practices in Gosplan's *Forms and Indicators*, cited above, (and given on p. 235 of the French translation) in which base and projected data on gross industrial activity in constant prices are given for the 11 years 1955–65.

[17] *Ekonomicheskaya gazeta*, no. 6, February 1969, p. 11.

[18] Sipov, N. E. "On the Calculation of the Social Product and National Income in Union Republics" in Akademiya nauk Kirgizskaya, *Voprosy ekonomiki sotsialisticheskoy promyshlennosti (Economic Problems of Socialist Industry)* Frunze, 1958, p. 56. Sipov states: "In a country where a comprehensive planned economy and a centralized system of recordkeeping obtain, the use of two methods of calculating a number of the most important indicators of the aggregate social product and of the national income is astonishing. It is necessary to note that not much is written on the uniformity of the indicators of the plan and of recordkeeping."

[19] Professor I. G. Malyy (quoted by Yevdokimenko, Iv. "In the Statistical Section of the Moscow House of Scientists," *Vestnik statistiki*, no. 7, 1969, pp. 72–73) states that "the statistical handbook provides no information on changes in the methodology of the plan" and requests "appropriate clarification and distinctions."

[20] Gusev, V. "Problems of Developing a Physical-Product Intersectoral Table," *Planovoye khozyaystvo*, no. 7, 1967, pp. 64–70. The basic difficulty is due to the formulation of gross industrial activity by Gosplan on the basis of data in physical terms and as an aggregate reported by enterprises to the Central Statistical Administration. In reference to the latter, Professor Kvasha writes that "it is difficult to find any large enterprise or institution, whether it is a metallurgical plant, a confectionery factory, or an administrative institution, which would not contain a number of auxiliary, subsidiary, and by-product production units and, of course, a repair shop, often with as many machine tools and wageworkers as would be typical for an average machine-building plant in the West." (Kvasha, Ya. and V. Krasovskiy, "Long-Term Planning and Economic Measurement," *Voprosy ekonomiki*, no. 4, 1968, pp. 26–31. See also the discussions of Professor Savinskiy, who compares the industrial enterprise to a feudal manor, and Mme. Shteyniger in *Ekonomicheskaya gazeta*, November 27, 1961, p. 33.

Thus, the machine-building output of each industrial enterprise is included in the gross industrial activity of the enterprise and of its branch of industry. Accordingly, defense industry, although a machine-building activity, is substantially grouped in the gross industrial activity of other industries. Hence, the great difficulty of reconciliation of production by product and by enterprise in Gosplan's Chief Computer Center.

[21] Gosplan's *Forms and Indicators* (cited above) refers to the planning of defense industry as a branch of the machine-building industry and Gosplan's *Methodological Regulations* (p. 23), states (in discussing the composition of industrial production in physical terms in the national economic plan) that it includes: "production for requirements of defense and special needs." Further (p. 24) it states that "the volume of production in the national economic plan is established for the country as a whole and by ministry and administrative agency which are basic suppliers of the given output." Also (p. 614) defense products are explicitly included in material balances.

On the other hand, in the statistics of the Central Statistical Administration, defense industry as a secondary industrial activity would be included in the gross industrial activity of many branches of industry. Even with the total arsenal of Gosplan statistics, these kinds of deconsolidation are most difficult. The Central Statistical Administration, because it collects very few data in physical terms (less than 100 and largely consisting of basic materials and different kinds of consumer goods) is unable to show the composition of industrial production. In particular, a decree of the Council of Ministers establishing the kinds of data to be submitted by enterprises, construction sites, and organizations to the Central Statistical Administration contains this proviso: "with the exception of defense production in physical terms." A careful examination of the kinds of data received by the Central Statistical Administration has found no references to the subject except oblique statements on the lack of the data. The text of the decree cited above is given by Romashkin, P. S. et al. (Eds.), *Zakonodatel'nyye akty po voprosam narodnogo khozyaystva SSSR (Legal Acts on National Economic Questions in the U.S.S.R.)*, Volume I, Moscow, Gosyurizdat, 1961, p. 687.

The decree states: "Enterprises, construction organizations, and other organizations . . . submit statistical reports to the statistical administrations . . . of the Central Statistical Administration with the exception of reporting on the production of defense production in physical terms which is submitted only to their branch administrative agencies . . . and ministries." In view of the collection, grouping, and compilation of data on defense in national tabulations of the Gosplan system, the decree acts to prevent a repetition of this procedure in another statistical system.

