JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY 11, 180–191 (1975)

Significance, Nonsignificance, and Interpretation of an ESP Experiment[1]

ANTHONY G. GREENWALD[2]

Ohio State University

Received February 13, 1974

The data of B. D. Layton and B. Turnbull's (1974) two extra sensory perception (ESP) experiments are used to illustrate Bayesian hypothesis tests that provide more useful information than is obtained from significance tests. In introducing these reanalyses, it is noted that the use of significance tests has tended unreasonably to foster negative evaluation of "nonsignificant" findings relative to "significant" findings. The present Bayesian hypothesis tests indicate that Layton and Turnbull's findings should mostly be taken as support for their null hypotheses (no ESP effects) rather than for the weakest alternatives that their experiments had reasonable power to detect.

The reader may recognize the preceding article by Layton and Turnbull (1974) as somewhat unusual content for a social science journal, not only for its ESP subject matter, but also for its nonsignificant results. The Layton–Turnbull article therefore provides a context for considering some of the issues related to the approach to "nonsignificant" results. In this comment, I shall (a) attempt to dispel some myths that seem to be at the root of the dislike for nonsignificant results, and (b) give an example, using Layton and Turnbull's data, of a method of analysis that permits extraction of more meaning or significance from "negative" results than do our currently fashionable tests of significance. I should make clear at the outset that my illustrative use of the Layton–Turnbull study should not be construed as a criticism of their study. In fact, I find their study important (I am eschewing "significant" for the moment) and I hope, further, to enhance its contribution by means of the additional analyses I shall report.

Some misconceived prejudices against the null hypothesis

Consider the following observations on nonsignificant results.

(1) A null finding is only a basis for uncertainty. Conclusions about relationships among variables should be based only on rejections of null hypotheses.

[1] This report was supported in part by PHS Grant MH-20527-02.

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

(2) Little is gained 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.

Although the above statements can find some support in published works on methodology (some partial sources for them are Aronson and Carlsmith (1969, p. 21); Festinger (1953, pp. 142–143); Mills (1969, pp. 442–448); Wilson and Miller (1964)), I think they can be convincingly refuted. I shall indicate these refutations only briefly here (the numbers below correspond to the numbers of the above statements, see also Greenwald (1973)).

(1) To the extent that you can't prove the null hypothesis because it is an exact (point) hypothesis, neither can you prove any nonnull exact hypothesis. Statistical techniques not based on the test of significance (see below for more on this) do allow ways of describing acceptability of null hypotheses.

(2) As Platt (1964) has convincingly argued, scientific advance is often most powerfully achieved by rejecting theories (as distinct from rejecting null hypotheses). A major strategy for rejecting a theory is to demonstrate that relationships predicted by the theory are not obtained, and this may often mean accepting a null hypothesis.[3]

(3) A significant result indicates (barring Type I error) only that some relationship or effect was observed. It does not show what variables were related. The researcher who would claim that a significant result in-

[3] To spell this out just a bit more: Any number of consistent results may support a theory, but do not prove it. Just one solid result that is inconsistent with a theory serves to disprove the theory and indicates both the need for a superior formulation and at least one new result that must be accommodated by the new theory. A classic example in physics is the experimental disproof, by Michelson and Morley (1887), of the hypothesis that light is propagated by a medium ("ether") that is at rest in relation to the orbital movement of the earth. The Michelson–Morley experiment showed that the speed of light did not vary as a function of the cardinal direction in which it was measured; had the light been propagated by a medium at rest relative to the earth, the speed of light should have been affected (differently in different directions) by variations in the relative motion of the hypothesized medium. In psychology, Thorndike's (1911) law of effect, which stipulated an automatic association-strengthening effect of rewards, has been gradually discredited as a principle of human learning by the accumulation of results (Buchwald, 1969; Estes, 1969; Nuttin and Greenwald, 1968) favoring a null hypothesis of no special connection-strengthening effect of rewards. In the place of Thorndike's automatic-effect theory, cognitive interpretations (see articles just cited) of reward effects are now being employed.

dicates a relationship between given conceptual variables X and Y should be as clearly obliged to show that his experimental operations corresponded to those conceptual variables as should the researcher who would claim the absence of a relationship.

