Open access • Journal Article • DOI:10.1037/H0076157

Consequences of Prejudice Against the Null Hypothesis — Source link

Anthony G. Greenwald

Institutions: Ohio State University

Published on: 01 Jan 1975 - Psychological Bulletin (American Psychological Association)

Topics: Size, Null hypothesis, Alternative hypothesis, Statistical hypothesis testing and P-rep

Psychological Bulletin
1975, Vol. 82, No. 1, 1–20

Consequences of Prejudice Against the Null Hypothesis

Anthony G. Greenwald
Ohio State University

The consequences of prejudice against accepting the null hypothesis were examined through (a) a mathematical model intended to stimulate the research–publication process and (b) case studies of apparent erroneous rejections of the null hypothesis in published psychological research. The input parameters for the model characterize investigators' probabilities of selecting a problem for which the null hypothesis is true, of reporting, following up on, or abandoning research when data do or do not reject the null hypothesis, and they characterize editors' probabilities of publishing manuscripts concluding in favor of or against the null hypothesis. With estimates of the input parameters based on a questionnaire survey of a sample of social psychologists, the model output indicates a dysfunctional research–publication system. Particularly, the model indicates that there may be relatively few publications on problems for which the null hypothesis is (at least to a reasonable approximation) true, and of these, a high proportion will erroneously reject the null hypothesis. The case studies provide additional support for this conclusion. Accordingly, it is concluded that research traditions and customs of discrimination against accepting the null hypothesis may be very detrimental to research progress. Some procedures that can help eliminate this bias are prescribed.

In a standard college dictionary (*Webster's New World,* College Edition, 1960), *null* is defined as "invalid; amounting to nought; of no value, effect, or consequence; insignificant." In statistical hypothesis testing, the *null hypothesis* most often refers to the hypothesis of no difference between treatment effects or of no association between variables. Interestingly, in the behavioral sciences, researchers' null hypotheses frequently satisfy the nonstatistical definition of *null,* being "of no value," "insignificant," and presumably "invalid." My aims here are to document this state of affairs, to examine its consequences for the archival accumulation of scientific knowledge, and lastly, to make a positive case for the formulation of more potent and acceptable null hypotheses as a part of an overall research strategy.

Because of my familiarity with its literature, most of the illustrative material I use is drawn from social psychology. This should not be read as an implication that the problems being discussed are confined to social psychology. I suspect they are equally characteristic of other behavioral science fields that are lacking in well-established organizing theoretical systems.

The Lowly Null Hypothesis

My paraphrasing of some widespread beliefs of behavioral scientists concerning the null hypothesis appears below. Some partial sources for the content of this listing are

Preparation of this report was facilitated by grants from National Science Foundation (GS-3050) and U.S. Public Health Service (MH-20527-02). Although they should not be held responsible for positions espoused herein, I am very grateful to the following for providing comments on earlier drafts: Mari R. Jones, Paul Isaac, David Bakan, Timothy C. Brock, Bibb Latané, Thomas M. Ostrom, Hanan C. Selvin, Martin Fishbein, Zick Rubin, and Richard A. Zeller.

Requests for reprints should be sent to Anthony G. Greenwald, Department of Psychology, Ohio State University, 404C West 17th Avenue, Columbus, Ohio 43210.

Festinger (1953, pp. 142–143), Wilson and Miller (1964), Aronson and Carlsmith (1969, p. 21), and Mills (1969, pp. 442–448).

1. Given the characteristics of statistical analysis procedures, a null result is only a basis for uncertainty. Conclusions about relationships among variables should be based only on rejections of null hypotheses.

2. Little knowledge is achieved by finding out that two variables are unrelated. Science advances, rather, by discovering relationships between variables.

3. If statistically significant effects are obtained in an experiment, it is fairly certain that the experiment was done properly.

4. On the other hand, it is inadvisable to place confidence in results that support a null hypothesis because there are too many ways (including incompetence of the researcher), other than the null hypothesis being true, for obtaining a null result.

Given the existence of such beliefs among behavioral science researchers, it is not surprising that some observers have arrived at conclusions such as:

> Many null hypotheses tested by classical procedures are scientifically preposterous, not worthy of a moment's credence even as approximations. (Edwards, 1965, pp. 401–402)

> It [the null hypothesis] is usually formulated for the express purpose of being rejected. (Siegel, 1956, p. 7)

Refutations of Null Hypothesis "Cultural Truisms"

I am sure that many behavioral science researchers endorse the beliefs previously enumerated but would have difficulty in providing a rational defense for these beliefs should they be strongly attacked. That is, these attitudes toward the null hypothesis may have some of the characteristics of cultural truisms as described by McGuire (1964). Cultural truisms are beliefs that are so widely and unquestioningly held that their adherents (a) are unlikely ever to have heard them being attacked and may therefore (b) have difficulty defending them against an attack. If I am correct, the reader will have difficulty defending the preceding beliefs against the following attacks (the numbered paragraphs correspond to those in the preceding listing.) Briefly stated, these attacks are:

1. The notion that you cannot prove the null hypothesis is true in the same sense that it is also true that you cannot prove *any* exact (or point) hypothesis. However, there is no reason for believing that an estimate of some parameter that is near a zero point is less valid than an estimate that is significantly different from zero. Currently available Bayesian techniques (e.g., Phillips, 1973) allow methods of describing acceptability of null hypotheses.

2. The point is commonly made that theories predict relationships between variables; therefore, finding relationships between variables (i.e., non-null results) helps to confirm theories and thereby to advance science. This argument ignores the fact that scientific advance is often most powerfully achieved by *rejecting* theories (cf. Platt, 1964). A major strategy for doing this is to demonstrate that relationships predicted by a theory are not obtained, and this would often require acceptance of a null hypothesis.

3. I am aware of no reason for thinking that a statistically significant rejection of a null hypothesis is an appropriate basis for assuming that the conceptually intended variables were manipulated or measured validly. The significant result (barring Type I error) does indicate that some relationship or effect was observed, but that is all it indicates. The researcher who would claim that his data show a relationship between two variables should be as clearly obliged to show that those variables are the ones intended as should the researcher who would claim that his data show the absence of a relationship.

4. Perhaps the most damaging accusation against the null hypothesis is that incompetence is more likely to lead to erroneous nonsignificant, "negative," or null results than to erroneous significant or "positive" results. There is some substance to this accusation—when the incompetence has the effect of introducing noise or unsystematic error into data. Examples of this sort of incompetence are the use of unreliable paper-and-pencil measures, conducting research in a "noisy" setting (i.e., one with important extraneous variables uncontrolled), unreliable apparatus

functioning, inaccurate placement of recording or stimulating electrodes, random errors in data recording or transcribing, and making too few observations. These types of incompetence are often found in the work of the novice researcher and are proper cause for caution in accepting null findings as adequate evidence for the absence of effects or relationships. Some other very common types of incompetence are much more likely to produce false positive or significant results. These types of incompetence result in the introduction of *systematic* errors into data collection. Examples of such sources of artifact (cf. Rosenthal & Rosnow, 1969) are experimenter bias, inappropriate demand characteristics, nonrandom sampling, invalid or contaminated manipulations or measures, systematic apparatus malfunction (e.g., errors in calibration), or systematic error (either accidental or intentional) in data recording or transcribing. This latter category of incompetence is by no means confined to novices and may be quite difficult to detect, particularly since our existing customs encourage greater suspicion of null findings than of significant findings.

Behavioral Symptoms of Anti-Null-Hypothesis Prejudice

We should not perhaps be very disturbed about the existence of the beliefs previously listed if those beliefs would prove to be unrelated to behavior. The following is a list of some possible behavioral symptoms of prejudice against null hypotheses: (a) designing research so that the personal prediction of the researcher is identified with rejection rather than acceptance of the null hypothesis; (b) submitting results for publication more often when the null hypothesis has been rejected than when it has not been rejected; (c) continuing research on a problem when results have been close to rejection of the null hypothesis ("near significant"), while abandoning the problem if rejection of the null hypothesis is not close; (d) elevating ancillary hypothesis tests or fortuitous findings to prominence in reports of studies for which the major dependent variables did not provide a clear rejection of the null hypothesis; (e) revising otherwise adequate operationalizations of variables when unable to obtain rejection of the null hypothesis and continuing to revise until the null hypothesis is (at last!) rejected or until the problem is abandoned without publication; (f) failing to report initial data collections (renamed as "pilot data" or "false starts") in a series of studies that eventually leads to a prediction-confirming rejection of the null hypothesis; (g) failing to detect data analysis errors when an analysis has rejected the null hypothesis by miscomputation, while vigilantly checking and rechecking computations if the null hypothesis has not been rejected; and (h) using stricter editorial standards for evaluating manuscripts that conclude in favor of, rather than against, the null hypothesis.

