University of Rhode Island

DigitalCommons@URI

Open Access Dissertations

1984

Statistical Power of Psychological Research: The Artifactual Basis of Controversial Results

Joseph S. Rossi

University of Rhode Island

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Recommended Citation

Rossi, Joseph S., "Statistical Power of Psychological Research: The Artifactual Basis of Controversial Results" (1984). *Open Access Dissertations*. Paper 962.
https://digitalcommons.uri.edu/oa_diss/962

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

STATISTICAL POWER OF PSYCHOLOGICAL RESEARCH:
THE ARTIFACTUAL BASIS OF CONTROVERSIAL RESULTS

BY

JOSEPH S. ROSSI

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

1984

ABSTRACT

The statistical power of psychological research was examined in three studies. The purpose of Study I was to provide a general assessment of statistical power in psychological research. All of the studies using statistical tests in the 1982 issues of the Journal of Abnormal Psychology and the Journal of Personality and Social Psychology (N = 142) were evaluated and compared to the results of a previous survey. Results indicated that the power of psychological research has increased over the past 20 years but that power must still be characterized as low. These results are limited because accurate estimates of effect size, on which power analysis must be based, are not given in most research reports. The purpose of Studies II and III was to use recently devised techniques of meta-analysis to provide estimates of effect size for subsequent power analyses. Use of this procedure allowed power analyses to be conducted using precise estimates of effect size, rather than a wide range of effect sizes, as in previous power surveys. In Study II, a meta-analysis of 48 studies on the spontaneous recovery of verbal learning was conducted. This area of research was selected because previous studies of spontaneous recovery have produced conflicting results concerning the existence of the phenomenon. Results indicated that the evidence for spontaneous recovery is statistically significant, but that the magnitude of the effect is small. A power analysis was

then performed on the spontaneous recovery studies using the effect size information obtained by the meta-analysis. Results indicated that the power of research on spontaneous recovery was low, and implied that contradictory findings may be an artifact of low statistical power. This implication was examined in Study III. Power analyses were conducted on each of 12 published meta-analyses, all of which had been performed on research literatures characterized by inconsistent results. The results of Study II were generally confirmed. For six of the meta-analyses statistical power was similar to the proportion of significant results. In only three of the meta-analyses was power relatively high ($> .75$), while for the remaining meta-analyses power was very low ($< .55$). The low power of these research areas is due to the small size of the effects that were investigated. Average effect size (η^2) across all meta-analyses was only .019. For effects of this magnitude only very large sample sizes will yield acceptable levels of statistical power (e.g., .80). It is recommended, therefore, that psychologists increase power by making effect sizes larger, that is by increasing the signal-to-noise ratio of research designs. Ultimately desirable is the construction of "strong" theories, which would enable researchers to make specific predictions concerning the magnitudes of the effects they study.

DEDICATION

To my parents

For lighting the spark of my curiosity so many years ago.

ACKNOWLEDGEMENTS

First, I'd like to acknowledge the time, advice, and friendship of my committee members: Dr. Charles Collyer, Dr. Peter Merenda, Dr. Janet Kulberg, Dr. Choudary Hanumara, Dr. Susan Brady and Dr. Ray Hurley. I'd especially like to thank Dr. Collyer and Dr. Merenda for their patience with my slow progress: their help and friendship has meant a great deal to me. Beyond their help on my research, they have taught me what psychology is all about. Although not a committee member, Dr. Wayne Velicer has also been a friend and mentor throughout my graduate career. The friendship of many of my fellow students has made graduate school a truly worthwhile experience. Especially I'd like to thank Dr. Pat Seymour, Dr. Kathi O'Neill, Dr. Sandy Neill, Sally Cottrill, Ed Guadagnoli, Dr. Steven Ginpil, and Dr. Gene Heitz. I'd also like to thank Sally Cottrill for helping in the collection of data for Study I. My deepest and most lasting appreciation goes to my wife Sue. She participated in data collection for Study I, typed large portions of the original draft manuscript, and has single-handedly prepared the final version of the manuscript, which required her to learn a new computer system under considerable time pressure. Most of all, she had to put up with me while my concerns seemed focused almost entirely on my work. Her efforts and sacrifice have made it possible.