(4) It is, indeed, possible to obtain null results through incompetence, particularly by using weak experimental or measurement operations. However, the most common types of incompetence (experimenter bias, inappropriate demand characteristics, subject self-selection, sampling fault, improper data analysis) are considerably more likely to result in erroneous "significant" results than in nonsignificance.

Consequences of prejudice against the null hypothesis

In another paper (Greenwald, 1973) I have presented the results of a survey of research practices of social psychologists. The survey results indicated the existence of several behavioral biases against null results. The strongest bias was in the reported probability of submitting (for publication) a report of research that rejected a null hypothesis ($\bar{P} = .48$) vs one that did not ($\bar{P} = .06$). This was perhaps related to the fact that investigators tended to identify their personal predictions with a rejection (rather than a nonrejection) of the null hypothesis with a mean probability of .80. Additional bias against the null hypothesis was evident in a greater reported probability of abandoning research on a problem when initial results indicated nonrejection (rather than rejection) of the null hypothesis.

With some reasonable assumptions about operation of the research-publication system (Greenwald, 1973), it can be demonstrated that these biases produce a state of affairs in which (a) very little is expected to be published on problems for which the null hypothesis is, to a reasonable approximation, true; (b) a distressingly high proportion of publications on problems for which the null hypothesis is true will erroneously report rejection of the null hypothesis; and (c) published rejections of null hypotheses are apt to be of very much less generality than is claimed for them.

To remedy this situation, it is important for researchers (and editors) to become aware that null results may often be very valuable in terms of advancing knowledge, and also to avoid biased practices that lead to selective publication of null-hypothesis rejections (such practices are catalogued in Greenwald (1973)). These remedies, although necessary, are not sufficient. A remaining barrier to overcome is the fact that our ritualistically employed statistical tests of significance make it considerably more convenient for us to characterize a set of results as rejecting, rather than accepting, a null hypothesis. Layton and Turnbull provide an illustration of the difficulty of dealing with nonsignificant results when they remark, ". . . we are left with no alternative but to consider these

studies *inconclusive* regarding the effects of the experimental manipulations of ESP performance" (Layton and Turnbull).

The argument that test-of-significance analyses represent an existing paradigm that may be unduly restrictive on interpretation is hardly new (cf. (Bakan, 1966; Edwards, Lindman, and Savage, 1963; Morrison and Henkel, 1970)). An advantage of some of the more recently developed Bayesian analytic methods (see (Edwards *et al.*, 1963; Hays, 1973, Chap. 19; or Mosteller and Tukey, 1969, pp. 160–183) for an introduction to Bayesian procedures) is that they make it just as easy to describe the acceptability of a null hypothesis as of any other hypothesis. The remainder of this comment will illustrate the application of alternative analysis procedures, including Bayesian hypothesis tests, to Layton and Turnbull's data.

Brief summary of the Layton–Turnbull study

Layton and Turnbull conducted two near-identical experiments, the first involving 179, the second 235 college subjects. Two orthogonally manipulated independent variables were *Belief* (in the existence of ESP or not) and *Evaluation* (of ESP as good or harmful). Belief was manipulated by having the experimenter describe a past series of *successes* (or *failures*) in attempts to demonstrate ESP and, also, state that he personally *believed* (or *disbelieved*) in ESP and was confident that it *would* (or *would not*) be demonstrated in the present experiment. The Evaluation manipulation was accomplished by the experimenter's enumeration of either possible beneficial or harmful effects of the use of ESP. The ESP task then consisted of the subject's attempting to reproduce an ordered 100-item list of the digits 1, 2, 3, 4, and 5. The list was contained in a sealed envelope handed to each subject, with adequate precaution against cheating. The expected number of hits by chance on this test was 20, so that clairvoyance could be evidenced by either upward or downward deviations from this expectation.