Perhaps the enumeration of the items on this list will arouse sufficient recognition of symptoms in readers to convince them that the illness of anti-null-hypothesis prejudice indeed exists. However, just as a hypochondriac should have better evidence that he is ill than that the symptoms he has just heard about seem familiar, so should we have better evidence than symptom recognition for making conclusions about the existence of prejudice against the null hypothesis.

A Survey to Estimate Bias Against the Null Hypothesis

In order to obtain some more concrete evidence regarding the manifestations of anti-null-hypothesis prejudice, I conducted a survey of reviewers and authors of articles submitted to the *Journal of Personality and Social Psychology* (*JPSP*). The sample included the primary (corresponding) authors and the reviewers for all manuscripts that I processed as an associate editor of *JPSP* during a 3-month period in 1973. The sample thus consisted of 48 authors and 47 reviewers to whom I sent a questionnaire. Returns were obtained from 36 authors (75%) and 39 reviewers (81%). The major items in the questionnaire assessed behavior in situations in which bias for or against the null hypothesis could occur. These situations were (a) initial formulation of a problem, (b) setting probabilities of Type I and Type II error, and (c) deciding what action to pursue once results were obtained. All questions were stated with refer-

TABLE 1

Results of Survey of *JPSP* Authors and Reviewers to Determine Prejudice Toward or Against the Null Hypothesis

Question	Mean responses for			
	Reviewers	Authors	All	$SD_{M_{all}}$
1. What is the probability that your typical prediction will be for a rejection (rather than an acceptance) of a null hypothesis?	.790(39)	.829(35)	.803(74)	.021
2. Indicate the level of alpha you typically regard as a satisfactory basis for rejecting the null hypothesis.	.043(39)	.049(35)	.046(74)	.002
3. Indicate the level of beta you would regard as a satisfactory basis for accepting the null hypothesis.	.292(18)	.258(19)	.274(37)	.045
4. After an initial full-scale test of the focal hypothesis that allows rejection of the null hypothesis, what is the probability that you will				
(a) submit the results for publication before further data collection,	.408(38)	.588(35)	.494(73)	.033
(b) conduct an exact replication before deciding whether to submit for publication,	.078(38)	.069(35)	.074(73)	.009
(c) conduct a modified replication before deciding whether to submit,	.437(38)	.289(35)	.366(73)	.027
(d) give up the problem.	.077(38)	.053(35)	.066(73)	.012
Total	1.000	1.000	1.000	
5. After an initial full-scale test of the focal hypothesis that does not allow rejection of the null hypothesis, what is the probability that you will				
(a) submit the results for publication before further data collection,	.053(37)	.064(35)	.059(73)	.014
(b) conduct an exact replication before deciding whether to submit for publication,	.107(37)	.098(36)	.102(73)	.013
(c) conduct a modified replication before deciding whether to submit,	.592(37)	.524(36)	.558(73)	.025
(d) give up the problem.	.248(37)	.314(36)	.280(73)	.023
Total	1.000	1.000	1.000	

Note. Table entries are means of respondents' estimates of probabilities, based on the number of responses given in parentheses.

ence to a test of the "focal hypothesis" for a new line of research. The focal hypothesis test was further defined as "the one hypothesis test that is of greatest importance" to the line of investigation. Responses were indicated on probability scales that could range from 0 to 1.00. The major results are given in Table 1.

With the exception of responses to two questions, the results for authors and reviewers were quite similar. This was not terribly surprising because there was substantial overlap between the populations from which these two subsamples were drawn. From Questions 4a and 4d it can be seen that authors reported they were more likely to report null hypothesis rejections and less likely to abandon the problem following a null hypothesis rejection than were reviewers. Given these rather limited differences, the following discussion of these data treats only the overall responses for the combined sample.

The questionnaire results gave several strong confirmations of existence of prejudice against the null hypothesis. In the stage of formulation of a problem, respondents indicated a strong preference for identifying their own predictions with an expected rejection, rather than an acceptance of the null hypothesis. The mean probability of the researcher's personal prediction being of the null hypothesis rejection (Question 1: $\bar{X} = .81 \pm .04$) is substantially greater than .50.[1] This state of affairs is consistent with supposing that researchers set themselves the goal of confirming a theoretically predicted relation between variables more often than refuting one, despite good reason to believe that knowledge may advance more rapidly by the latter strategy (Platt, 1964).

In setting the probability of Type I error, respondents indicated relatively close adherence to the .05 alpha criterion (Question 2: $\bar{X} = .046 \pm .004$). Responses to Question 3 indicated a substantial lack of standard practice with regard to Type II errors (i.e., accepting the null hypothesis when in truth it should be rejected). About 50% of the respondents failed to answer the question requesting specification of a preferred Type II error (beta) criterion. Those who did indicate a Type II error criterion indicated much more tolerance for this type of error than for a Type I error, the resulting estimate of beta being approximately .30 (Question 3: $\bar{X} = .27 \pm .09$). This estimate, it should be noted, is in line with Cohen's (1962) conclusion that studies published in the *Journal of Abnormal and Social Psychology* were relatively low on power (probability of rejecting the null hypothesis when the alternative is true; power = 1.00 − beta). In regard to tolerance for Type I and Type II errors then, the questionnaire respondents appeared biased *toward* null hypothesis acceptance in the sense that they reported more willingness to err by accepting, rather than rejecting, the null hypothesis. Such a conclusion would, I think, be quite misleading. Rather, responses to other questions not summarized in Table 1 and the frequency of nonresponse to Question 3 indicated that most respondents did not take seriously the idea of setting a Type II error criterion in advance. For example, the responses to questions asking for probability of setting alpha and beta criterions in advance of data collection indicated a .63 (±.09) probability that alpha would be set in advance of data collection, compared with only a .17 (±.06) probability that beta would be set in advance. Rather than indicating a prejudice toward acceptance of the null hypothesis then, I think the responses to the questions on alpha and beta indicate that acceptance of the null hypothesis is not usually treated as a viable research outcome.

In terms of what is done after completion of a full-scale data collection to test a focal hypothesis, a major bias is indicated in the .49 (±.06) probability of submitting a rejection of the null hypothesis for publication (Question 4a) compared to the low probability of .06 (±.03) for submitting a nonrejection of the null hypothesis for publication (Question 5a). A secondary bias is apparent in the probability of continuing with a problem and is computed conditionally upon the decision to write a report having *not* been made following data collection. This derived index has a value of .86 (±.05) when the initial result is a rejection of the null hypothesis, compared to .70 (±.05) when the initial result is a nonrejection of the null hypothesis, indicating greater likelihood of proceeding in the former case.[2]

In sum, the questionnaire responses of a sample of contributors to the social psychological literature gave self-report evidence of substantial biases against the null hypothesis in formulating a research problem and in deciding what to do with the data once collected. In the following section, the impact of these biases on the content of the archival literature is considered.

A MODEL OF THE RESEARCH–PUBLICATION SYSTEM

The alpha criterion most commonly employed in the behavioral sciences is .05. Without giving the matter much thought, one may guess on this basis that approximately 1 in 20 publications may be an erroneous rejection of a true null hypothesis. However, some thought on the matter soon brings the discovery that the probability of a published article being a Type I error depends on much

[1] The errors of estimates given are equal to the limits of 95% confidence intervals, approximately plus or minus twice the standard deviation of the estimated mean.

[2] In the case of a result rejecting the null hypothesis, this index is computed as [(4b + 4c) ÷ (4b + 4c + 4d)], the numbers referring to the responses to the questions given in Table 1.

more than (a) the researcher's alpha criterion. The other determinants include (b) the probability of accepting the null hypothesis when it is false (Type II error or beta), (c) the a priori probability of an investigator selecting a problem for which the null hypothesis is true or false, (d) the probability of rejections versus nonrejections of the null hypothesis being submitted for publication, (e) the probability of the researcher's giving up in despair after achieving a rejection versus a nonrejection of the null hypothesis, and (f) the probability of an editor's accepting an article that reports a rejection versus a nonrejection of the null hypothesis. All of these probabilities represent opportunities for the occurrence of strategies that discriminate against the null hypothesis. The model I develop functions to derive consequences for the content of published literature from assumptions made about these strategies.