A Proposal for a New Editorial Policy in the Social Sciences[1]

G. WILLIAM WALSTER and T. ANNE CLEARY
The University of Wisconsin, Madison

". . . there's this desert prison, see, with an old prisoner, resigned to his life, and a young one just arrived. The young one talks constantly of escape, and, after a few months, he makes a break. He's gone a week, and then he's brought back by the guards. He's half dead, crazy with hunger and thirst. He describes how awful it was to the old prisoner. The endless stretches of sand, no oasis, no signs of life anywhere. The old prisoner listens for a while, then says, 'Yep. I know. I tried to escape myself, twenty years ago.' The young prisoner says, 'You did? Why didn't you tell me, all these months I was planning my escape? Why didn't you let me know it was impossible?' And the old prisoner shrugs, and says, 'So who publishes negative results?' " (Hudson, 1968, p. 168)

A virtual prerequisite for the publication of research in the social sciences is the attainment of statistical significance. Anyone who has read this literature knows that it contains few articles that fail to report statistically significant results. Systematic evidence to support this contention has been provided by Theodore Sterling (1959). In a survey of four journals of the American Psychological Association he found that, out of 294 articles using statistical tests, only eight reported results that failed to reach the .05 level of significance.

Moreover, Sterling was unable to find a single replication of previously published research.

The fact that this literature contains virtually no articles that report replications or statistically nonsignificant results is undoubtedly a consequence of current editorial policy. This policy appears to be based on the belief that a single instance of statistical significance can establish the importance of a research finding.

Unfortunately, because of its consequences, this editorial policy precludes use of journals to record and communicate accumulated knowledge.[2] The first of these consequences is that it is impossible for a reader to differentiate those articles that report important findings from those that report Type I errors. For example, if twenty studies testing the same hypothesis are executed, on the average, one of them will attain the .05 level of significance by chance alone. If this study is published, a result that is a Type I error, and nothing more, will appear in a professional journal. Not only will this Type I error be published but editorial policy guarantees that this error will never be publicly exposed! Only by allowing publication of replications and failures to replicate will Type I errors in the literature be uncovered. At present, only by consulting the underground network of people who have worked in a particular area, can one uncover the information required to discriminate between Type I errors and important findings.

The second consequence of current editorial practice is that it leads researchers to *search* for statistical significance. This search is motivated by the knowledge that statistical significance is a prerequisite for publication and by the awareness that, in many academic environments, one is required to publish research in order to survive, let alone prosper. The necessity to "publish or perish," combined with the practice of requiring statistical significance for publication, surely influences the attitudes and practices of even the most conscientious and objective scientist. Unfortunately, the very fact that researchers engage in this search destroys the meaning of any reported significance levels.

Consider two tactics that researchers have employed to attain statistical significance. First, researchers may collect their data in a way that almost guarantees significant results. One way is to begin with, say, five subjects in each condition of an experiment. If the results are significant, terminate the experiment. However, if results are not significant, add five more subjects per condition and pool these with the initial five to yield ten subjects per cell. This process is repeated until significant results are obtained or the cost of adding another five subjects per cell is not worth a publication. In order for a sequential sampling technique like this to be used legitimately, methods of analysis very unlike those normally employed are required. If classical methodology is applied to a study in which the sampling has been sequential, the probability of committing Type I errors may be much greater than that announced.

A second way of generating statistical significance is to perform many different analyses on a large amount of data. This is a particularly efficient method, as computers are readily available to eliminate the tedious aspects of performing many analyses. As an example, in a particular study a researcher may collect ten dependent measures and perform multiple comparisons on each of them. Because of the large number of tests performed, statistical significance is almost assured, a priori.

In pointing out these techniques that researchers have employed to attain statistical significance, we intend not to chastise them, but to illustrate an effect of current editorial policy: not only does this policy insure that Type I errors are indistinguishable from important findings, but it markedly increases the frequency with which these errors are made.

A third consequence of current editorial policy is that a large number of studies are not published even though they contain important, but statistically insignificant findings. An experiment whose results did not achieve some arbitrary level of significance may contain valuable information for interested readers. For example, a reader may wish to search for a critical modification of the experiment or the theory that led to its design. Or, if a particular hypothesis is not rejected in several well-designed and executed studies, a reader may conclude that the hypothesis is, at most, trivially false. Neither of these alternatives is available to the researcher, unless he has the information contained in the studies that currently are not being published.

A fourth undesirable consequence of current editorial practice also stems from the loss of research that does not attain statistical significance. Journal readers are now, in effect, letting editors and reviewers decide for them the level of significance that is indicative of an important finding. It is well known that the power of a statistical test is a very fragile quantity: it can be affected by a variety of factors, such as sample size, reliability of the dependent measure, and the strength of the various treatments used in the experiment. Upon consideration of these factors, a knowledgeable reader of the literature will wish to determine for himself the level of statistical significance that indicates what to him is an important finding. If the studies whose results fall below a given arbitrary level of significance are not published, it is impossible for a reader to make rational decisions concerning the presence or absence of effects of interest to him.