Layton and Turnbull analyzed their data in a $2 \times 2 \times 2$ factorial analysis of variance design (Belief X Evaluation X Sex of Subject). In Experiment I, of seven significance tests obtained from this 3-factor design, the only one significant at an $\alpha = .05$ criterion was an interaction of Belief X Sex. This finding indicated that the expected facilitating effect of belief in ESP on performance was obtained for females, but was unexpectedly reversed for males (and nonsignificant for the sample as a whole). A marginally significant finding ($p < .06$) was in the predicted direction of better performance in the ESP-beneficial ($X = 20.66$) than the ESP-harmful ($X = 19.49$) condition of the Evaluation factor. In Experiment II there were no significant effects at all. The means of the basic 2×2 design (Belief X Evaluation) for the two experiments are presented in Table 1.

TABLE 1
MEAN DEVIATION FROM CHANCE EXPECTATION ON THE MATCHING TASK

Evaluation	Belief								
	Experiment I			Experiment II			Combined experiments		
	Sheep	Goat	Ave.	Sheep	Goat	Ave.	Sheep	Goat	Ave.
Positive	+.51	+.82	+.66	.00	+.15	+.08	+.22	+.41	+.32
	(39)	(34)	(73)	(51)	(53)	(104)	(90)	(87)	(177)
Negative	−.81	−.18	−.51	+.10	+.08	+.09	−.28	−.03	−.16
	(36)	(33)	(69)	(50)	(48)	(98)	(86)	(81)	(167)
Average	−.12	+.33	+.09	+.05	+.12	+.08	−.02	+.20	+.09
	(75)	(67)	(142)	(101)	(101)	(202)	(176)	(168)	(344)

Note. Table entries are mean deviations from the chance expectation of 20 correct responses out of 100 trials (n's in parentheses). For the control condition in Experiment I, the mean deviation was −.38 ($n = 37$) and in Experiment II the control mean was +.70 ($n = 33$). The average within cells error mean square was 12.70 for Experiment I and 19.13 for Experiment II.

Interpreting the findings

Because significance tests of the replication failed to match those of Experiment I, the conclusion that the original results were spurious became quite plausible. Layton and Turnbull point out that it is equally plausible to regard the replication's results as erroneous. Does this mean that we have learned nothing from the data provided by 414 subjects? One point that might be made in answer to this question is that there were a lot of possible outcomes of the two experiments with which the obtained data were grossly inconsistent. Surely, one thing we might do with the data, then, is to discredit the hypotheses that would have predicted such outcomes.

As will be seen, it is necessary to be specific in stating alternative hypotheses in order to extract the most useful information from the data. The remainder of this note will, first, take up the problem of formulating alternative hypotheses for the Layton–Turnbull study and, then, attempt to use the Layton–Turnbull data to achieve some more conclusive interpretations than can be based just on significance tests.

Formulating an alternative hypothesis

The basic question is: Just how small a clairvoyance effect should we want to detect? One possible answer to this question is that we want to detect *any* effect, no matter how small. In principle, this answer is quite valid but, in practice, it is not helpful because there would be a limit on either our budget or our patience that would prevent us from looking for effects below some minimum magnitude. In attempting, then, to answer

the question for the purpose of doing some alternative analyses, I asked: What magnitude of effect were Layton and Turnbull looking for? While Layton and Turnbull did not give an explicit answer to this question in their article, there is an answer implicit in their selection of task and sample size. From this information we can determine, that is, what magnitude of predicted clairvoyance effect the Layton–Turnbull experiments would have detected at $\alpha = .05$ with some reasonably high probability, say, .90.