Model Description

In the model employed for the research–publication system (see Figure 1), a critical notion is that of a *focal hypothesis test.* It is assumed that in any line of investigation, there is one statistical test that is of major interest. This may be a test for a main or interaction effect in an analysis of variance, a test of the difference between two groups or treatments, a test of correlation between two variables, and the like. This statistical test is assumed to be made in terms of a rejection or acceptance (nonrejection, if you prefer) of a null hypothesis of no main effect, no interaction, and so forth. In conducting this focal hypothesis test, the researcher is assumed to have formulated an extent of deviation from the null hypothesis (an alternative hypothesis, H_1) that he would like to be able to detect with probability (power) $1 - \beta$. In practice, this formulation of H_1 may often be an implicit consequence of setting a critical region for rejection of the null hypothesis with a given risk, α, of Type I error. For example, assuming $\beta = \alpha$, the start of the critical region is effectively a midpoint between the null hypothesis and H_1.

In the model, the fate of a research problem is traced in terms of the probabilities of alternative outcomes at four types of choice points: (a) the researcher's formulation of a hypothesis, (b) his collection of data, (c) his evaluation of obtained results, and (d) an editor's judgment of a manuscript reporting the research results. At each of these points in the research–publication process, behavioral bias relating to the null hypothesis may enter. The model incorporates parameters that serve to quantify these biases, and these are listed here in their sequential order of occurrence in the research–publication process.

The probability that the null hypothesis is true for the focal hypothesis. Specification of this parameter requires a clear definition of the null hypothesis. 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, as would be the probability of any other point hypothesis. In most cases, however, the investigator should not be concerned about the hypothesis that the true value of a statistic equals exactly zero, but rather about the hypothesis that the effect or relationship to be tested is so small as not to be usefully distinguished from zero. For the purposes of the model then, the probability of the null hypothesis being true becomes identified with the probability that the true state of affairs underlying the focal hypothesis is within a *null range* (cf. Hays, 1973, pp. 850–853). In the model, the probability that the investigator's focal hypothesis is one for which truth is within such a null range is represented as h_0. The probability that truth is outside this range is $h_1 = 1.00 - h_0$. One would have to be omniscient to be assured of selecting accurate values for the h_0 and h_1 parameters. It seems, however that, the values of these parameters should be clearly weighted in the direction of starting with a false null hypothesis (i.e., $h_1 > h_0$). Some reasons for this are that (a) researchers identify their personal predictions predominantly with the falsity of the null hypothesis (see Table 1, Question 1), and there may often be good reason for them to make these predictions; and (b) as argued by McGuire (1973), there is usually at least a narrow sense in which

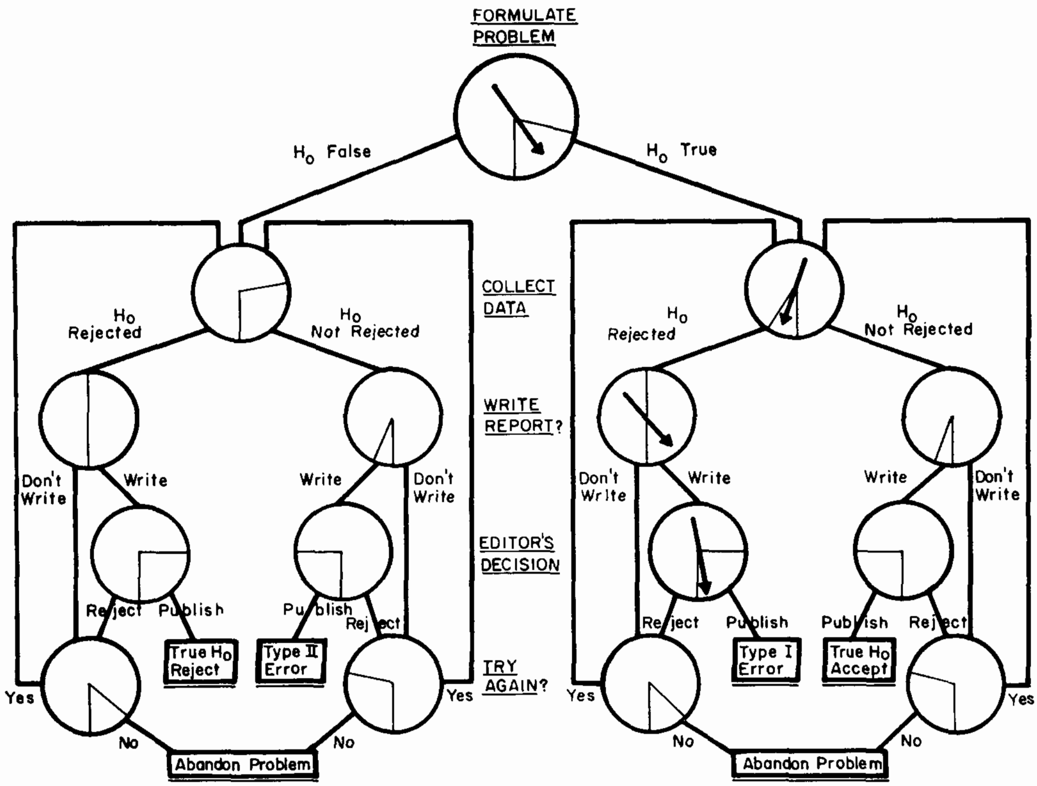

FIGURE 1. Model of research–publication system. (Five types of sequential decision points in the research–publication process are represented by rows of circles that can be thought of as spinners in a board game, each spinner selecting one of two departures from the decision point. The spinners shown on four of the circles depict a published Type I error resulting from a researcher's first data collection on a problem.)

most researchers' predictions are correct. For no outstandingly good reason, the values of .20 and .80 were selected for h_0 and h_1, respectively. To compensate for the difficulty of justifying this initial assumption, system results are given below for other values of these parameters.[3]

Outcome of data collection. As used here *data collection* refers to the researcher's activities subsequent to problem formulation, up to and including the statistical analysis of results. It is assumed that any such data collection can be characterized by probabilities of Type I and Type II errors that are either explicitly chosen by the investigator or else follow implicitly (cf. Cohen, 1962) from his choices of sample size, dependent measures, statistical tests, and the like. (Because investigators may often examine data midway in a planned piece of research and thereupon terminate or otherwise alter plans, the notion of a data collection is somewhat vague in practice and must necessarily be so in the model). The outcome of a data collection will either be a rejection or a nonrejection of the null hypothesis. The probability of rejection if the null hypothesis is true is characterized in the model as r_0 and is approximately equivalent to the researcher's alpha criterion. Based

[3] The equations for computing system output indexes have been prepared as a computer program in the BASIC language. This program generates system output indexes in response to values of the system input parameters entered at a terminal by the user. The program therefore permits ready examination of consequences of assumptions other than those made presently about values of the system's input parameters. A listing of this program may be obtained from the author.

on the questionnaire responses, r_0 is estimated at .05. If the null hypothesis is false, then the probability of its rejection is characterized as r_1 and this should be approximately 1.00 minus the researcher's beta criterion. This value is estimated at .70 based on the questionnaire responses. Probabilities of nonrejection of the null hypothesis are 1.00 minus r_0 (which equals .95) or 1.00 minus r_1 (which equals .30), respectively.[4]

Probability of writing a report. The model assumes that upon completing a data collection, the researcher examines his results and decides whether or not to write a report. The probability of deciding to write if the null hypothesis has not been rejected is represented as w_0 and is estimated at .06, based on the questionnaire results (see Table 1, Question 5a). When the null hypothesis has been rejected the probability of deciding to write is represented as w_1 and is estimated at .49, based on the questionnaire responses (Question 4a).

Probability of editorial acceptance. In order for the result of a data collection to appear in print, it has to be accepted for publication by an editor. In the model, an editor accepts an article with probability e_0 if it reports a nonrejection of the null hypothesis and e_1 if it reports a rejection of the null hypothesis. The questionnaire data did not permit any estimates of these parameters and they have been estimated, somewhat arbitrarily, as both being equal to .25. Thus, although the model permits analysis of the consequences of editorial discrimination for or against the null hypothesis, no initial assumption has been made regarding the existence of such bias.

If at first you don't succeed. The researcher may be left holding a bagful of data if (a) he has decided not to report the results or (b) he has decided to report them but has been unable to obtain the cooperation of an editor. At this point, the model allows the researcher to decide whether to continue research or to abandon the problem. If the result of the preceding data collection was a nonrejection of the null hypothesis, the probability of continuing is represented as c_0; if the result was a rejection of the null hypothesis, the probability of continuing is represented as c_1. Estimates of these parameters have been derived from the questionnaire responses by computing the probability of continuations b and c in response to Questions 4 and 5, conditional on a decision to write *not* having been made. The resulting estimates are .70 for c_0 and .86 for c_1.

