Psychological Bulletin
1989, Vol. 105, No. 2, 309-316

0033-2909/89/$00.75

Do Studies of Statistical Power Have an Effect on the Power of Studies?

Peter Sedlmeier and Gerd Gigerenzer
University of Konstanz, Federal Republic of Germany

The long-term impact of studies of statistical power is investigated using J. Cohen's (1962) pioneering work as an example. We argue that the impact is nil; the power of studies in the same journal that Cohen reviewed (now the *Journal of Abnormal Psychology*) has not increased over the past 24 years. In 1960 the median power (i.e., the probability that a significant result will be obtained if there is a true effect) was .46 for a medium size effect, whereas in 1984 it was only .37. The decline of power is a result of alpha-adjusted procedures. Low power seems to go unnoticed: only 2 out of 64 experiments mentioned power, and it was never estimated. Nonsignificance was generally interpreted as confirmation of the null hypothesis (if this was the research hypothesis), although the median power was as low as .25 in these cases. We discuss reasons for the ongoing neglect of power.

Since J. Cohen's (1962) classical study on the statistical power of the studies published in the 1960 volume of the *Journal of Abnormal and Social Psychology,* a number of power analyses have been performed. These studies exhorted researchers to pay attention to the power of their tests rather than to focus exclusively on the level of significance. Historically, the concept of power was developed within the statistical theory of Jerzy Neyman and Egon S. Pearson but was vigorously rejected by R. A. Fisher. Fisher compared statisticians concerned with Type II errors (beta) or power (1-beta) to "Russians," who were trained in technological efficiency as in a 5-year plan, rather than in scientific inference (Fisher, 1955, p. 70). Neyman, also with reference to the issue of power, called some of Fisher's testing methods "worse than useless" in a mathematically specifiable sense (Stegmüller, 1973, p. 2). The unresolved controversies about statistical inference within statistics proper had their impact on psychological research, although here the controversial issues have been for the most part neglected (Gigerenzer, 1987). In psychological textbooks, an apparently uncontroversial hybrid theory is taught, which contains concepts from both camps (e.g., null hypothesis testing, following Fisher, and Type II error and specification of the Type I error before the data has been obtained, following Neyman and Pearson). This hybrid statistics is usually taught as statistics per se, without mention of the originators of the respective ideas, and this mixture of concepts certainly would not have been approved by either of the originators (Gigerenzer & Murray, 1987, chap. 1).

It is important to understand the unresolved issue of power in psychological studies against the background of this unresolved debate in statistics rather than as an isolated issue. We shall come back to this connection in the Discussion section.

This research was supported by an Akademie Stipendium from the Volkswagen Foundation and by Grant 17900585 from the University of Konstanz to the second author.

We are grateful to Jacob Cohen and to an anonymous reviewer for many helpful comments and to Phil Magin for his assistance in data analysis.

Correspondence concerning this article should be addressed to Gerd Gigerenzer, Fachgruppe Psychologie, Universität Konstanz, Postfach 5560, 7750 Konstanz, Federal Republic of Germany.

Power Studies

In the Neyman-Pearson theory, power (1-beta) is defined as the long-run frequency of acceptance of H_1 if H_1 is true. Recall that in their theory two point hypotheses, often called H_0 and H_1, are formulated, and this allows both alpha and beta to be determined. Beta cannot be calculated in Fisher's theory of null hypothesis testing, where only one point hypothesis, the null, is specified and tested (Fisher, 1935/1966). There are three major factors that influence the magnitude of the power: effect size in the population, level of significance, and number of observations. The effect size expresses the discrepancy between H_0 and H_1; for example, for a t test between sample means, it is the standardized difference between the two population means posited by H_0 and H_1 (J. Cohen, 1977). Everything else being constant, the greater the effect size, the greater the power. The level of significance, alpha, is the long-run frequency of rejecting H_0 if H_0 is true, which must be posited before the data are obtained, according to Neyman-Pearson theory. Note that in Fisher's theory of null hypothesis testing, the effect size is not a concept, and the level of significance can be determined after the data have been obtained (compare his later writings, e.g., Fisher, 1956). Everything else being constant, the smaller the level of significance, the smaller the power. Finally, when the number n of observations increases, the standard deviations of the sampling distributions for H_0 and H_1 become smaller. Thus the distributions will overlap less, and the power will increase. Besides these three major factors, the assumptions of the statistical model are important insofar as they affect the power if they do not hold. Examples are violation of independence (Edgell & Noon, 1984), false assumptions concerning the equality of variances (Petrinovich & Hardyck, 1969), and false assumptions about measurement scales and shapes of distributions (Trachtman, Giambalvo, & Dippner, 1978).

