, M., & Sechrest, L. (1986). Use of psychotherapy research by professional psychologists. *American Psychologist, 41,* 198–206.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 42,* 145–151.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178,* 471–479.

Garvey, W., & Griffith, B. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist, 26,* 349–362.

Gottfredson, S. (1978). Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist, 33,* 920–933.

Greenwald, A. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Harnad, S. (1982). Peer commentary on peer review [Special symposium issue]. *The Behavior and Brain Sciences, 5*(2).
Jacobson, N., Follette, W., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336–352.
Kuhn, R. (1970). *The structure of scientific revolutions.* Chicago: University of Chicago Press.
Lindsey, D. (1977). Participation and influence in publication review proceedings: A reply. *American Psychologist, 32,* 579–586.
Lykken, D. (1970). Statistical significance in psychological research. In D. Morrison & R. Henkel (Eds.), *The significance test controversy* (pp. 267–279). Chicago: Aldine. (Original work published 1968)
Markle, H., & Rinn, R. (1977). *Author's guide to journals in psychology, psychiatry, and social work.* New York: Haworth Press.
Marsh, H., & Ball, S. (1981). Interjudgmental reliability of reviews for the *Journal of Educational Psychology. Journal of Educational Psychology, 73,* 872–880.
Meehl, P. (1978). Theoretical risk and tabular asterisk: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.
Morrow-Bradley, C., & Elliot, R. (1986). Utilization of psychotherapy research by practicing psychotherapists. *American Psychologist, 41,* 188–197.
Nisbett, R. (1978). A guide for reviewers: Editorial hardball and the '70s. *American Psychologist, 33,* 519–520.
Peters, D., & Ceci, S. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *The Behavioral and Brain Sciences, 5,* 187–195.
Presser, S. (1982). Reviewer reliability: Confusing random error with systematic error or bias. *The Behavioral and Brain Sciences, 5,* 234–236.
Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185–193.
Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.
Rozeboom, W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428.
Smith, M., Glass, G., & Miller, T. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.
Sterling, T. (1970). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. In D. Morrison & R. Henkel (Eds.), *The significance test controversy* (pp. 295–300). Chicago: Aldine. (Original work published 1959)
Surwillo, W. (1986). Anonymous reviewing and the peer-review process. *American Psychologist, 41,* 218.
Watkins, M. (1979). Chance and interrater agreement on manuscripts. *American Psychologist, 34,* 796–798.
Whitehurst, G. (1982). The quandary of manuscript reviewing. *The Behavioral and Brain Sciences, 5,* 241–242.
Whitehurst, G. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39,* 22–28.