The model assumes that the researcher continues research by returning to the data collection stage, at which point the fate of his research is subject to the r, w, e, and c parameters as before. In carrying out computations based on the model, a three-strikes-and-out rule was assumed. That is, if the researcher has not achieved publication after three data collections, it is assumed that he will abandon the problem. With parameter values estimated for the present system, 62% of lines of investigation are published or abandoned after three attempts in any case. The limitation to three data collections is of little practical importance since the major output indices of the model (see below) change little with additional iterations.

The Figure 1 representation of the model portrays the researcher's choice points as

[4] The .05 level is probably a conservatively low estimate of alpha employed by the researchers to whom questionnaires were sent. In response to a question that asked for an estimate of a level of alpha which "although not satisfactory for rejecting the null hypothesis, would lead you to consider that the null hypothesis is sufficiently likely to be false so as to warrant additional data collection before drawing a conclusion," the mean response was .11 (±.02). This suggests that researchers may be willing to treat "marginally significant" results more like null hypothesis rejections than like nonrejections. Further, the .05 estimate of r_0 is based on the classical hypothesis-testing assumption of an exact null hypothesis, rather than a range null hypothesis, as is employed in the model. The adoption of the range hypothesis framework has the effect of increasing alpha over its nominal level, the extent of the increase being dependent on the width of the null range in relation to the power $(1-\beta)$ of the research. Since full development of this point is beyond the scope of the present exposition, it shall simply be noted that the presently employed estimates of r_0 and r_1 are at best approximate. The estimates actually employed, as derived from the questionnaire responses, are conservative in the sense that they probably err by leading to an overly favorable estimate of system output.

spinners in a game of chance, the parameter values then being represented by the areas in which each spinner may stop. This illustration is intended to make it clear that the model parameters are conditional probabilities, each indicating the probability of a specific departure from a choice point once that choice point has been reached, rather than being an attempted judgment of the research process.

Limitations of the Model

No pretense is made for this model providing anything more than a potentially useful approximation to the research–publication system. Limitations in the accuracy with which some central model parameters can be estimated have already been mentioned. Perhaps the most glaring weakness in the model is its assumption that the probability of editorial acceptance of a report is independent of the sequence of events that precede submission to a journal. The model considers all manuscripts that reject the null hypothesis to be equivalent before the editorial process regardless of the number of data collections in which the null hypothesis was rejected. Similarly, all manuscripts that report acceptance of the null hypothesis are regarded as equivalent. Perhaps even more importantly, the model assumes the editorial process to be insensitive to the actual truth–falsity of the null hypothesis. The performance of the system would be at least a little better, on the various output criteria to be reported, if the model assumed some success of the editorial process in weeding out Type I and Type II errors rather than these having a likelihood of acceptance equal to true rejections and acceptances of the null hypothesis, respectively. These modifications have not been made partly because they would add complexity and also because the elaboration of additional parameters for the editorial process would not seriously affect the relations between the model's input parameters and its ouput indices. (They would affect absolute values of the output indices.)

Note a general caution: The model parameter estimates based on questionnaire responses are certainly more appropriate to some areas of behavioral science research than to others. Particularly, they are appropriate to areas of research in which null hypothesis decision procedures (Rozeboom, 1960) are dominant. Further, given the use of null hypothesis decision procedures, assumptions made about the present state of the system are most appropriate for those areas of research in which measurement error is substantial in relation to the magnitude of theoretically or practically meaningful effects. These are areas in which investigators are prone to work with relatively high risks of Type I error and to proceed otherwise in ways that tend to discriminate against acceptance of the null hypothesis. Within psychology, for example, much research in psychophysics, neuropsychology, and operant behavior would not properly be considered in terms of the present model. On the other hand, much research in social, developmental, experimental, clinical, industrial, and counseling psychology would, I expect, be reasonably well simulated by the model.

Model Output Indices

In order to illustrate how the model's output indices respond to change in model parameters, Figure 2 presents seven output measures as a function of the model parameter h_0 (probability that the null hypothesis is true for the focal hypothesis test). These results have been obtained with model parameters other than h_0 held constant at their previously described values (estimated from questionnaire responses).

If Type I and Type II errors are examined as a percentage of total journal content, it may be seen that these represent a gratifyingly small proportion of total published content (upper portion of Figure 2), given the estimated present-system value of $h_0 = .20$. It then becomes a bit disturbing to note that the Type I error rate of the system (system alpha) is rather high, .30. (System alpha is computed as the proportion of all publications on the right side of Figure 1 that are Type I errors.) System beta (the proportion of all publications on the left side of Figure 1 that are Type II errors) is quite low, .05.

It is somewhat coincidental, but nonetheless remarkable, that the system output levels of

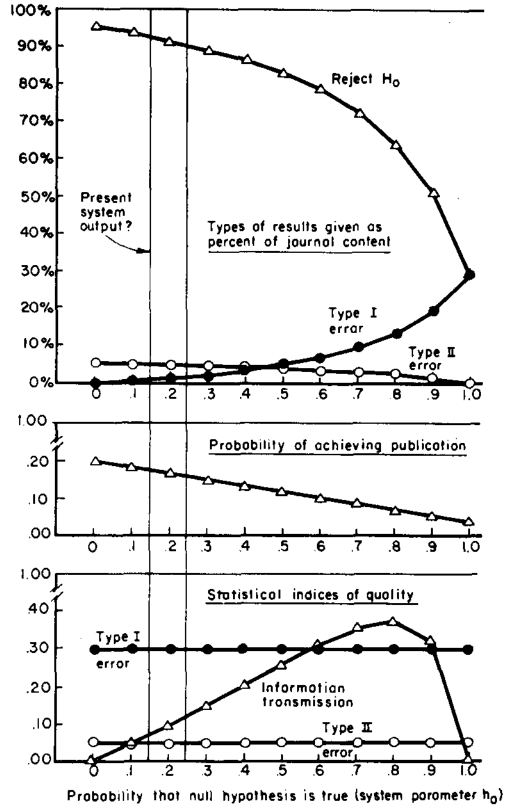

FIGURE 2. Seven output indices for the research-publication system model. (To illustrate responsiveness of output indices to an input parameter, the seven indices are plotted as a function of system parameter h_0 [which equals the probability that the researcher formulates a problem for which the null hypothesis is, in fact, true].)

alpha (.30) and beta (.05) are exactly the reverse of the alpha ($r_0 = .05$) and beta ($1.00 - r_1 = .30$) levels used as estimates of model input parameters. The explanation for the discrepancy between the system alpha and system beta indices, on the one hand, and Type I and Type II errors considered as a percentage of all publications, on the other, can be found in an index giving the percentage of all publications in which the null hypothesis is reported as rejected for the focal hypothesis test (upper portion of Figure 2). This index has the quite high value of 91.5% when $h_0 = .20$. It is apparent then that the high value of system alpha, despite the low proportion of publications that are Type I errors, is a consequence of the fact that system output includes very few publications of true acceptances of the null hypothesis.

An information transmission index. Because it is difficult to interpret the Type I and Type II percentage error indices or the system alpha and beta indices directly as measures of the quality of functioning of the research–publication system, it is desirable to have an index that better summarizes the system's accuracy in communicating information about the truth and falsity of researchers' null hypotheses to journal readers. An information transmission index, computed as shown in Table 2, can partially serve this purpose.