Consider first the magnitude of sought effect implicit in their Experiment I. Let us treat this from the perspective of either of their two major independent variable manipulations (Belief or Evaluation). Given their sample sizes and the variability of the number-of-hits-in-100-tries measure ($SD = (n\,p\,q)^{1/2} = (100(.2)(.8))^{1/2} = 4.00$), the minimum true difference between two subsamples of $n = 70$ each that would achieve significance ($\alpha = .05$, 1-tailed) 90% of the time is about 2.0 hits. Accordingly, the sought magnitude of predicted effect implicit in the Experiment I design was about 1.0 hits (i.e., equal and opposite effects of this magnitude of two treatments would be detected with probability .90). Since the range of the dependent measure was 100.0, Experiment I thus had reasonably good sensitivity for treatment effects averaging at least 1% of the measurable range of treatment effect.

If we apply similar calculations to the data of the two experiments combined, we find that the magnitude of sought average treatment effect implicit in the overall sample of both experiments was about 0.6 hits. In the following analyses, I shall assume that a slightly weaker treatment effect, averaging only 0.5 hits per subject (1/2% of the measurable range of effect), is (a) large enough to be of interest and (b) not too small to be beyond the sensitivity of a reasonable (but large) experiment.

Reanalyses of the Layton–Turnbull results[4]

Significance tests of combined experiments. Analysis of variance on the two experiments combined indicated no significant (or near significant) effects. The main effect of Evaluation that was near significant for Layton and Turnbull's Experiment I yielded an $F(1,394)$ of only 1.17 for the two experiments combined, clearly nonsignificant. The overall hit rate for the 414 subjects who participated in the two experiments (including 70 control subjects) was 20.094, or +.094 hits over the average of 20 to be expected by chance in the absence of ESP ability. The difference from chance of +.094 was not significantly greater than zero ($F < 1$).

[4] I thank Bruce Layton for providing the data needed to conduct the reanalyses reported here.

TABLE 2
NULL HYPOTHESES (H_0), ALTERNATIVE HYPOTHESES (H_1), AND 95% CONFIDENCE INTERVALS (CI) FOR MAIN EFFECTS AND GRAND MEAN OF LAYTON AND TURNBULL'S COMBINED EXPERIMENTS

Effect	H_0	H_1	Observed mean effect[a]	SD_m[b]	95% CI	
					Lower bound	Upper bound
Belief	0.0	+1.0	−.225	.436	−1.083	+.633
Evaluation	0.0	+1.0	+.472	.436	−.386	+1.330
Grand Mean[c]	0.0	+0.5	+.094	.199	−.296	+.484

[a] Observed mean effect = difference between means for two treatment levels or difference of grand mean from chance value of 20, based on data of both experiments.
[b] SD_m = Standard deviation of observed mean effect (394 df).
[c] Includes 70 control subjects.

Confidence interval analysis. The problem with summarizing the results in terms of significance tests (as has just been done) is that significance tests do not provide either the researcher or the reader with the information needed to decide whether the results are most compatible with a null hypothesis or a meaningful alternative. In the present case, we have formulated (see above) an alternative hypothesis (of treatment affects averaging 0.5 hits), but do not know from the significance test results, considered alone, whether the findings are more consistent with the truth of that alternative or of the null hypothesis (treatment effects averaging 0.0). Presentation of the results in terms of confidence intervals is somewhat more informative. Table 2 gives the major results of the combined Layton–Turnbull experiments expressed in the form of 95% confidence intervals. In the case of the main effect of Belief, the 95% confidence interval included both the null hypothesis and the reverse of the alternative (mean treatment difference of −1.0). For the main effect of Evaluation, the confidence interval included both the null hypothesis and the alternative. For the overall sample mean, the 95% confidence interval included the null hypothesis but not the alternative. A 95% confidence interval includes all hypotheses (points) that would not be rejected at the .05 level by a significance test. The confidence interval analysis does not, however, allow a conclusion in terms of preference among the hypotheses that would not be rejected by a significance test; nor, for that matter, does it allow a conclusion about the relative acceptability of hypotheses that would not be rejected compared to those that would, beyond this information itself (cf (Mosteller and Tukey, 1969, pp. 180–183)).