Some of these same problems have been raised in the context of criticisms of classical hypothesis-testing methodology (Bakan, 1966; Faia, 1966; Kish, 1959; Lykken, 1968; Meehl, 1967; Selvin, 1957; Selvin, 1966; Sterling, 1959). However, no one has pointed out that these problems are caused by current editorial policy or sought a feasible alternative to present publication procedures. If the social sciences are to proceed in a rational fashion, an alternative must be found and implemented.

In the remainder of this paper, we shall consider an alternative publication policy for articles that report inferential statistical analyses of data. In proposing this alternative policy, we argue that all decisions involving the treatment of data should be considered design decisions. Then, since the decision to publish the results of a study is a particular treatment of data, it follows that the same limitations should be imposed on publication decisions as are imposed on all design decisions. When one views publication in this way, it becomes immediately clear that a specific change should be made in current policy. There is a cardinal rule in experimental design that any decision regarding the treatment of data must be made *prior* to an inspection of the data. If this rule is extended to publication decisions, it follows that when an article is submitted to a journal for review, the data and the results should be withheld. This would insure that the decision to publish, or not to publish, would be unrelated to the outcome of the research. The decision to publish would be based upon such factors as the adequacy of the design, and the relevance of the research to current theoretical and topical issues.[3]

With the exception of the results and conclusions, any information that tends to support the publication of the proposed study could be included in that portion of the article submitted for review. The following topics could be covered:

1. Theoretical relevance and/or justification;
2. Relevance to applied and/or topical issues;
3. Predicted outcomes and the implications for the theoretical and/or applied problems;
4. If the study is or includes a replication of previously published research, a discussion of the need for the proposed replication;
5. Detailed description of the procedure of the study, including the source of subjects, description of randomization scheme, transcript of instructions given subjects, description of independent and dependent variables;
6. Any previous research and/or data which indicates the extent to which the independent variables are validly being manipulated;
7. Any previous research and/or data which indicate the extent to which the dependent variables are reliable and/or valid;
8. Discussion of the proposed data analysis and its relevance to the predicted outcomes;
9. Pilot data that tend to support the predicted outcome, especially if the predictions are at variance with existing theories or published results.

This change in review procedures, along with a marked increase in the frequency with which replications are accepted for publication, would have many beneficial effects on the literature in the social sciences. First, an examination of the literature would enable one to determine which studies are reporting Type I errors and which are demonstrating important findings. A result that is statistically significant by chance alone would be accompanied in the literature by a number of failures to replicate, while important findings would be accompanied by several confirmations. Second, the pressure to publish would be re-directed so that researchers would feel a need to design substantively important and methodologically sound studies rather than merely to achieve statistical significance. Third, the information contained in studies which do not achieve statistical significance would not be lost as it is now. Fourth, it would be possible for each reader to judge for himself the ultimate importance of the results of published research.

This change in editorial policy would have an additional important effect on research practices in the social sciences. The obvious strategy of the researcher would be to submit for review a proposed study for which the data have not been collected. After a proposal is accepted, a researcher could execute his study with the guarantee that it would be published whether or not the results are statistically significant. If the paper is rejected, the effort involved in executing the study would not be lost. If it is accepted on the condition that some methodological or design change be made, such changes could be feasibly incorporated into the study before it is conducted. This review procedure, which does not require execution of studies prior to their acceptance for publication, would eliminate the waste of executed but unpublished research. (Current rejection rates in APA journals run from 50 to 88 percent [Newman, 1966].)

Of course, we recognize that any disruption of a long-established tradition may cause some problems. For example, Rosenthal (1966) suggests that a similar review procedure might lead to an increased demand for journal space.[4] Initially there may be an increase in the number of proposals submitted. However, when researchers become aware of the necessary preparation and possible pre-testing that will be requisite for a competitive proposal, we expect that the number of proposals requiring serious review will be less than it is now. Moreover, even if this and other unforseen practical problems arise, the benefit of a truly useful literature must far exceed any temporary inconvenience associated with the transition.

NOTES

[1] An earlier draft of this paper titled *"Current Editorial Policy: A Type I Error"* was reviewed at the Nineteenth International Congress of Psychology, London, 1969.

[2] The need for a change in editorial policy has been recognized by others. In 1959, Tullock encouraged journals to make space for replications. In 1960, Quinn McNemar discussed many of the dangers inherent in current editorial policy in his Presidential Address of the Western Psychological Association. More recently, students in social psychology at the University of North Carolina have organized a new journal, *Representative Research in Social Psychology*. It is the stated editorial policy of this new journal to publish not only results that are not statistically significant, but replications and failures to replicate.