To interpret the information transmission index, assume that a journal reader is presented with a list of the focal hypotheses tested in an upcoming journal issue. Maximally, reading the journal might reduce the reader's uncertainty about the truth–falsity of the several focal hypotheses by an average of 1.00 bit. The information transmission index will approach this maximum value to the extent that (a) there is a fifty–fifty likelihood that the null hypothesis is true or false for the published articles (i.e., the reader's uncertainty is maximal), (b) there is a fifty–fifty true–false reporting ratio for the null hypothesis in the published articles (i.e., the journal's content is maximally uncertain), and (c) the published conclusions are perfectly accurate (or perfectly inaccurate!) regarding the truth–falsity of the null hypothesis. It is important to note that this index bears little direct relation to the percentage of articles reporting a correct result. To appre-

TABLE 2

COMPUTATION OF PUBLICATION INFORMATION TRANSMISSION INDEX

	Published result		
Truth of H₀	Not reject H₀	Reject H₀	Sum
H₀ true	p_{00}	p_{01}	$p_{0.}$
H₀ false	p_{10}	p_{11}	$p_{1.}$
Sum	$p_{.0}$	$p_{.1}$	1.00

Note: Table entries are proportions of only those lines of investigation that have reached the stage of journal publication. The index is computed as (cf. Attneave, 1969, pp. 46 ff):

$$(-\sum_{i=0}^{1} p_{i.}\log_2 p_{i.}) + (-\sum_{j=0}^{1} p_{.j}\log_2 p_{.j}) - (-\sum_{i=0}^{1}\sum_{j=0}^{1} p_{ij}\log_2 p_{ij}).$$

ciate this, consider that a journal may print nothing but correct rejections of the null hypothesis. By definition then, all of its content would be correct. However, the reader who had an advance list of the focal hypotheses of the to-be-published articles would gain *no* information regarding the truth–falsity of any focal null hypothesis from actually reading the journal, since he could know, by extrapolation from past experience, that the null hypothesis would invariably be rejected.

In Figure 2 (lower part), it is apparent that the information transmission index has a very low value (about .10 bits) given the present-system assumption that $h_0 = .20$. The fact that the information transmission index increases dramatically as h_0 increases reflects primarily some virtue in compensating, at the problem formulation stage, for biases against the null hypothesis residing elsewhere in the system.

Comment on the information transmission index. A few of my colleagues have objected to the information transmission index as a summary of system functioning because it takes no account of their primary criterion for evaluating published research—the importance of the problem with which the research is concerned. These colleagues pointed out that archives full of confirmations and rejections of trivial null hypotheses would get high marks on the transmission index but would make for poor science. I am in full sympathy with this view and would not like readers to construe my preference for the information transmission index as a call for journals to catalog trivial results. Thus, it should be emphasized that the information transmission index is insensitive to several possible system virtues. Particularly, (a) it takes no account of the value to readers of the conceptual content of journal articles; (b) it ignores the information contained in tests of nonfocal hypotheses; and (c) by conceptualizing the test of the focal hypothesis as having just an accept–reject outcome, it ignores possible information in the direction or magnitude of effect shown by the focal hypothesis test. Further, the assumption implicit in the index—that readers can be aware in advance of articles' focal hypotheses—is obviously out of touch with reality. Despite these limitations, it is difficult to formulate an index that better summarizes functioning of the research–publication system.

There is an alternative form of the information transmission index that may seem preferable to the one shown in Table 2. This alternate index is based on *all* lines of investigation (not just those that reach the stage of publication) and classifies the outcomes of these lines as published rejection of the null hypothesis, published nonrejection, and also nonpublication. This index has the virtues of (a) summarizing activity in the whole system (rather than just the published portion) and (b) allowing nonpublication to provide information about the truth–falsity of the null hypothesis. Computations have been made for this index, the results indicating system functioning at about as poor a level as does the index described in Table 2. The alternate index has not been presented in Figure 2 chiefly because its implicit assumption—that system output watchers can keep track of lines of research that do not achieve publication—seems too unreasonable.

A final index shown in Figure 2, the probability of achieving a publication given embarcation on a research problem, is one that ought to be of practical concern to researchers. This index, plotted as a function of the h_0 parameter, indicates interestingly that the system "rewards" researchers with publications to the extent that they formulate a problem for which the null hypothesis is false.

A Check on the Model's Accuracy

One means of obtaining a rough check on the model's validity is to compare its predicted proportion of articles for which a focal null hypothesis is accepted against the actual content of the literature. With the assistance of John A. Miller and Karl E. Rosenberg, such a check was made for the *Journal of Personality and Social Psychology* for the year 1972. Every article published that year was read to determine, first, what the focal null hypothesis was and, second, whether the article concluded in favor of acceptance or rejection of that null hypothesis. Out of 199 articles for which a focal null hypothesis was identified, 24 reported acceptance (or nonre-

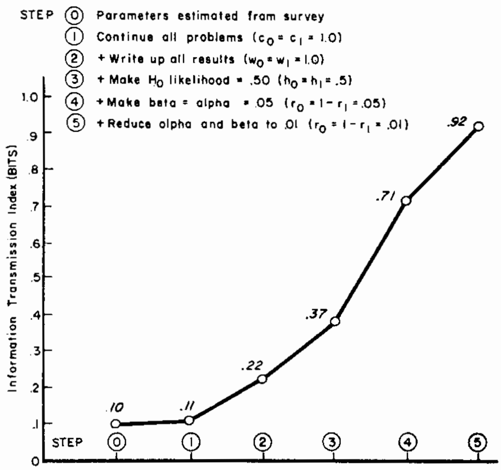

FIGURE 3. Effects on information transmission index of step-by-step alterations in research–publication system parameters to reduce bias against the null hypothesis. (Present-system parameters estimated from survey results are given in the text. Hypothetical changed parameter values are indicated in parentheses in the legend, and characterize also all points to the right of the one in which the change is first indicated.)

jection) of that hypothesis. A 95% confidence interval for the proportion of articles reporting null hypothesis acceptance (12.1% ± 4.5%) included the model's estimated value for the present system of 8.5%, providing some evidence supporting the model's validity. A similar check of four psychological journals in the mid 1950s by Sterling (1959) yielded a lower estimate of 8 out of 294 (which equals 2.7% ± 1.9%) articles that reported nonrejection of a focal hypothesis test. However, it is possible that Sterling may have used a more lenient criterion for declaring that an article rejected the null hypothesis for a focal hypothesis test (cf. Sterling, 1959, pp. 31–32).[5]

Toward a More Satisfactory System

The foregoing results strongly suggest that the research–publication system is functioning well below its potential in research areas characterized by prejudice against the null hypothesis. With the system model it is easy to demonstrate the improvement in system functioning that is potentially possible if biases against the null hypothesis are eliminated. Figure 3 shows the consequences of step-by-step restoration of equal status to the null hypothesis, as reflected in values of the information transmission index. It is quite apparent from Figure 3 that unbiased behavior at the various stages of the research–publication process can have highly desirable effects on the informativeness of published research. The methods of achieving such unbiasedness are considered in more detail below.

System Effect on Generality of Research Findings

The information transmission and other system output indexes are insensitive to what may be the worst consequence of prejudice against the null hypothesis—the archival accumulation of valid results with extremely limited generality.

Consider the situation of the researcher who starts off with the hypothesis that an increase in variable *x* produces an increase in variable *y*. Since he is very convinced of the virtues of the theory that led to this prediction, he is willing to proceed through a number of false starts and pilot tests during which he uses a few different experimenters to collect data, a few different methods of manipulating variable *x*, a few different measures of variable *y*, and a few different settings to mask the true purpose of the experiment. At last, he obtains the result that confirms the expected impact of *x* on *y* but is properly concerned that the result may have been an unreplicable Type I error. To relieve this concern, he conducts an exact replication using the same combination of experimenter, operationalization of *x*, measure of *y*, and experimental setting that previously "worked," and is gratified to discover that his finding is replicated. Concerned about the validity of his procedures and measures, he also obtains evidence indicating that the manipulation of *x* was perceived as intended and that the measure of *y* had adequate reliability and validity. He then publishes his research concluding that increases in *x* cause increases in *y* and, therefore, that his theory, which predicted this relationship, is supported.

[5] It gives me pause, in reading over this paragraph, to consider whether or not I would have reported the results of the content check of *JPSP* if it had not been confirming of the model.

The potential fault in this conclusion should be obvious from the way I have presented the problem, but it is not likely to be obvious to the researcher conducting the investigation. Because of his investment in confirming his theory with a rejection of the null hypothesis, he has overlooked the possibility that the observed x–y relationship may be dependent on a specific manipulation, measure, experimenter, setting, or some combination of them. In his eagerness to proclaim a general x–y relationship, he has been willing to attribute his previous false starts to his (or, better, his research assistants') incompetence and, on this basis, does not feel it either necessary or desirable to inform the scientific community about them.

This style of researcher's approach has been well described by McGuire (1973):

> The more persistent of us typically manage at last to get control of the experimental situation so that we can reliably demonstrate the hypothesized relationship. But note that what the experiment tests is not whether the hypothesis is true but rather whether the experimenter is a sufficiently ingenious stage manager. (p. 449)

For further discussion of situations in which findings of limited generality appear to be much more general, I refer the reader to Campbell's (1969, pp. 358–363) typology of threats to valid inference.