Bayesian hypothesis tests. In contrast with the significance test results and the somewhat more informative confidence-interval analyses, the

Bayesian hypothesis tests presented in Fig. 1 provide refreshingly direct confrontations between null and alternative hypotheses. Bayesian analyses differ most fundamentally from standard statistical tests by using a degree-of-belief (rather than a limit-of-relative-frequency) interpretation of probability. This permits the attachment of probabilities other than 0 or 1 to hypotheses, and these (variable) probabilities of hypotheses appear in the prior and posterior distributions of a Bayesian analysis. The steps of the Bayesian analysis are the following (cf (Phillips, 1973)).

(1) Formulate a *prior probability distribution* for each experimental effect. This is a probability density function indicating distribution of initial belief across the range of hypothetical values for the given effect. (In the present case, the three effects of interest were the differences between the contrasted pairs of Belief and Evaluation treatments and the deviation of the Grand Mean from its chance value of 20.)

(2) Compute a *likelihood function* for each effect from the data. This probability density function indicates the relative likelihood of the actual data for each point in the range of hypothetical values of a given effect.

(3) Obtain a *posterior probability distribution* for each effect by multiplying, point by point, the corresponding ordinates of the appropriate prior distribution and likelihood function, adjusting further by a constant to set the area under the posterior distribution equal to 1.

(4) Test hypotheses in terms of *posterior odds* computed from the posterior distribution. For example, the posterior odds favoring one exact hypothetical value for an effect over another would be obtained by dividing the posterior distribution ordinate (i.e., probability density) value of the first by that for the second. For inexact (interval) hypotheses, posterior odds would be obtained by comparing the areas under the posterior distribution defined by the boundaries of the respective intervals.

Because of the controversial nature of existing evidence on ESP, and also the lack of previous tests for the specific treatments employed by Layton and Turnbull, prior opinion is most reasonably regarded as being well scattered across a broad range of hypothetical values for the effects of Belief, Evaluation, and the overall Grand Mean. These prior beliefs are indicated by the dashed prior probability distributions in Fig. 1.

Since the means and standard deviations of the various effects are all unknown in advance of data collection, the likelihood functions for the several effects are described by Student's t distribution, with means, standard deviations, and degrees of freedom obtained from the data in the usual way (Phillips, 1973, p. 279). In Fig. 1, the lighter solid curves are the likelihood functions obtained from the data of Layton and Turnbull's Experiment I, considered by itself.