In addition, two journals in education, *The Journal of Experimental Education* and *The Journal of Educational Research*, are changing their editorial policies on the basis of the recommendations in this paper.

[3,4] A similar editorial policy has been recommended before. Rosenthal (1966), concerned that researchers focus almost exclusively on outcomes, advocates a review procedure that excludes results:

> "What we need is a system for evaluating research based only on the procedures employed. If the procedures are judged appropriate, sensible, and sufficiently rigorous to permit conclusions from the results, the research cannot then be judged inconclusive on the basis of the results and rejected by the referees or editors. Whether the procedures were adequate would be judged independently of the outcome. To accomplish this might require that procedures only be submitted initially for editorial review or that only the result-less section be sent to a referee or, at least, that an evaluation of the procedures be set down before the referee or editor reads the results. This change in policy would serve to decrease the outcome-consciousness of editorial decisions, but it might lead to an increased demand for journal space. This practical problem could be met in part by an increased use of "brief reports" which summarize the research in the journal but promise the availability of full reports to interested workers."

REFERENCES

Bakan, David. "The test of significance in psychological research." *Psychological Bulletin*, 1966, *66* (6), 423–437.

Faia, Michael A. "A proposal for the standardization of Type II error in the replication of social research." *Sociology and Social Research*, 1966, *51*, 87–93.

Hudson, Jeffrey. *A Case of Need*. New York: The New American Library, Inc., 1968.

Kish, Leslie. "Some statistical problems in research design." *American Sociological Review*, 1959, *24*, (3), 328–338.

Lykken, David T. "Statistical significance in psychological research." *Psychological Bulletin*, 1968, *70*, (3), 151–159.

McNemar, Quinn. "At random: Sense and nonsense." *American Psychologist*, 1960, *15*, 295–300.

Meehl, P. E. "Theory testing in psychology and physics: A methodological paradox." *Philosophy of Science*, 1967, *34*, 103–115.

Newman, S. H. "Improving the evaluation of submitted manuscripts." *American Psychologist*, 1966, *21*, 980–981.

Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Croft, 1966.

Selvin, Hanan C. "A critique of tests of significance in survey research." *American Sociological Review*, 1957, *22*, (5), 519–527.

Selvin, Hanan C. and Stuart, A. "Data-dredging procedures in survey analysis." *American Statistician*, *20*, (3), 1966, 20–23.

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

Tullock, Gordon. "Publication decisions and tests of significance —a comment." *Journal of the American Statistical Association*, 1959, *54*, 593.

Is Kurtosis Really "Peakedness?"[1]

RICHARD B. DARLINGTON
Cornell University

Abstract

Kurtosis is best described not as a measure of peakedness versus flatness, as in most texts, but as a measure of unimodality versus bimodality.

In a recent survey of the elementary statistical texts on his shelf, the present writer found 11 which attempted to explain the concept of kurtosis. Ten of these unambiguously used terms like "peaked" and "flat-topped" to describe high-kurtosis (leptokurtic) and low-kurtosis (platykurtic) distributions respectively. One pointed out that *platy* is the Greek word for flat. Only one specialized text [1, p. 68] suggested that the description "flat" does not adequately describe distributions with low kurtosis, and even there the details were unclear. The purpose of the present paper is to make somewhat clearer what kurtosis measures, what sorts of distributions have high and low kurtosis, and how kurtosis is altered when observations are added to an existing distribution.

The formal definition of kurtosis

For simplicity of notation we shall confine the discussion to distributions with finite numbers of observations. The generalization to infinite distributions is straightforward.

If m and s denote the mean and standard deviation of a distribution (where the denominator in s^2 is N rather than $N - 1$), then the usual measure of kurtosis[2] is

$$k = \frac{N^{-1}\sum(X - m)^4}{s^4}.$$

k is unaffected by changes in the mean or standard deviation of the distribution. k can be simply expressed as a function of the z scores; it is easy to show that

$$k = N^{-1}\sum z^4.$$

What does kurtosis mean in intuitive terms?

Most elementary texts describe kurtosis as a measure of the "peakedness" of a distribution. We shall attempt to show that this term is misleading, and that a far better term for describing kurtosis is "bimodality," where the lower the kurtosis, the greater the bimodality. We shall make this point in three ways:

(a) by an algebraic analysis of the formula for k, (b) by examining distributions with low and high kurtosis, and (c) by examining the change in k resulting from modifying an existing distribution by adding observations at various points.

Our "algebraic analysis of the formula for k" can be completed in one short paragraph. In a distribution of z scores, it is true by definition that

$$\text{Mean}(z^2) = 1.$$