Some Epidemics of Type I Error

If the results generated by the model are to be believed, then the existing archival literature in the behavioral sciences should contain some blatant Type I errors. Although the absolute frequency of Type I error publications is not expected to be high, there should be some true null hypotheses for which only rejections of the null hypothesis have been published. About the only way to demonstrate the existence of Type I errors conclusively is to demonstrate that "established" findings cannot be replicated and that such failures to replicate cannot easily be regarded as Type II errors. As mentioned before, the fact that two of the three following cases are drawn from social psychology reflects only my familiarity with this field, not any belief that social psychology is more prone to such errors than are other areas of behavioral science research.

Attitude and Selective Learning

Between 1939 and 1958, approximately 10 studies (referenced in Greenwald & Sakumura, 1967) reported the consistent finding that subjects, when exposed to information on a controversial topic, more easily learned information that was agreeable rather than disagreeable to their existing attitude on the issue. This selective learning effect was regarded as sufficiently established to appear in many introductory psychology and social psychology textbooks, the study of Levine and Murphy (1943) particularly being regarded as somewhat of a classic.

Starting in 1963, however, almost all published studies that included a test of this hypothesis failed to confirm it (Brigham & Cook, 1969; Fitzgerald & Ausubel, 1963; Greenwald & Sakumura, 1967; Waly & Cook, 1966). In one study (Malpass, 1969) the hypothesis appeared to be confirmed in only one of three conditions in which it was tested. In general, the experiments reported since 1963 have been quite carefully done, each publication typically reporting the results of more than one replication of the hypothesis test and with careful attempts to control extraneous variables that might contaminate the tests. Therefore it does not seem reasonable to suggest that these recent findings should be regarded uniformly as Type II errors. Because the recent investigations have also made strenous attempts, generally unsuccessful, to explain the earlier findings in terms of interactions with previously uncontrolled factors, the possibility that most of the earlier results were Type I errors is currently very plausible. This apparent epidemic of Type I error can be readily understood in terms of the hypothesized present research–publication system. Several of the earlier publications reported rejections of the null hypothesis with an alpha criterion greater than .05. After the selective learning effect had thus established some precedent in the literature, presumably researchers and editors were more disposed to regard a rejection of the null hypothesis as true than false. Possibly, also, investigators who could not obtain the established finding were content to regard their experiments as inadequate in some respect or other and did not even bother to seek publication for what they may have believed to be Type II errors, nor did they bother to conduct further research that might have explained their failure to replicate published findings.

The Sleeper Effect

The sleeper effect in persuasion is said to occur when a communication from an untrustworthy or inexpert source has a greater persuasive impact after some time delay than it does on original exposure. That is, the communication presumably achieves its effect while the audience "sleeps" on it. This result is established well enough so that it is described in most introductory social psychology texts. The research history of the sleeper effect demonstrates a variety of ways in which Type I publication errors may occur (if one assumes, that is, that the effect is not a genuine one).

The original report of a sleeper effect by Hovland, Lumsdaine, and Sheffield (1949) involved the use of an alpha criterion that was inflated by selective sampling from multiple post hoc tests of the hypothesis. That is, the effect was not predicted and was found on only a subset (not an a priori one) of the opinion items used by the investigators. In subsequent years, experimental investigators have chosen to look for the sleeper effect in terms of a comparison between the temporal course of opinion changes induced by the same communication from a trustworthy, versus an untrustworthy, source. That is, the increase in effect over time with the untrustworthy source should not be matched by a similar increase when the source is trustworthy. Significant interaction effects involving these two variables of source credibility and time since communication have been reported in a number of studies (e.g., Gillig & Greenwald, 1974; Hovland & Weiss, 1951; Kelman & Hovland, 1953; Shulman & Worrall, 1970; Watts & McGuire, 1964). However, in *none* of these studies was there reported a significant increase in impact, with passage of time since the communication, for subjects receiving the communication from an untrustworthy source. That is, the interaction effects were due primarily or entirely to loss of effects, with passage of time, for subjects receiving the communication from a trustworthy source.

The sleeper effect, it is clear, was established in the literature by a series of studies, each of which employed an ostensible alpha criterion of .05 but for which the effective alpha criterion was substantially higher. In the original Hovland et al. (1949) study, alpha was inflated through the selective reporting of post hoc significance tests; in the later studies, it was inflated by use of an inappropriate overall interaction effect test instead of the simple effect of the time variable within the untrustworthy source conditions. Evidence that the original and subsequent sleeper effect reports are likely to have been Type I errors has come recently from a series of seven investigations by Gillig and Greenwald (1974) involving a total of 656 subjects. With their procedures, a true sleeper effect (increase over time) of .50 points on the 15-point opinion measure they used would have been detected with better than .95 probability. A 95% confidence interval (±.27 scale points) around the observed mean change of +.14 clearly included the hypothesis of zero change.

Quasi-Sensory Communication

A perennially interesting subject for behavioral science research concerns the possibility of perception of events that provide no detectable inputs to known sensory receptors. Research on extrasensory perception or quasi-sensory communication (Clement, 1972; McBain, Fox, Kimura, Nakanishi, & Tirada, 1970) is so plagued with research–publication system problems that no reasonable person can regard himself as having an adequate basis for a true–false conclusion. This state of affairs is not due to lack of research. It would be difficult to estimate, on the basis of the published literature, the amount of research energy that has been invested in parapsychological questions, and this is precisely the problem. It is a certainty that the published literature, both in favor of and against quasi-sensory communication, represents only a small fraction of the total research effort. Two anecdotes in my own experience are illustrative:

1. A physicist at Ohio State University once described to me an investigation, conducted with a colleague as a digression at their laboratory, into the detection of human-expressed affect by plants.[6] They happened to have electronic apparatus of sufficient accuracy to detect electric potential charges of as small a magnitude as 10 nV between two points on the same or opposite surfaces of a leaf. This was approximately one part in 10^7 of the baseline voltage. They failed to detect responses of this magnitude reliably in a number of tests involving verbally and facially communicated threats to the plant. When I learned of this (at a cocktail party, of course) I asked if they had any intention either to publish their results or to repeat the experiment. The reply was negative, although I expect the scientific community would have been informed had their results been positive.

2. As an editorial consultant to a journal, I was asked to review an article that obtained an extrasensory perception effect that would reject the hypothesis of no effect if alpha were set at .10. I advised the editor that the result was one that had a higher probability of being Type I error than the ostensible .10, but the appropriate editorial response, since the study was competently done and the problem was interesting, was to guarantee to publish the results if the investigators would agree (a) to conduct a replication and (b) to publish the outcome of the replication (as well as the already submitted study). Two years later, the study was published (Layton & Turnbull, 1974; see also Greenwald, 1974) with the results of the replication *failing* to confirm the original findings.

Now we all know that anecdotes are unacceptable as scientific evidence because of the inflated probability that unusual events will be noticed and propagated as anecdotes. What is distressing is that the published literature on quasi-sensory communication (and other topics) also seems to be highly likely to detect and communicate relatively unlikely events. As it is functioning in at least some areas of behavioral science research, the research–publication system may be regarded as a device for systematically generating and propagating anecdotal information.

Rational Strategies Regarding the Null Hypothesis

My criticisms of researchers' null-hypothesis-related strategies are not new. They have been expressed, in part, by several previous writers, the article by Bakan (1966) being perhaps closest to the approach I have taken. The point that Type I publication errors are underestimated by reported alpha criteria has been made also in critiques of the use of significance tests in sociology (Selvin, 1966) and psychology (Sterling, 1959) (see also the anthology edited by Morrison & Henkel, 1970). What I have attempted to add to the previous critiques is a quantitative assessment of the magnitude of the problem as it exists, by means of (a) a questionnaire survey and (b) a system simulation employing system parameters derived from the survey results. The obtained quantitative estimates must be regarded as frightening, even calling into question the scientific basis for much published literature.

Previous critics have not been negligent in suggesting remedies for what they too have regarded as an undesirable situation. Some suggestions have been intended for use in conjunction with the standard significance testing approach. For example, Cohen (1962) has pointed out that social psychological experiments often have power adequate to detect only relatively large effects. His suggestion for higher powered experiments, if adopted, should be expected to result in an increase in the frequency of null hypothesis rejections relative to nonrejections. However, it is also possible that increased awareness of experimental power may lead to taking null results more seriously. Hays (1963) has suggested using estimates of magnitude of association to accompany the standard reports of alpha levels. This would help to assure that trivial effects associated with a rejection of

[6] I am grateful to James T. Tough and James C. Garland for permission to give this informal report of their investigation.

the null hypothesis would be recognized as such, but might have no systematic effect on the treatment of null results.