For the purpose of power analysis, it is assumed that the assumptions of the statistical model hold. In that case, any of the four variables—effect size, n, alpha, and power—can be determined as a function of the other three. The usual procedure for calculating the power of published studies is to determine the n of a test, to assume that alpha has been set as .05 in advance (which is not clearly stated in many studies), and to calculate

Table 1
Results of Power Studies (Arithmetical Means)

Journal	Study	Effect size		
		Small	Medium	Large
Journal of Abnormal and Social Psychology	J. Cohen (1962)	.18	.48	.83
American Journal of Educational Research	Brewer (1972)	.14	.58	.78
Journal of Research in Teaching	Brewer (1972)	.22	.71	.87
The Research Quarterly	Brewer (1972)	.14	.52	.80
Journal of Communication	Katzer & Sodt (1973)	.23	.56	.79
Counselor Education and Supervision	Haase (1974)	.10	.37	.74
American Sociological Review	Spreitzer (1974; cited in Chase & Tucker, 1976)	.55	.84	.94
American Forensic Association Journal, Central States Speech Journal, Journal of Communication, The Quarterly Journal of Speech, Southern Speech Communication Journal, Speech Monographs, The Speech Teacher, Today's Speech, Western Speech	Chase & Tucker (1975)	.18	.52	.79
American Speech and Hearing Research, Journal of Communication Disorders	Kroll & Chase (1975)	.16	.44	.73
Journalism Quarterly, The Journal of Broadcasting	Chase & Baran (1976)	.34	.76	.91
Journal of Applied Psychology	Chase & Chase (1976)	.25	.67	.86
Journal of Marketing Research	Sawyer & Ball (1981)	.41	.89	.98

the power depending on an assumed effect size. Generally, three levels of effect size, as suggested by J. Cohen (1962, 1969), are used. For the purpose of planning research, that is, for prospective rather than retrospective power analysis, the procedure is different. There are two major possibilities. The first is to calculate n, given a conventional alpha, a size of effect estimated from previous research, and a desired power. J. Cohen (1965), for instance, recommended .80 as a convention for a desirable power. The second possibility is to calculate alpha, given n, a size of effect, and a desired power. The second possibility has almost never been considered in actual research. The reason for this neglect can be seen in the widespread interpretation of alpha as a conventional yardstick for inductive inference. The calculations can be facilitated using the tables provided by J. Cohen (1969, 1977), in which (a) the power of commonly used (parametric) tests is listed as a function of effect size, n, and three levels of alpha (.01, .05, .10; one- and two-tailed tests) and (b) n is listed as a function of the other three variables. J. Cohen (1970) also offered rule-of-thumb procedures that allow a rough estimation of power for seven standard test statistics.

The first systematic power analysis was conducted by J. Cohen (1962), analyzing all studies published in the 1960 volume of the *Journal of Abnormal and Social Psychology.* He distinguished between major and minor research hypotheses and calculated the mean and median power of all significance tests performed in an article, for each kind of hypothesis. The calculations were based on alpha = .05 and on three effect sizes—small, medium, and large—corresponding to the dimensionless Pearson correlations .20, .40, and .60, respectively. His results, together with those of subsequent studies that appeared after a time lag of a decade, are depicted in Table 1. Cohen found a mean power of only .18 for detecting small effects, of .48 for medium effects, and of .83 for large effects. If one follows his judgment that medium effects are what one can expect in this area of research, then it follows that the experiments were designed so that the researcher had less than a 50% chance of obtaining a significant result if there was a true effect. Note that Cohen did not calculate the sample effect sizes in the 1960 volume. His definitions of small, medium, and large effect sizes were based on judgment but seem to be indirectly supported from calculations of sample effect sizes in related areas of research (Haase, Waechter, & Solomon, 1982), although these are not without problems (Murray & Dosser, 1987). Cooper and Findley (1982) concluded from their effect size study in social psychological research that it seems reasonable to assume a medium effect size (Cohen's definition) in power analysis. Most power studies have used Cohen's definitions; and the present study gives direct evidence for the validity of his definition with respect to his own study (J. Cohen, 1962).

The other studies shown in Table 1 are comparable to Cohen's study, because those authors followed Cohen's procedure, with the exception of Brewer (1972), who used single tests rather than articles as units.[1] Four other studies could not be

[1] The power values of Cohen's original study are not strictly comparable with those of the later studies summarized in Table 1. The reason is that Cohen has lowered over time the criteria of what constitutes a small, medium, or large effect for some statistics, and the later studies all used his criteria published in 1969 (or 1977). For instance, in 1962, a Pearson correlation of .2, .4, or .6 was defined as a small, medium, or large effect, respectively, whereas in 1969 (and 1977) the corresponding values were .1, .3, and .5. This systematic lowering of the effect size conventions has the effect of slightly lowering the calculated power, too. We have taken account of this problem in comparing the states of affairs of now and then (1984 and 1960) by using Cohen's original criteria whenever possible.

included in Table 1 because they were not comparable; these studies did not use dimensionless measures of effect size (Freiman, Chalmers, Smith, & Kuebler, 1978), confused population effect size with obtained sample effect size (Ottenbacher, 1982), used only one "median total sample size" (Arvey, Cole, Hazucha, & Hartanto, 1985), or examined problems in a specific area of research (King, 1985). Most of these studies were conducted in the 1970s and show about the same low average power as that in Cohen's study, although there are a few journals with considerably more powerful tests, such as the *American Sociological Review* and the *Journal of Marketing Research.*

In general, these studies reveal an apparently paradoxical state in research. Given the high premium placed by both researchers and journal editors on significant results (see Atkinson, Furlang, & Wampold, 1982; Bredenkamp, 1972; L. H. Cohen, 1979; Coursol & Wagner, 1986; Greenwald, 1975; Melton, 1962), it seems strange that research was planned and conducted to give only a low chance of a significant result if there was a true effect. Researchers paradoxically seem to prefer probable waste of time, money, and energy to the explicit calculation of power.

