The Journal of Mind & Behavior
Spring, 1981, Volume 2, Number 1
ISSN 0271-0137

Toward a Reformulation of Editorial Policy

C. Raymond Millimet
University of Nebraska at Omaha

Based on the understanding that an unacceptably large number of Type I errors enter the scientific literature, it was proposed that all manuscripts submitted for publication be accompanied by an independent replication supporting the initial findings. A replication that could provide an estimate of the generality of the phenomenon would be even more desirable. Furthermore, it was recommended that the editorial board of scientific journals contract to accept a study for publication solely on the basis of its soundness and importance to the scientific community *independently* of the statistical significance of the findings. That is, to evaluate the study *prior* to data collection, thereby insuring publication to all studies judged to be acceptable regardless of the ultimate probability of the effects.

A student and I recently performed an extensive empirical investigation. An analysis of variance applied to the data indicated two statistically significant third-order interactions. Following standard statistical procedure, the appropriate simple effects analyses were performed. To our astonishment, not only were the results of these analyses inconsistent with expectation, but the pattern of the treatment effects made no sense at all.

It appeared that only a programming error could be responsible for such a creation. And yet everything seemed to be in order, as no error statement was in evidence. Witness to this was the neatly printed analysis of variance summary table, as well as cell means and cell variances, normal by-products of such an analysis. Could the computer output have materialized *sans* error statement if something had been out of phase in the programming? I knew the answer was yes, but my impatience to get on with things undermined my better judgment, so we looked elsewhere for an answer.

My next thought was to attribute the improbable findings to a Type I error—rejecting the null hypothesis when it is true. But one never knows when a Type I error is being committed, and after a while we tend to lose sight of the possibility of making them. In addition, we are conditioned to revere statistically significant findings, as publication practices demand that they be obtained if we wish to see our efforts reach the printing house. Yet it has been recognized for some time that if decision errors are being published, they are of the Type I variety (Bakan, 1966). But

Requests for reprints should be sent to C. Raymond Millimet, Ph.D., Department of Psychology, University of Nebraska at Omaha, Omaha, Nebraska 68182, U.S.A.

Type I errors should be vigorously avoided, for in addition to the false information they provide, Type I errors tend to bring research to a premature and undeserved conclusion.

We decided to take a strong look at the possibility that the results of the study were due to a Type I error. What could be wrong with attributing the problematic findings to a Type I error and let things go at that? Simply report that the findings represent nothing more than chance differences among the treatment conditions and that they should not be taken seriously. Such a course of action was tempting. But it wouldn't do. Upon posing a similar argument to my dissertation advisor a number of years ago, he advised me to interpret all significant effects as if they were just that. He said that to speculate about the presence or absence of a Type I error because the experimental effects are not in line with theory and expectation is playing a very dangerous game.

The point is that even though an investigator recognizes the likelihood of committing a Type I error, one cannot allow oneself to proclaim that a Type I error has been made. At first glance, such a comment seems to smack of Catch-22 logic and Orwellian double-think. But it really makes good sense! If an investigator reported and emphasized only those statistically significant effects that are consistent with theory and expectation, while relegating to Type I error status those statistically significant effects that are not, then one may as well not have bothered to do the experiment at all, but simply declare the theory correct on the basis of a keen insight into the matter.

The latter form of reasoning was exhibited by a member of my dissertation committee who patently refused to accept my research proposal because he didn't think people functioned in the way in which I saw them. Moreover, he said that should I find results consistent with my theoretical position, he wouldn't believe them. His view of humankind was different than mine and from such differences springs good science. However, take notice of the implicit Type I error reasoning in his argument—the inclination to deny a statistically significant effect because it is inconsistent with expectation or belief. This is not to say that empiricism is the *sine qua non* for establishing truth. But at least an empirical approach has objectivity on its side. Or should.

There seemed to be no other choice but to attack the problem directly. The typical ploy often used at this point is to not only devise some acceptable post hoc explanation of the results, but to make the after-the-fact explanation sound as if that is what would have been predicted if one had thought things through a bit better in the first place.

This was no easy exercise. Our thoughts extended over great periods of time. Slowly, we began to formulate an explanation that seemed to make sense. As we continued to talk about it, the better it sounded. In time, we

came to accept the explanation fully. And that seemed to be that. Then we found the programming error.

I won't burden you with the extended details of the error. Simply stated, the error stemmed from the data cards being out of proper alignment. As luck would have it, the data cards were isomorphic to the parameter and format cards and so the program ran. To detect this kind of error, all one need do is compute a portion of the analysis by hand. Usually a check of the cell means will suffice. If an inconsistency is noted, further examination should lead to the source of the problem. However, this is not the point I wish to make.

The disturbing thing is that my student and I had begun to make sense out of nonsense. That is to say, we had found meaning where there was little or no meaning to be found. One's search for meaning is no doubt an aspect of human endeavor of considerable importance to those prone to value an existential point of view. And perhaps our activity presents grist for the existentialist's mill. But what concerns me is not our activity and the impetus behind it. It is the result of this activity and its effect on scientific progress that troubles me.