Other writers have recommended departures from the significance testing framework. Particularly, suggestions for the use of interval estimation (Grant, 1962) or Bayesian analytic techniques (Bakan, 1966; Edwards, Lindman, & Savage, 1963) would help to avoid prejudice against the null hypothesis, because with these procedures, results need not be stated in terms of acceptance or rejection of a null hypothesis. Despite the good reasons for using interval estimation and Bayesian techniques that have been advanced by several writers, inspection of current journals makes it apparent that tests of significance against point null hypotheses remain the predominant mode of analysis for behavioral data. (Further, there is little evidence that behavioral researchers have given increased attention to the power of their research designs or to magnitude of association, in pursuit of the suggestions by Cohen, Hays and others.)

It would be a mistake, I think, to expect that a recommendation to adopt some analysis strategy other than (or in addition to) significance testing might, by itself, eliminate bias against accepting the null hypothesis. This is because, as has been shown here, *the problem exists as much or more in the behavior of investigators before collecting and after analyzing their data as in the techniques they use for analysis.* Further, since a research enterprise may often be directed quite properly at the determination of whether a given relationship or effect does or does not approximate a zero value, it seems inappropriate to urge the dropping of methods of analysis in which null hypotheses are compared with alternatives. As noted earlier, a research question stated in null hypothesis versus alternative hypothesis form is especially appropriate for theory-testing research. In such research, a result that can be used to accept a null hypothesis may often serve to advance knowledge by disproving the theory.

For these reasons, *my basic recommendation is a suggested attitude change of researchers (and editors) toward the null hypothesis. Support for the null hypothesis must be regarded as a research outcome that is as acceptable as any other.* I cannot leave this recommendation just baldly stated because I suspect that most readers will not know how to go about analyzing and reporting data in a fashion that can lead to the acceptance of the null hypothesis. I conclude, therefore, by considering some technical points related to acceptance of the null hypothesis. It should be clear to readers, as it is to me that what follows is a rather low-level consideration of technical matters, directed at users around my own level of statistical naiveté but nonetheless accurate as far as I can determine through consultation with more expert colleagues.

How to Accept the Null Hypothesis Gracefully

Use a range, rather than a point, null hypothesis. The procedural recommendations to follow are much easier to apply if the researcher has decided, in advance of data collection, just what magnitude of effect on a dependent measure or measure of association is large enough not to be considered trivial. This decision may have to be made somewhat arbitrarily but seems better to be made somewhat arbitrarily before data collection than to be made after examination of the data. The minimum magnitude of effect that the researcher is willing to consider nontrivial is then a boundary of the null range. The illustrations that follow employ a "two-tailed" null range that is symmetric around the zero point of a test statistic.

Select N *on the basis of desirable error of estimate of the test statistic.* Assume, for example, that in an experiment with one treatment condition and a control condition, the researcher had decided that a treatment versus control difference of .50 units on his dependent measure is a minimum nontrivial effect. (Therefore, the null range is (−.50, +.50) on this measure). It would seem inappropriate to collect data with N only large enough so that the estimate of the treatment effect would have a standard error of, say, .50. To appreciate this, consider that a 95% confidence interval based on this imprecise an estimate would encompass about twice the width of the null range. I can think of no hard and fast way of specifying a desirable degree of precision, but I would suggest that an error of estimate of effect on the order of 10%–20% of the width of the null range may often be appropriate. (A 95% confidence interval then would be 40%–80% of the null range's width.) More precision than this may often be desirable, but the researcher has to make such decisions based on the cost of obtaining such precision relative to the value of the knowledge obtained thereby.[7]

Have convincing evidence that manipulations and measures are valid. Whether the data are to be used to accept or reject a null hypothesis or to make some other conclusion, it seems essential that the researcher be able to document the validity of his procedures relative to the conceptual variables being studied. In the case of accepting the null hypothesis, the results are patently useless if the researcher has not defended himself against the argument that his operations lacked correspondence with the variables that were critical to his hypothesis test. However, the researcher drawing a conclusion that rejects a null hypothesis should feel equal compulsion to demonstrate that his procedures were valid.

Compute the posterior probability of the null (range) hypothesis. I refer the reader to statistical texts (e.g., Hays, 1973, chap. 19; Mosteller & Tukey, 1969, pp. 160–183; Phillips, 1973) for an introduction to Bayesian methods (see also Edwards et al., 1963). Figure 4 offers a comparison of three modes of analysis—significance testing, interval estimation, and Bayesian posterior probability computation—for some hypothetical data. These hypothetical data are for the difference between two correlated means on a measure for which the researcher's null range is (−.50, +.50). The standard error of the difference scores is assumed to be 1.00, and the obtained sample mean difference (M_D) is +.25, a point clearly within the null range. Each analysis method is presented for three sample sizes as an aid to comparing the different analysis procedures.

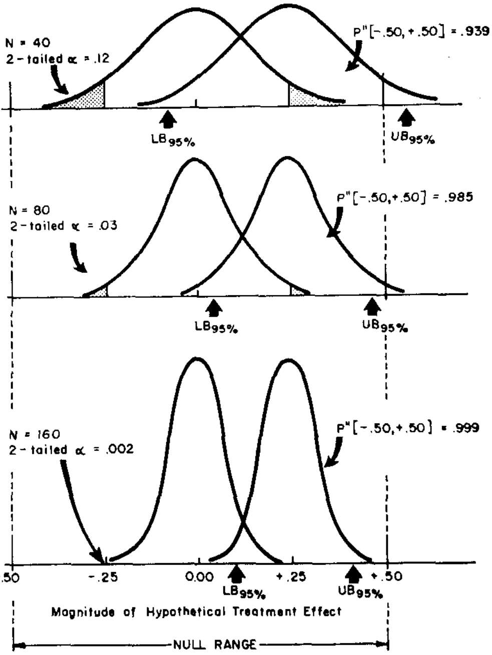

FIGURE 4. Comparison of significance testing, confidence interval estimation, and posterior probability estimation for three sample sizes. (The example assumes a null hypothesis range of (−.50, +.50), a variance of 1.00, and an obtained sample estimate of +.25 for a hypothetical treatment effect. The distribution centered over 0.00 is the expected distribution of sample mean estimates of the effect if the point null hypothesis of 0.00 is true. This is used to compute significance levels [αs]. The distribution centered over +.25 is the Bayesian posterior likelihood distribution, the posterior probability [P″] estimate being computed as the fraction of the area under this distribution falling in the interval (−.50, +.50). LB and UB are lower and upper confidence interval boundaries.)

The first analysis employs a standard two-tailed significance test for a *point* (not range) null hypothesis. This would seem to be the analysis currently preferred by most behavioral scientists. At $\alpha = .05$, this analysis does not reject the null (point) hypothesis for the smallest sample size shown but does do so for the two larger sample sizes, despite (a) the observed data point being well within the null range and (b) the fact that with the larger sample sizes we should have more confidence in the accuracy of this estimate.

[7] Setting N to achieve a given level of precision requires some advance estimate of variability of the data. If such information is unavailable at the outset of data collection, it may then be necessary to determine this variability on the basis of initial data collection.

Clearly, computation of the significance level of an obtained result relative to an exact null hypothesis is not a useful way of going about accepting a range null hypothesis. With a relatively large N it is, rather, a good means of exercising prejudice against the null hypothesis.

The second analysis shown in Figure 4 presents 95% confidence intervals for the $M_D = +.25$ result for the three sample sizes. If we consider the containment of the 95% confidence interval within the null range as a criterion for accepting the null hypothesis, then we should accept the null hypothesis for the two larger sample sizes. This is definitely an improvement over the significance test analysis, but it still has some drawbacks. Particularly, (a) we are at a loss to make direct use of the data for the smallest sample size, for which the 95% confidence interval overspreads the null range; and (b) the conclusion does not reflect the increase in confidence that should be associated with the result for $N = 160$ relative to that for $N = 80$. It is apparent that these drawbacks of the confidence interval procedure stem from the awkwardness of relating the interval estimation procedure to a decision relative to the null hypothesis (cf. Mosteller & Tukey, 1969, pp. 180–183).