24 Years Later: A Case Study

The question of interest is whether studies of power have an effect on the power of studies. Because methodological innovations, be they wrong or right, often have an incubation time on the order of decades rather than years, it is fair to use as a case study the influence of Cohen's analysis, which is both the oldest and the most prominent analysis. What impact did Cohen's analysis have on the power of studies published later in the same journal? That journal has meanwhile been divided into the *Journal of Abnormal Psychology* and the *Journal of Personality and Social Psychology.* We decided to analyze all studies published in the 1984 volume of the *Journal of Abnormal Psychology.* Twenty-four years should be sufficient time for a methodological innovation to be established. Of course, our comparison between the 1960 and the 1984 volumes cannot prove a causal connection; all that we can do is to establish whether there is a change or not.

Method

J. Cohen (1962) calculated the power using articles as units, but 24 years later, articles often contain more than one experiment. We have treated different experiments within the same article as separate units if they used different samples of subjects. Power calculations were made using two different units: articles, as in J. Cohen (1962), and experiments. For each unit, all tests were classified into tests either of major or of minor hypotheses, following J. Cohen (1962). Then the power of each test was calculated for small, medium, and large effect sizes; for most tests, J. Cohen's (1977) tables could be used. For all calculations, alpha was assumed to be .05 (two-tailed for the unadjusted procedures; for the multiple comparison procedures the corresponding error rates [see, e.g., O'Neill & Wetherill, 1971] were used). Nonparametric tests that occasionally occurred were treated like their corresponding parametric tests (e.g., Mann-Whitney *U* tests, Kruskal-Wallis tests, and Spearman correlations were treated like *t* tests, *F* tests, and Pearson correlations, respectively); this usually results in a slight overestimation of power (J. Cohen, 1965).

The major difference between the tests published in 1960 and 1984 was the frequent use of alpha-adjusted procedures. These were practically nonexistent in 1960. In order to control for the effect of alpha adjustment, we calculated the power of each unit of analysis both including and excluding these procedures. In the latter case, alpha-adjusted tests like *t* tests, *F* tests, and chi-square were treated as if they were not adjusted, and multiple comparison procedures were excluded. For the Newman-Keuls, Duncan, and Tukey procedures the tables of Harter (1969) were used, and for the Scheffé test the tables of Hager and Möller (1985, 1986) were used.[2]

Results

The 1984 volume of the *Journal of Abnormal Psychology* contained 56 articles. One of them discussed previous research and contained no statistical tests; another used descriptive statistics concerning the major hypotheses and only one marginal test. These two articles were not evaluated. The remaining 54 articles employed statistical testing of hypotheses in 64 experiments. Alpha-adjusted procedures were used in 28 articles (31 experiments). In seven articles (seven experiments) at least some research hypotheses were stated as statistical null hypotheses. In these cases, any conclusion that nonsignificance supports the research hypotheses is unwarranted without power considerations.

Remarks on power were found in only two cases, and nobody estimated the power of his or her tests. In four additional articles, alpha was mentioned, either by saying that it was set at a certain level (.05) before the experiment or by referring to the danger of alpha inflation. No author discussed why a certain alpha or *n* was chosen or what effect size was looked for. This first result shows that concern about power is almost nonexistent, at least in print.

Our calculations of power, based on either experiments or articles as units of analysis, resulted in practically identical values, the median and maximum absolute deviations being .01 and .04, respectively. Similarly, the tests of major and minor hypotheses were indistinguishable with respect to average power. For these reasons, we do not distinguish between these here, and we report the power of major tests using experiments as a unit. Table 2 reports the power of major tests using experiments as units, excluding alpha-adjusted procedures. This simulates the situation 24 years earlier, when those procedures were not common. A small number of experiments had very high power (see columns labeled *Small effects* and *Medium effects*); because of these outliers, we consider the medians rather than the means as representing the average experiment's power. The median power for small, medium, and large effects was .14, .44, and .90, respectively. Twenty-four years earlier, the median power was .17, .46, and .89, respectively. As a general result, therefore, the power has not increased after 24 years. Although there are now a small percentage of experiments in which the chance of finding a significant result if there is an effect is high, even for small and medium effects, the respective median power for medium effects is slightly lower than that 24 years earlier.

For the purpose of comparison with the 1960 volume, where

[2] The number of tests, the test statistic used, the *n* used and whether a test concerned a major or minor hypothesis were evaluated by an independent rater. This rater analyzed 10 randomly chosen articles. The average amount of agreement between the authors and the independent rater was 98%.