Upon reanalyzing the data, correctly this time, the "significant" interactions disappeared. After all the turmoil, the study showed very little. Such are the misfortunes of science. But think of what could have happened. Had the incorrect analysis gone unnoticed, and had our verbal gymnastics concerning the unexpected effects been acceptable to a journal editor, the results and their discussion would have been disseminated to the world.

While it is probably true that most published research does not exert a major influence on the work of others, the potential for influence is always present and cannot be anticipated effectively. If nothing else, experimental results have a strong continuing influence on one's own research program. Unfortunately, there do not appear to be satisfactory safeguards consistent with current editorial policy for uncovering or undoing misinformation provided by empirical investigations that are the product of statistical decision errors. In too many instances science is betrayed by the .05 level of significance.

It is indeed a rare event when a published study features as its major finding the acceptance of the null hypothesis. Editorial policy does not readily welcome such an intruder. Strong evidence for this lack of respect for nonsignificant findings has been noted by Sterling (1959) who found that only 2.7% of the studies reported in four psychology journals failed to surpass the .05 level of significance. Although Greenwald (1975) has shown that the rate has risen to 12.1%, this is convincing documentation to the fact that editorial policy strongly favors studies reporting statistically significant effects. It follows that there must be a large

number of articles published each year that are based on erroneous information resulting from the false rejection of the null hypothesis. In fact, the number of Type I errors housed in our journals is undoubtedly much greater than the .05 level would lead one to expect. Indeed, the inflation rate is very likely related to the differential submission rate built into our system of publication. Because it is understood that statistically significant results are more likely to be published, statistically significant results are more readily submitted for publication (Rosenthal, 1979).

In short, it appears that an exceedingly large number of Type I errors find a home in our scientific journals as a direct result of the publication practices employed by the editorial staffs of these journals. The problem stems from the extreme reluctance of editors to accept studies limited to nonsignificant effects. Consequently, it does not take very long for the scientific investigator to become part of the scenario by becoming overly dependent on the outcome of the null hypothesis test. Such dependency is bound to result in experimental practices that tend to increase the likelihood of falsely rejecting the null hypothesis. These tactics undoubtedly would lose their appeal if the concept of statistical significance was not revered so highly.

Can the exact frequency of published studies based on chance occurrences be tallied? More importantly, is it possible to distinguish between studies characterized by statistical decision error and studies resulting in correct statistical decision? I think it is possible.

In large part, the ability to establish the frequency of published Type I errors is directly related to the ability to isolate their individual occurrence by replicating each study. Performing one or more replications is probably the best (and perhaps only) method for uncovering erroneous statistical decisions that have made their way into the research literature.

Lykken (1968) has outlined three kinds of replication: a) literal replication, where an investigator duplicates exactly his/her own experimental method; b) operational replication, where an investigator duplicates as best he/she can the sampling and experimental procedures indicated in the method section of a published article, and c) constructive replication, where a second investigation avoids an exact imitation of the first investigation's methods, but develops a different method for sampling, measurement, and data analysis. Lykken concludes that:

> Since constuctive replication has greater generality, its success strongly implying that an operational replication would have succeeded also, one should usually replicate one's own work constructively, using different sampling and measurement procedures within the purview of the same constuctive hypothesis. (p. 159)

Investigators must be given an incentive to perform and publish replications of previously published research. Unfortunately, editorial policy rarely provides the forum for such activity. In fact, Sterling (1959)

could find no evidence of a single replication within the sample of 294 studies he reviewed. And yet the dedicated investigator who doubts the veracity of published research, whether it is one's own or that of someone else, should not have to feel that the effort will go unrewarded and unnoticed should a repliction be performed. Of course, the real loser is the discipline itself. The person who said that the wonderful feature of science is that it is self-correcting probably never tried to publish a replicated study.

A compelling logistical argument against multiple replication is that space for publishing them is limited and the publication lag is already quite pronounced. A reasonable solution to this problem is to require all manuscripts submitted for publication to possess a replicated study, preferably a constructive replication, with results consistent with the initial findings. Without an accompanying replicated study establishing evidence for the reliability and generality of the initial findings, the principal investigator would be advised not to submit the manuscript to a journal for consideration. This practice would surely keep many Type I errors from appearing in print.