The final procedure illustrated in Figure 4 involves the computation of posterior likelihood distributions based on the obtained data. When in a Bayesian analysis one starts from ignorance (a "diffuse," "uniform," or "gentle" prior likelihood distribution), the posterior likelihood distribution is constructed directly from the mean and variability of the obtained data, much as is a confidence interval. A critical difference from the confidence interval analysis is that the assumptions underlying the Bayesian analysis facilitate drawing a conclusion about the acceptability of the null hypothesis. For the posterior distributions presented in Figure 4, a uniform prior distribution is assumed. The resulting posterior probability statements have the desirable feature of allowing us to conclude that the (range) null hypothesis is considerably more likely than its complement for all three sample sizes, while at the same time allowing expressions of the increased certainty afforded by the larger sample sizes for the $M_D = +.25$ result.

To provide a more concrete illustration of a posterior probability computation used as the basis for accepting a null hypothesis, consider the data from the Gillig and Greenwald (1974) sleeper effect study described in the earlier section on Type I errors. In this study, Gillig and Greenwald were employing a 15-point opinion scale as the dependent measure. They considered that a change from an immediate posttest to a delayed posttest of less than .50 on this scale was a trivial effect (.50 was less than 25% of the standard deviation of the obtained difference scores). They employed 273 subjects to estimate this change, so that the standard error of their estimate of the effect was .134 (which equals 13.4% of the (−.50, +.50) null range). Computation of the posterior likelihood distribution of the hypothesis, assuming a uniform prior distribution, indicated that 99.6% of the area under the posterior distribution was within the null range. The .996 figure can therefore be taken as a posterior probability measure of acceptance of the null (range) hypothesis for these data. This figure can be expressed alternately as a posterior odds ratio of .996/(1 − .996) = 249:1 in favor of the null hypothesis. For comparison, an odds ratio of 19:1 ($\alpha =$ .05) is frequently considered "significant" in rejecting a point null hypothesis (as contrasted with all possible alternatives).

Report all results of research for which conditions appropriate to testing a given hypothesis have been established. As has been demonstrated earlier, successful communication of information through archival publication is severely threatened by self-censorship on the part of investigators who obtain unpredicted (often meaning null) findings. The only justifiable basis for withholding a report of the results of a data collection should be that the hypothesis intended for testing was not actually tested. This could come about through failures of manipulation, measurement, randomization, and so forth. As previously noted the investigator should be prepared for these possibilities, meaning that he should be able to support a decision to withhold data by demonstrating that such an invalidating condition obtained. Given a valid hypothesis test, the only justifiable procedure for reporting less than all of the data obtained is the decidedly dubious one of discarding portions of the data randomly; any nonrandom decision procedure with widespread application would result in publications being a biased sample of actual research results. Therefore, researchers should make a point of including at least brief mentions of findings of preliminary data collections, explaining why these results have been ignored (if they have), in reports of data on which more final conclusions have been based. It will be obvious that the admonition to publish all one's data fails to take into account the reality of editorial rejection. This point prompts a few final comments. First, it is a truly gross ethical violation for a researcher to suppress reporting of difficult-to-explain or embarrassing data in order to present a neat and attractive package to a journal editor. Second, it is to be hoped that journal editors will base publication decisions on criteria of importance and methodological soundness, uninfluenced by whether a result supports or rejects a null hypothesis.

Conclusions

As has, I hope, been clear there is a moral to all this. In the interest of making this moral fully explicit (and also for the benefit of the reader who has started at this point), I offer the following two boiled-down recommendations.

1. Do research in which any outcome (including a null one) can be an acceptable and informative outcome.

2. Judge your own (or others') research not on the basis of the results but only on the basis of adequacy of procedures and importance of findings.[8]

[8] Concluding note: Although I have not had occasion to cite their work directly in this report, the articles of Binder (1963), Campbell and Stanley (1963), Lykken (1968), and Walster and Cleary (1970) have stimulated some of the ideas developed here.

References

Aronson, E., & Carlsmith, J. M. Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (2nd ed., Vol. 2), Reading, Mass: Addison-Wesley, 1969.

Attneave, F. *Applications of information theory to psychology*. New York: Holt, 1959.

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

Binder, A. Further considerations of testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review,* 1963, *70*, 107–115.

Brigham, J. C., & Cook, S. W. The influence of attitude on the recall of controversial material: A failure to conform. *Journal of Experimental Social Psychology,* 1969, *5*, 240–243.

Campbell, D. T. Prospective: Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral reseach.* New York: Academic Press, 1969.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963.

Clement, D. E. Quasi-sensory communication: Still not proved. *Journal of Personality and Social Psychology,* 1972, *23*, 103–104.

Cohen, J. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,* 1962, *65*, 145–153.

Edwards, W. Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin,* 1965, *63*, 400–402.

Edwards, W., Lindman, H., & Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review,* 1963, *70*, 193–242.

Festinger, L. Laboratory experiments. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences.* New York: Holt, 1953.

Fitzgerald, D., & Ausubel, D. P. Cognitive versus affective factors in the learning and retention of controversial material. *Journal of Educational Psychology,* 1963, *54*, 73–84.

Gillig, P. M., & Greenwald, A. G. Is it time to lay the sleeper effect to rest? *Journal of Personality and Social Psychology,* 1974, *29*, 132–139.

Grant, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review,* 1962, *69*, 54–61.

Greenwald, A. G. Significance, nonsignificance, and interpretation of an ESP experiment. *Journal of Experimental Social Psychology,* 1974, *10*, in press.

Greenwald, A. G., & Sakumura, J. S. Attitude and selective learning: Where are the phenomena of yesteryear? *Journal of Personality and Social Psychology,* 1967, *7*, 387–397.

Hays, W. L. *Statistics for psychologists.* Nev York: Holt, Rinehart & Winston, 1963.

Hays, W. L. *Statistics for social scientists* (2nd ed.). New York: Holt, Rinehart & Winston, 1973.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. *Experiments on mass communication.* Princeton: Princeton University Press, 1949.

Hovland, C. I., & Weiss, W. The influence of source credibility on communication effectiveness. *Public Opinion Quarterly,* 1951, *15,* 635–650.

Kelman, H. C., & Hovland, C. I. "Reinstatement" of the communicator in delayed measurement of opinion change. *Journal of Abnormal and Social Psychology,* 1953, *48,* 327–335.

Layton, B. D., & Turnbull, B. Belief, evaluation, and performance on an ESP task. *Journal of Experimental Social Psychology,* 1974, *10,* in press.

Levine, J. M., & Murphy, C. The learning and forgetting of controversial material. *Journal of Abnormal and Social Psychology,* 1943, *38,* 507–517.

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

Malpass, R. S. Effects of attitude on learning and memory: The influence of instruction induced sets. *Journal of Experimental Social Psychology,* 1969, *5,* 441–453.

McBain, W. N., Fox, W., Kimura, S., Nakanishi, M., & Tirado, J. Quasi-sensory communication: An investigation using semantic matching and accentuated effect. *Journal of Personality and Social Psychology,* 1970, *14,* 281–291.

McGuire, W. J. Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 1). New York: Academic Press, 1964.

McGuire, W. J. The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology,* 1973, *26,* 446–456.

Mills, J. The experimental method. In J. Mills (Ed.), *Experimental social psychology.* Toronto: Macmillan, 1969.

Morrison, D. E., & Henkel, R. E. (Eds.). *The significant test controversy.* Chicago: Aldine, 1970.

Mosteller, F., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (2nd ed., Vol. 2). Reading, Mass: Addison-Wesley, 1969.

Phillips, L. D. *Bayesian statistics for social scientists.* New York: Crowell, 1973.

Platt, J. R. Strong inference. *Science,* October 1964, pp. 347–353.

Rosenthal, R., & Rosnow, R. L. (Eds.). *Artifact in behavioral research.* New York: Academic Press, 1969.

Rozeboom, W. R. The fallacy of the null-hypothesis significance test. *Psychological Bulletin,* 1960, *57,* 416–428.

Selvin, H. C., & Stuart, A. Data-dredging procedures in survey analysis. *American Statistician,* 1966, *20,* 20–23.

Shulman, G. I., & Worrall, C. Salience patterns, source credibility, and the sleeper effect. *Public Opinion Quarterly,* 1970, *34,* 371–382.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

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. *American Statistician,* 1970, *24,* 16–19.

Waly, P., & Cook, S. W. Attitudes as a determinant of learning and memory: A failure to confirm. *Journal of Personality and Social Psychology,* 1966, *4,* 280–288.

Watts, W. A., & McGuire, W. J. Persistency of induced opinion change and retention of the inducing message contests. *Journal of Abnormal and Social Psychology,* 1964, *68,* 233–241.

Wilson, W. R., & Miller, H. A note on the inconclusiveness of accepting the null hypothesis. *Psychological Review,* 1964, *71,* 238-242.

(Received February 18, 1974)