Table 2
Power of 64 Experiments Published in the 1984 Volume of the Journal of Abnormal Psychology, *Excluding Alpha-Adjusted Procedures (Alpha = .05, Two-Tailed Tests)*

	Small effects		Medium effects		Large effects	
Power	Frequency	Cumulative percentage	Frequency	Cumulative percentage	Frequency	Cumulative percentage
.99	1	100	5	100	21	100
.95–.98	—	98	—	92	4	67
.90–.94	—	98	2	92	8	61
.80–.89	—	98	6	89	12	48
.70–.79	3	98	4	80	7	30
.60–.69	1	94	6	73	6	19
.50–.59	—	94	4	64	2	9
.40–.49	1	92	7	58	2	6
.30–.39	5	91	10	47	1	3
.20–.29	8	83	15	31	1	2
.10–.19	34	70	4	8	—	—
.05–.09	11	17	1	2	—	—
N	64	—	64	—	64	—
M	.21		.50		.84	
Mdn	.14		.44		.90	
SD	.19		.27		.18	
Q_1	.10		.28		.76	
Q_3	.22		.70		.99	

alpha adjustment was practically nonexistent, we have ignored these techniques in calculating the power in Table 2. However, as was mentioned previously, about 50% of all articles used at least one of the various alpha-adjusted procedures to test major research hypotheses. Because the power increases with alpha, the real power of the 64 experiments is smaller than that in Table 2. The effect of the alpha-adjusted procedures on the power of the tests is shown in Table 3, both for the entire set of experiments and for those using adjustment procedures. Results show that the use of alpha adjustment in 1984 decreased the real median power considerably. Because of the emphasis on alpha adjustment, the median power of detecting a medium-sized effect if there is one is only .37. Thus, we must conclude that 24 years later, the power of the tests has decreased instead of increased. The two bottom rows in Table 3 show an interesting state of affairs. Experiments using alpha-adjusted procedures have on the average smaller power than those that do not, even when these procedures are excluded from our calculations. This smaller power is then once more diminished by the adjustment, resulting in a median power of .28 for medium effects. Thus the researchers using alpha adjustment designed their experiments as if they believed that alpha adjustment compensated for the factors that increase power (such as large *n*), whereas in fact it decreases power.

Table 3
Impact of Alpha Adjustment on the Power of the 64 Experiments From the 1984 Volume of the Journal of Abnormal Psychology *(Values Are Medians)*

	Effect size		
Condition	Small	Medium	Large
All experiments (N = 64)			
Including alpha adjustment	.12	.37	.86
Excluding alpha adjustment	.14	.44	.90
Experiments using alpha adjustment (N = 31)			
Including alpha adjustment	.10	.28	.72
Excluding alpha adjustment	.14	.37	.85

The ratio of beta to alpha implies a conception of the relative seriousness of the two possible errors. This ratio varies between 14:1 and 11:1 for the conditions in Table 3, assuming a medium effect size. This means that researchers act as if they believe that mistakenly rejecting the null hypothesis is 11 to 14 times more serious than mistakenly accepting it.

We now shall address a specific issue: null hypotheses as research hypotheses. As was stated previously, in seven experiments (11%) at least some null hypotheses (no difference between treatments) were operationalizations of the research hypotheses. None of these tests became significant, and this result was unanimously interpreted by the authors as a confirmation of their research hypothesis. The power of these tests should have been particularly high in order to justify such conclusions (e.g., .95, which would correspond roughly to the case in which alpha = .05 and H_1 is the research hypothesis). Otherwise, in the case of unknown and probably low power, a nonsignificant result signifies that no conclusion should be drawn, that is, that one should not affirm the null hypothesis with an uncontrolled error rate (beta error) and that the experiment probably was a waste of time and money. The actual median power in these nonsignificant tests was .25, with a range between .13 and .67 for a medium-sized effect. This means that the experimental conditions, such as number of observations, were set up in such a way that given a true medium effect, the research (null) hy-

pothesis would nevertheless be "confirmed" in 75% of the cases. There can be no doubt that such tests of research hypotheses are empirically worthless and the positive conclusions unwarranted, and the question arises, How can such tests and conclusions be accepted and published in a major journal?

Discussion

We found almost no concern with power expressed in the 1984 volume, and no increase in the power of tests from 1960 to 1984, but rather a considerable decrease of power due to the frequent use of alpha-adjusted procedures. J. Cohen's (1962) seminal power analysis of the *Journal of Abnormal and Social Psychology* seems to have had no noticeable effect on actual practice as demonstrated in the *Journal of Abnormal Psychology* 24 years later.[3] Must we conclude that researchers stubbornly neglect a major methodological issue over decades? Or should we assume that they are intuitively right and that we really do not need more power than .37?