It may be seen, in Fig. 1, that the prior distributions are quite flat over

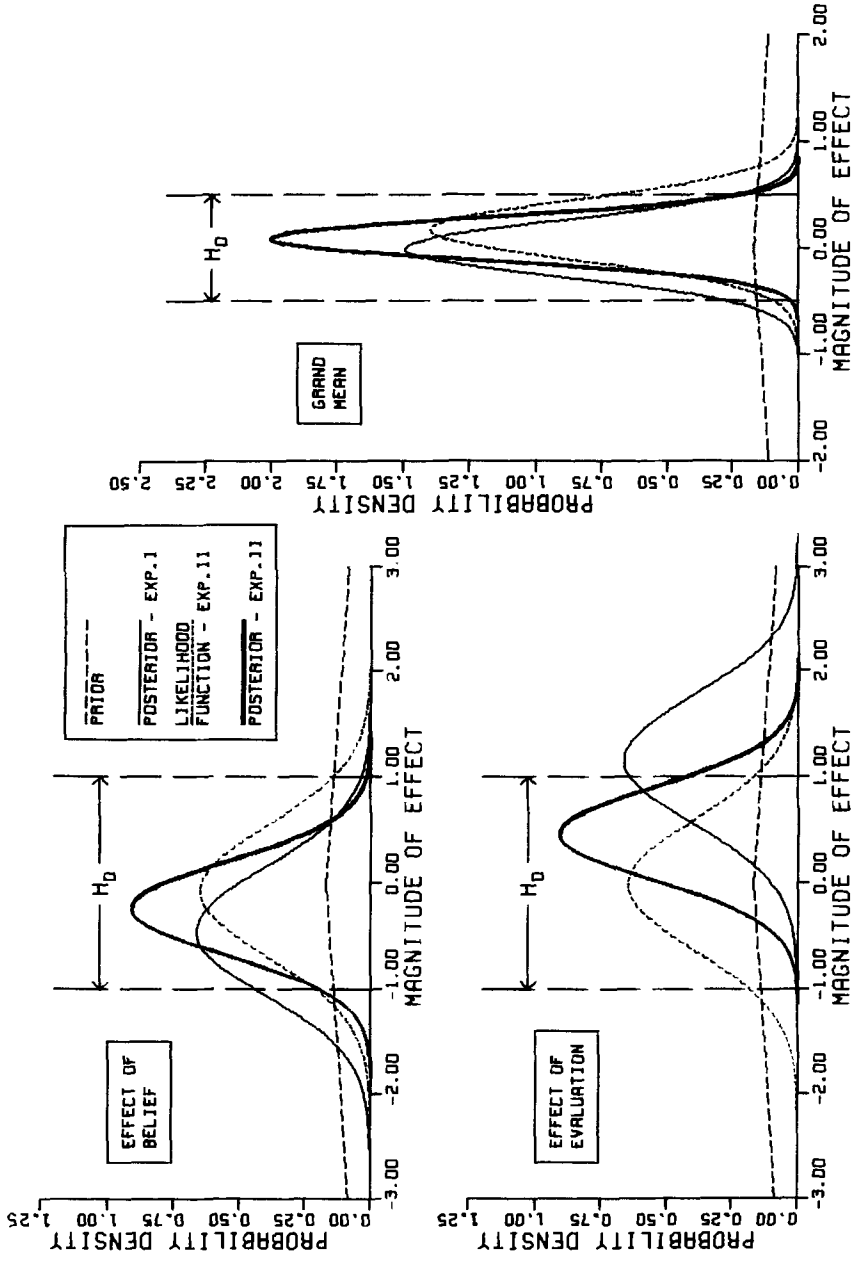

FIG. 1. Bayesian analysis of the Layton–Turnbull data.

the regions of values favored by the data (i.e., the regions covered by most of the likelihood functions). Consequently, the *principle of stable estimation* (Edwards, Lindman, and Savage, 1963, pp. 201–208) applies. This principle permits operation under the assumption that the prior distribution is effectively flat. In turn, this assumption simplifies subsequent computations, since the posterior distribution is identical to the likelihood function when the prior is flat. Therefore, the Experiment I likelihood functions are also Bayesian posteriors for Experiment I.

To continue the Bayesian analysis, the posteriors of Experiment I serve as priors for Experiment II. The likelihood functions for Experiment II are indicated as dotted curves in Fig. 1, and the heavy solid curves are the resulting final posterior distributions, which turn out to be *t* distributions with means, standard deviations, and degrees of freedom based on the combined data of the two experiments.

Table 3 summarizes hypothesis tests computed on the various Bayesian posteriors of Fig. 1. For these tests, each null hypothesis specifies that the true effect is smaller (absolutely) than the minimum alternative that the combined experiments had reasonable power to detect. For example, for the test of the effect of Belief, the (interval) null hypothesis extended from a difference between the Sheep and Goat conditions of −1 to a difference of +1. The hypothesis test shows that this "null range" is preferred over all alternative hypotheses by posterior odds of 23.29 : 1 on the basis of the two experiments' data. In only one case, the test of the effect of Evaluation for the Experiment I data, did the posterior odds even slightly (1.55 : 1) go against the null hypothesis. For the test of the Evaluation effect based on the final posterior, the null hypothesis was preferred over the alternative by odds of 7.76 : 1. Because 7.76 : 1 are not particularly large odds, it remains reasonable to entertain the hypothesis of some small effect of Evaluation for further research. It