No less a figure than Sir Ronald Fisher (1929) made the following statement in which the dubiousness of the unreplicated study is highlighted:

> In the investigation of living beings by biologial methods statistical tests of significance are essential. Their function is to prevent us being deceived by accidental occurrences, due not to the causes we wish to study, or are trying to detect, but to a combination of the many other circumstances which we cannot control. An observation is judged significant, if it would rarely have been produced, in the absence of a real cause of the kind we are seeking. It is a common practice to judge a result significant, if it is of such magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator, but it does not mean that he allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained. *He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation.* (p. 190, italics added)

I would like to propose a more radical editorial reform, albeit one that has been advanced before (e.g., Greenwald, 1975; Lykken, 1968; Walster & Cleary, 1970). Instead of accepting or rejecting an article on the basis of its significance level, would it not be more desirable to accept it on the merit of its experimental design and potential contribution to the scientific community. According to Lykken (1968):

> Statistical significance is perhaps the least important attribute of a good experiment; it is never a significant condition for concluding that a theory has been corroborated, that a useful empirical fact has been estabilshed with reasonable con-

> fidence—or that an experimental report ought to be published. The value of any research can be determined, not from the statistical results, but only by skilled, subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied, and so on....Editors must be bold enough to take responsibility for deciding which studies are good and which are not, without resorting to letting the *p* value of the significance tests determine the decision. (pp. 158-159)

I am in complete agreement with Lykken's position. However, I would much prefer to eliminate the possible biasing effect of a reported probability value by requiring that the study be evaluated independently of it. That is, prior to data collection

Rather than submitting the *outcome* of an experiment for editorial review, the investigator would submit a comprehensive research proposal which would normally include the theoretical perspective, hypotheses, experimental design, and scientific implications of the study. The editor may accept the proposal outright or accept it with reservation that such things as one or more control groups be added, or a different criterion measure be used, or a particular statistical analysis by considered. Just as a thesis committee considers a graduate student's research proposal, judging its merits and deficiencies, making constructive additions and deletions before accepting it (or rejecting it as the case may be), so should a journal editor consider a manuscript submitted for publication. Furthermore, a graduate student's degree is not denied simply because the data do not achieve statistical significance. Why should such a stipulation be placed on the conscientious researcher who proposes to perform an empirical study that is recognized to have scientific merit?

Of course, an editor can reject the proposal. In this case, the researcher may elect to modify the proposal and submit it elsewhere or scuttle it entirely. Hopefully, the researcher will have learned a great deal from the experience and, if nothing else, has been saved the time and expense of an extended empirical investigation. The latter point is no small consideration. As the sciences do not enjoy the affluence of years gone by, it has become increasingly more important that resources not be depleted needlessly. If a study can be improved before it is carried out, think of the time, money, material, and energy that can be preserved. It is in this regard that an experienced editor can make the greatest contribution.

The investigator whose proposal has been accepted will continue with the project. The effort will include data collection, statistical analysis, and final report writing. But regardless of the statistical outcome of the experiment, the investigator need not fear for its publication. The publication was assured when the research proposal was accepted. Of course, the completed manuscript will be examined by the editor to insure that no unforeseen events have befallen the data collection and final report. In addition, the manuscript would have to conform to the

editorial standards and guidelines of the journal in which the article is to be published.

It may be argued that such an editorial policy is highly impractical if for no other reason than the greater amount of time it would consume compared to present procedures. I do not agree. Shifting the review process from the time period which follows the completion of a study to the time period prior to data collection would not require more time or effort, primarily because the product of the endeavor would not be appreciably different than that which would result from present procedures. It has been my experience, as well as that of my colleagues, that when an experimental study is rejected for publication, the reason for the rejection rarely is the result of the statistical outcome of the study, but is more likely to be concerned with methodological issues or the relative contribution the study makes to the scientific community. Actually, this should not be surprising. Because studies submitted for publication usually provide evidence for the rejection of the null hypothesis, it stands to reason that the high rejection rate characteristic of most scientific journals would reflect concerns *other* than the level of significance of the data. It may be concluded, therefore, that most experimental studies submitted for publication can be evaluated just as meaningfully without data as with it. More importantly, such an evaluation procedure would not suffer from anti-null hypothesis prejudice.

If this revision in editorial policy was adopted, one could calculate with certainty the number of Type I errors entering the literature. The number would be approximately 5% of all published research. Should editorial policy also demand that the investigator submit a replication attesting to the reliability of the findings, the percentage of Type I errors entering the literature would approach zero.

A possible flaw in this approach is that an acceptable research design does not insure care in the execution of the study. Carelessness in conducting the study would increase error variance and decrease the probability of a statistically significant outcome. A replication could contain the same systematic errors found in the original study. Hence, an increase in Type II errors—accepting a false null hypothesis. The best solution for this problem lies in the procedures for training competent researchers and not in editorial policy.

References

Bakan, D. The test of significance in psychological research. *Psychological Bulletin*, 1966, *66*, 432-437.

Fisher, R.A. The statistical method in psychical research. *Proceedings of the Society for Psychical Research*, 1929, *39*, 189-192.

Greenwald, A.G. Consequences of the prejudice against the null hypothesis. *Psychological Bulletin*, 1975, *82*, 1-20.

Lykken, D.T. Statistical significance in psychological research. *Psychological Bulletin,* 1968, *70,* 151-159.

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

Sterling, T.D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association,* 1959, *54,* 30-34.

Walster, G.W. & Cleary, T.A. A proposal for a new editorial policy in the social sciences. *The American Statistician,* 1970, *24,* 16-19.