Comparability

One way to defend research practice against our conclusion of "no change in low power" would be to assume that Cohen's criterion for a medium effect does not hold for both the 1960 and the 1984 volumes and that effects actually studied in 1984 were considerably larger, which implies larger power. For instance, if the 1960 studies were primarily of problems that yielded small effect sizes, but the 1984 studies were of problems that yielded medium-sized effects, this would suggest a change in power. Because none of the articles specified the sought-after effect size before the experiment (as Neyman-Pearson logic implies one should), we can check this conjecture only with respect to the actual sample effect sizes (determined after the experiment). As we mentioned earlier, sample effect sizes were not determined by Cohen for the 1960 volume; he instead used rule-of-thumb definitions for effect sizes. Thus we calculated effect sizes for both the 1960 and the 1984 volumes, in order to determine whether there was an increase in sample effect size.

To test the comparability of actual sample effect sizes between 1960 and 1984, we drew random samples of 20 (experimental) articles from each of the volumes. Sample effect sizes were again calculated separately for major and minor tests and for articles and experiments as units (the 1960 sample contained 20 experiments, the 1984 sample, 25 experiments). The median effect sizes were .31 in the 1960 sample and .27 in the 1984 sample (all effect sizes are expressed as Friedman's r_m).[4] These median effect sizes were identical under all the conditions mentioned here. The ranges were .12 to .69 and .08 to .64, respectively. This shows (a) that Cohen's definition, assuming a medium effect size (Pearson *r*) of .40 (J. Cohen, 1962) and .30 (J. Cohen, 1969), was quite close to the actual median sample effect size found in our analysis, and most important, (b) that sample effect sizes did not increase from 1960 to 1984. In fact, our results show the opposite tendency: Median sample effect sizes decreased slightly. These results speak for the comparability of actual sample effect sizes and contradict the assumption of an increase in power due to an increase in actual sample effect size.

Furthermore, our analysis provides a check of Cohen's judgment of a medium effect size in the 1960 volume. Recall that in his original study, he defined $r = .40$ as a medium effect size, and so did we in the present study. Using his own criteria for comparing various measures of effect size (J. Cohen, 1977, p. 82), we calculated that a point biserial $r_p = .32$ corresponds to $r = .40$. Because Friedman's r_m is roughly equivalent to r_p (for this, see Cohen, 1965, pp. 104–105), we may conclude that Cohen's judgment of a medium effect size of $r_p = .32$ corresponds closely to the actual sample median of .31 found in our analysis of the 1960 volume. Although a strict numerical comparison poses numerous difficulties, we now have evidence that his judgment was very close to the true median sample effect size.

Intuitions on Compensation

As was mentioned previously, power is a function of effect size, *n*, and alpha. Assuming that alpha is constant, concern with power should lead experimenters to compensate for a small expected population effect size by obtaining a large *n*, and vice versa. Although we have established that with the exception of two articles in the 1984 volume, nobody talks about power in print, researchers might follow this compensation principle intuitively. In particular, if there was a change in intuitions, then the correlation between effect size and *n* should be negative and larger than that in 1960.

Because single experiments often involved numerous tests with varying *n*s, we checked the intuitions with respect to both *n* and the number of subjects *N* (as given in the subjects sections). The latter is probably the more salient figure for the experimenter. For the actual sample effect sizes available, we calculated a Pearson correlation (major hypotheses only) between sample effect size and *N* of −.35 in 1960 and .01 in 1984. In-

[3] Although the power did not increase over the years, at least in the present case study, references to power as measured by the citation frequency of J. Cohen's (1969, 1977) book multiply. The Science Citation Index (including additional citations in the Social Science Citation Index) reports 4, 13, 83, 193, and 214 citations for the years 1971, 1975, 1980, 1985, and 1987, respectively. This indicates growing attention to the issue and, possibly, differences between journals.

[4] Calculations of sample effect sizes were based on degrees of freedom, values of test statistics, *p* values, and *n* reported, following Friedman (1968). In cases of missing information, we proceeded in the following way. If only means and variances were specified in *t* tests, we calculated the point-biserial correlation coefficient, following J. Cohen (1977). If a test result was only described as not significant and the *n* for the test could be determined, we calculated the sample effect size for $p = .05$ and divided the value by a factor of 2. We consider this to be a reasonable approximation, but it results in a tendency to obtain larger sample effect sizes for smaller *n*s. Therefore, we also made a second calculation in which sample effect sizes for nonsignificant results were assumed to be zero. This second calculation resulted in median sample effect sizes that were .02 and .05 smaller than those reported in the text, for the 1960 and 1984 volumes, respectively. It should be noted that values in the text that follows are calculated using the first method. If authors reported *p* values only, but not the value of the test statistic, the latter was inferred using the tables of Dunn and Clark (1974) and Hager and Möller (1985). Tests for multivariate procedures (e.g., multivariate analysis of variance), alpha-adjusted procedures, and coefficients of reliability (e.g., interrater reliability) were not included in the analyses. The procedure for estimating sample effect sizes was the same for the 1960 and the 1984 volumes.