TABLE 3
BAYESIAN HYPOTHESIS TESTS FOR THE LAYTON AND TURNBULL EXPERIMENTS

	Posterior odds[a] in favor of H_0 after	
	Experiment I	Experiment II
Effect		
Belief	4.32 : 1	23.29 : 1
Evaluation	0.64 : 1[b]	7.76 : 1
Grand Mean	15.00 : 1	43.07 : 1

[a] Posterior odds are computed from the appropriate posterior distribution in Figure 1 as the area within the H_0 range divided by the area outside that range.

[b] These odds favor the region outside the null range by 1.55 : 1.

should be quite clear, however, that this further research will be rather expensive to conduct (properly), since any experiment should have to be quite sensitive (powerful) in order to provide a substantial improvement in precision relative to the Layton–Turnbull study.

CONCLUSION

The judgment, "inconclusive," for Layton and Turnbull's (1974) findings might be based on either the inconsistent results of their tests of significance or on the supposition that their procedures may not have been adequate to test their hypotheses. The latter basis for the inconclusive verdict has nicely been ruled out by Layton and Turnbull's employment of manipulation checks which showed that the conditions necessary for testing their hypotheses were established with some clarity. The present reanalyses have shown further that Bayesian hypothesis tests allow greater conclusiveness in favor of null hypotheses than was possible with just the tests of significance reported by Layton and Turnbull.

Postscript

Social scientists have a marked preference for finding relationships between variables over finding no relationships. Thus, we value "significant" or "positive" results over "nonsignificant" or "null" or "negative" results. Some of the reasons for this state of affairs are understandable, but they are not entirely sensible. That is, it does appear that scientists' reputations are more readily established by looking for and finding new relationships that require new explanations than by looking for and finding nonrelationships that would discredit old (particularly their own) explanations. But it is distressing that we accumulate new relationships and explanations without getting rid of corresponding numbers of old ones, that the new explanations are often difficult to make consistent with one another, and that we often fail to face important empirical and theoretical problems because our significance tests divert us from them.

REFERENCES

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

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

BUCHWALD, A. M. Effects of "right" and "wrong" on subsequent behavior: A new interpretation. *Psychological Review,* 1969, **76,** 132–143.

EDWARDS, W., LINDMAN, H., & SAVAGE, L. J. Bayesian statistical inference for psychological research. *Psychological Review,* 1963, **70,** 193–242.

ESTES, W. K. Reinforcement in human learning. *In* J. T. Tapp (Ed.), *Reinforcement and behavior.* New York: Academic Press, 1969.

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

GREENWALD, A. G. Consequences of prejudice against the null hypothesis. Paper read at meetings of Society of Experimental Social Psychology, Iowa City, October, 1973.

HAYS, W. L. *Statistics for social scientists.* New York: Holt, 1973. 2nd ed.

LAYTON, B. D., & TURNBULL, B. Belief, evaluation, and performance on an ESP task. *Journal of Experimental Social Psychology,* 1975 **11,** 166–179.

MICHELSON, A. A., & MORLEY, E. W. On the relative motion of the Earth and the luminiferous ether. *American Journal of Science* (3rd Series), 1887, **34,** 333–345.

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

MORRISON, D. E., & HENKEL, R. E. (Eds.) *The significance test controversy.* Chicago: Aldine, 1970.

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

NUTTIN, J., & GREENWALD, A. G. *Reward and punishment in human learning.* New York: Academic Press, 1968.

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

PLATT, J. R. Strong inference. *Science,* 1964, **146,** 347–353.

THORNDIKE, E. L. *Animal intelligence.* New York: Macmillan, 1911.

WILSON, W. R., & MILLER, H. A note on the inconclusiveness of accepting the null hypothesis. *Psychological Review,* 1964, **71,** 238–242.