spection of the data revealed that the zero correlation was due to one experiment that investigated an unusually large number of subjects (725) and that found a rather large sample effect size (.34). Excluding this experiment, the correlation was −.37 in the 1984 sample, similar to that in 1960. The corresponding values for all tests (including minor hypotheses) were −.34 and −.33 in the 1960 and 1984 samples, respectively. However, the *N*s reported in the subjects sections may not be the best guess for the actual *n*s, because there are procedures such as the analysis of variance (ANOVA) with repeated measures that use a much larger *n* for the single test. The correlations between average *n*s and sample effect sizes per experiment were −.34 in 1960 and −.47 in 1984.[5] The corresponding values for all tests (including minor hypotheses) were .07 in 1960 and −.36 in 1984. The latter, less negative values could be seen to indicate some sensitivity for the relationship between power and the relative importance of a test, especially in the 1960 sample. However, the small difference in the correlations for major tests does not seem to warrant any conclusions of a change in intuitive power considerations.

Tentative Explanation for the Zero Impact of Power Studies

The question of how psychologists came to neglect power in the first place is a historical one. The question of why they continue to neglect power seems to be an institutional one. For historical reasons, psychologists became familiar first with Fisher's theory of null hypothesis testing, from about 1935 on, and only later, during World War II, with Neyman and Pearson's statistical theory (Acree, 1979; Gigerenzer, 1987; Lovie, 1979; Rucci & Tweney, 1980). Textbooks of psychology and education first taught the Fisherian message, but after World War II, textbook writers realized the impact of Neyman-Pearsonian theory within statistics proper and began to supplement the null hypotheses testing theory with concepts of the latter theory, such as power. The incorporation of Type II error and power was done reluctantly. This is not surprising, because Fisher (1955) himself rejected Neyman and Pearson's emphasis on power and utility analyses, and power could not be calculated within null hypothesis testing theory. For instance, in the third edition of his influential *Fundamental Statistics in Psychology and Education,* Guilford (1956) still declared the concept of power as too complicated to discuss (p. 217). Finally, the concepts of Type II error and power were added by the majority of textbook writers to the framework of null hypothesis testing but could not be numerically determined, because most textbooks did not teach students to set up a second point hypothesis, as in Neyman-Pearson theory, which would have been necessary for the calculation. The resulting hybrid theory was usually anonymously presented as inferential statistics per se, and the controversial issues and the existence of alternative theories about statistical inference were neglected (Gigerenzer & Murray, 1987). This was a strange and rare event in psychology, where the presentation of alternatives and controversies had always been the rule rather than the exception and where no one would dare to mix, say, the theories of Freud and Adler and to present the result as psychoanalysis per se. With hindsight, the great error during that time (when inferential statistics became institutionalized) was the attempt to present ideas from two fighting camps as a single monolithic body of statistical knowledge rather than to present one theory after the other and to make their different concepts and points of views explicit. This historical accident suggested a single, mechanical solution to the problem of scientific inference, and there seemed to be no need for methodological alternatives, because no controversial issues seemed to exist.

The institutionalization of this hybrid theory, patched together from the opposing theories of Fisher and Neyman and Pearson and sometimes supplemented by a Bayesian interpretation of what significance means, was documented recently (Gigerenzer et al., 1989, chaps. 3 and 6). This attempt to fuse opposing theories into a single truth generated, as a necessary consequence, confusion and illusions about the meaning of the basic concepts. For instance, Fisher and Neyman and Pearson never agreed whether the level of significance should be determined before or after the experiment, whether it applied to the single experiment or to the long-run frequency of errors, whether significance generated new knowledge about hypotheses or not, and so on. Therefore, it is not surprising that the hybrid theory became a steady source of contradiction and confusion (see, e.g., Bakan, 1966; Oakes, 1986). The ongoing neglect of power seems to be a direct consequence of this state of affairs. With respect to this, important confusions are the ideas that the level of significance determines (a) the probability that a significant result will be found in a replication and (b) the probability that H_1 (or H_0) is true. These and related confusions can be found in well-known American and German textbooks (e.g., Bortz, 1985, p. 149; Brown, 1973, pp. 522–523; Nunnally, 1975, p. 195) and in editorials of major journals (e.g., Melton 1962, pp. 553–554). Furthermore, as research on statistical intuitions of researchers in psychology indicates, these confusions seem to be shared by many of our colleagues. Tversky and Kahneman (1971) inferred from a questionnaire distributed at meetings of the Mathematical Psychology Group and of the American Psychological Association that most respondents have wrong intuitions about the relationship between number of observations and power, that is, that they systematically overestimate the power of experiments and believe in the "law of small numbers." Oakes (1986) tested 70 academic psychologists and reported that 96% held the erroneous opinion that the level of significance specified the probability that either H_0 or H_1 was true. Given such misconceptions, the calculation of power may appear obsolete, because intuitively, the level of significance already seems to determine all we want to know.

Moreover, the average researcher is not entirely to blame for conducting studies with a power of only .37. It is a historical accident that Fisher's theory of null hypothesis testing, which opposed power calculations in the Neyman-Pearson framework, became the starting point of the inference revolution in

[5] For correlations and chi-square (in the sample almost exclusively with $df = 1$), we used the given *n* for the test. The *n* for *t* tests was $df + 1$ and the *n* for *F* tests was $df_{\text{denominator}} + df_{\text{numerator}} + 1$ for main effects and $df_{\text{denominator}}$ + number of relevant cells for interactions. Occasionally occurring nonparametric tests were treated like their parametric counterparts. The correlations were calculated, excluding the outlier in the 1984 sample (see text) and one 1960 experiment in which a self-constructed ANOVA technique was used. Inclusion of these two extreme outliers would lead again to zero correlations (−.09) and positive correlations (.35), respectively.

psychology. But the researcher's alertness to alternatives has been dulled by the presentation of a hybrid theory as a monolithic, apparently uncontroversial theory of statistical inference. This may be responsible for current conservative attitudes, which shy away from practical innovation. Innovations that are accepted, such as alpha adjustment, are those that adjust the theory that was historically the first: null hypothesis testing and its emphasis on the level of significance. The cumulations of beta errors, in contrast, have been payed almost no attention (see, however, Westermann & Hager, 1986).

This historical development explains why psychologists were not familiar with calculating the power of a test in the first place; and the merging of null hypothesis testing with Neyman-Pearson theory and the presentation of the resulting hybrid theory as a monolithic statistical theory explains to some degree the ongoing neglect. Even studies of power seem to have no effect on the power of studies, at least in the case investigated in this article. What can be done about this situation? We believe there is only one force that can effect a change, and that is the same force that helped to institutionalize null hypothesis testing as the sine qua non for publication, namely, the editors of the major journals. This situation will not change until the first editor of a major journal writes into his or her editorial policy statement that authors should estimate the power of their tests if they perform significance testing, and in particular if H_0 is the research hypothesis.

Do We Really Need Power?

Of the three major approaches to inductive inference and hypothesis testing—Bayes, Fisher, and Neyman-Pearson—power is a concept of central importance only in the latter. Thus, a fundamental question emerges: Do we need power at all? The answer to this provocative question depends on whether researchers believe that they have to make a decision after an experiment or not. Neyman-Pearson theory aims at a decision between hypotheses, and Neyman and Pearson's examples focus on applications such as quality control, in which the statistical procedure serves to direct a final practical decision, such as stopping the production if the quality has decreased. In fact, Neyman and Pearson's joint papers contain no application in which a scientific hypothesis was the sole or primary object (Birnbaum, 1977, p. 30). Although Fisher ridiculed their reject–accept notion in the context of scientific inference, his earlier work, in particular *The Design of Experiments* (Fisher, 1935/1966), which was so influential on psychology, could be understood by many as implying a reject notion on a conventional 5% level of significance. In fact, in the hybrid theory that was institutionalized in psychology, his null hypothesis testing became linked with the reject–accept notion. The essential question is whether psychological research needs yes–no types of decisions, as in quality control and related areas. We believe that there is no unique answer and that an answer depends on the specific content and will more often be positive in applied research than elsewhere. However, given the general belief among psychologists in the decision type of statistical inference, knowledge about power remains indispensable. If, at a future point, the influence of both Fisherian and Neyman and Pearsonian theories on psychological methodology can be transcended,[6] then the perceived importance of decisions based on significance might decrease, and other methodological principles could gain or regain importance. An example would be the fundamental principle of controlling the error before the experiment rather than after, that is, of manipulating conditions, tasks, and measurement procedures before the experiment until one has a very small error in the dependent variable. Today, fast data collection methods are often preferred, and the error is dealt with by inserting it into the t or F value after the experiment has been performed. One tends to wait to see whether it will turn out to be significant or not. Gosset, who developed the t test in 1908, anticipated this overconcern with significance at the expense of other methodological concerns: "Obviously the important thing . . . is to have a low real error, not to have a 'significant' result at a particular station. The latter seems to me to be nearly valueless in itself" (quoted in Pearson, 1939, p. 247). As long as decisions based on conventional levels of significance are given top priority, however, theoretical conclusions based on significance or nonsignificance remain unsatisfactory without knowledge about power.

[6] Here, we refer to those parts of Fisher's and of Neyman and Pearson's theories that have been institutionalized as an instrument of scientific inference in psychology; there is of course much more contained in their theories.

References

Acree, M. C. (1979). Theories of statistical inference in psychological research: A historico-critical study. *Dissertation Abstracts International, 39,* 5073B. (University Microfilms No. 7907000)

Arvey, R. D., Cole, D. S., Hazucha, J. F., & Hartanto, F. M. (1985). Statistical power of training evaluation designs. *Personnel Psychology, 38,* 493–507.

Atkinson, D. R., Furlang, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology, 29,* 189–194.

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

Birnbaum, A. (1977). The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindsley-Savage argument for Bayesian theory. *Synthese, 36,* 19–49.

Bortz, J. (1985). *Lehrbuch der Statistik: Für Sozialwissenschaftler* (2nd ed.). Berlin: Springer-Verlag.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt/Main, Federal Republic of Germany: Akademische Verlagsgesellschaft.

Brewer, J. K. (1972). On the power of statistical tests in the American Educational Research Journal. *American Educational Research Journal, 9,* 391–401.

Brown, F. L. (1973). Introduction to statistical methods in psychology. Appendix in G. A. Miller & R. Buckhout, *Psychology: The science of mental life.* New York: Harper & Row.

Chase, L. J., & Baran, S. J. (1976). An assessment of quantitative research in mass communication. *Journalism Quarterly, 53,* 308–311.

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61,* 234–237.

Chase, L. J., & Tucker, R. K. (1975). A power-analytic examination of contemporary communication research. *Speech Monographs, 42,* 29–41.

Chase, L. J., & Tucker, R. K. (1976). Statistical power: Derivation, development, and data-analytic implications. *The Psychological Record, 26,* 473–486.

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

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1970). Approximate power and sample size determination for common one-sample and two-sample hypothesis tests. *Educational and Psychological Measurement, 30,* 811–831.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, L. H. (1979). Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 47,* 421–423.

Cooper, H., & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 8,* 168–173.

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta analysis bias. *Professional Psychology: Research and Practice, 17,* 136–137.

Dunn, O. J., & Clark, V. A. (1974). *Applied statistics: Analysis of variance and regression.* New York: Wiley.

Edgell, S. E., & Noon, S. M. (1984). Effect of violation of normality on the *t* test of the correlation coefficient. *Psychological Bulletin, 95,* 576–583.

Fisher, R. A. (1966). *The design of experiments* (8th ed.). Edinburgh, Scotland: Oliver & Boyd. (Original work published 1935)

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Ser. B, 17,* 69–78.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Edinburgh, Scotland: Oliver & Boyd.

Freiman, J. A., Chalmers, T. C., Smith, H., Jr., & Kuebler, R. R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative trials." *The New England Journal of Medicine, 299,* 690–694.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin, 70,* 245–251.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences.* Cambridge, MA: MIT Press.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L. J., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life.* Cambridge, England: Cambridge University Press.

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

Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3rd ed.). New York: McGraw-Hill.

Haase, R. F. (1974). Power analysis of research in counselor education. *Counselor Education and Supervision, 14,* 124–132.

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in Counseling Psychology. *Journal of Counseling Psychology, 29,* 58–65.

Hager, W., & Möller, H. (1985). *Zentrale F-Verteilungen für zufällige Effekte und Signifikanzbeurteilungen.* Unpublished manuscript, Institute for Psychology, University of Göttingen, Göttingen, Federal Republic of Germany.

Hager, W., & Möller, H. (1986). Tables and procedures for the determination of power and sample sizes in univariate and multivariate analyses of variance and regression. *Biometrical Journal, 28,* 647–663.

Harter, H. L., (1969). *Order statistics and their use in testing and estimation: Vol. 1. Tests based on range and studentized range of samples from a normal population.* Washington, DC: U.S. Government Printing Office.

Katzer, J., & Sodt, J. (1973). An analysis of the use of statistical testing in communication research. *The Journal of Communication, 23,* 251–265.

King, D. S. (1985). Statistical power of the controlled research on wheat gluten and schizophrenia. *Biological Psychiatry, 20,* 785–787.

Kroll, R. M., & Chase, L. J. (1975). Communication disorders: A power analytic assessment of recent research. *Journal of Communication Disorders, 8,* 237–247.

Lovie, A. D. (1979). The analysis of variance in experimental psychology: 1934-1945. *British Journal of Mathematical and Statistical Psychology, 32,* 151–178.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64,* 553–557.

Murray, L. W., & Dosser, D. A., Jr. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology, 34,* 68–72.

Nunally, J. C. (1975). *Introduction to statistics for psychology and education.* New York: McGraw-Hill.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

O'Neill, R., & Wetherill, G. B. (1971). The present state of multiple comparison methods (With discussions). *Journal of the Royal Statistical Society, Ser. B, 33,* 218–250.

Ottenbacher, K. (1982). Statistical power and research in occupational therapy. *Occupational Therapy Journal of Research, 2,* 13–25.

Pearson, E. S. (1939). "Student" as statistician. *Biometrika, 30,* 210–250.

Petrinovich, L. F., & Hardyck, C. D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erraneous conclusions. *Psychological Bulletin, 71,* 43–54.

Rucci, A. J., & Tweney, R. D. (1980). Analysis of variance and the "second discipline" of scientific psychology: A historical account. *Psychological Bulletin, 87,* 166–184.

Sawyer, A. G., & Ball, A. D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research, 18,* 275–290.

Stegmüller, W. (1973). *"Jenseits von Popper und Carnap": Die logischen Grundlagen des statistischen Schliessens.* Berlin: Springer-Verlag.

Trachtman, J. N., Giambalvo, V., & Dippner, R. S. (1978). On the assumptions concerning the assumptions of a *t* test. *The Journal of General Psychology, 99,* 107–116.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.

Westermann, R., & Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11,* 117–146.

Received July 13, 1987
Revision received March 28, 1988
Accepted June 7, 1988 ■