PREFACE

It often happens in science that the actual conduct of research is not as orderly and straightforward as the presentation of the results might suggest. The research presented here is no exception. Study II was conducted first. As a result of a seminar on human learning and memory I wrote a paper reviewing the literature on the spontaneous recovery of verbal associations (Rossi, 1977). The experience left me puzzled: why should so many well-designed studies fail to produce a clear consensus among researchers? I concluded, with the majority of reviewers, that spontaneous recovery did not exist, or if it did, that it was too ephemeral to be meaningful. I first became aware of the techniques of meta-analysis several years later. Here, it seemed, was a quantitative means of demonstrating spontaneous recovery's lack of robustness. Instead, the statistical evidence for the effect emerged strongly. Only when the small magnitude of the effect was revealed did I begin to suspect the role statistical power might play in generating controversy among researchers. Study III followed naturally since by 1982 a sufficient number of studies had been published to make a power analysis of meta-analyses a feasible project. Study I was conceived last, almost as an afterthought since, as I have explained to friends, "I needed an EPA paper for 1984 to justify going to Baltimore and eating

crab cakes for four days." Study I was then included in the dissertation since it seemed an appropriate prelude to Studies II and III. Portions of all three studies have been presented at professional conferences: Study I at the 55th Annual Meeting of the Eastern Psychological Association, Baltimore, MD, April 1984; Study II at the 53rd Annual Meeting of the Eastern Psychological Association, Baltimore, MD, April 1982; Study III at the 54th Annual Meeting of the Eastern Psychological Association, Philadelphia, PA, April 1983; and portions of all three studies at the 150th National Meeting of the American Association for the Advancement of Science, New York, NY, May 1984.

31 May 1984

TABLE OF CONTENTS

ABSTRACT	ii
DEDICATION	iv
ACKNOWLEDGEMENTS	v
PREFACE	vi
TABLE OF CONTENTS	viii
LIST OF TABLES	x
1. STATISTICAL POWER	1
Summary	12
2. ROLE OF EFFECT SIZE IN STATISTICAL POWER ANALYSIS	14
Why Effect Sizes are not Reported	16
Availability of Information on Effect Size	22
Cohen's Definitions of Effect Magnitude	37
Criticisms of Effect Size Measures	46
Effect Size as an Index of Practical Significance	59
Summary	65
3. STUDY I: POWER SURVEY OF CURRENT PSYCHOLOGICAL	
RESEARCH	68
Cohen's Power Survey	68
Results of Other Power Surveys	71
Power Survey of Current Psychological Research	79
Method	85
Results	104
Discussion	107

4. STUDY II: STATISTICAL POWER AND ARTIFACTUAL	
CONTROVERSY IN RESEARCH ON SPONTANEOUS RECOVERY	115
Limitations of Power Surveys	115
Method	126
Results	133
Discussion	137
5. STUDY III: INADEQUATE STATISTICAL POWER AND	
ARTIFACTUAL CONTROVERSY IN PSYCHOLOGICAL RESEARCH	144
Method	145
Results and Discussion	150
6. BEHAVIORAL AND PHYSICAL SCIENCES COMPARED	159
7. SUMMARY OF FINDINGS	173
REFERENCES	177
APPENDIX	223

LIST OF TABLES

1. Possible Outcomes of Statistical Test Decisions	1
2. Outcomes of Statistical Test Decisions Against a Range of Effect Sizes	41
3. Power Analysis for a Correlation Problem	43
4. Effect Size Definitions in Terms of Proportion of Accounted for Variance	45
5. Power Surveys of Behavioral Research	73
6. Results of Power Surveys of Behavioral Research	77
7. Percentage of Studies with Power less than .80 and .50	80
8. Statistical Tests Included in the Power Survey	87
9. Frequency Distribution for Tests in Power Survey	105
10. Average Power of 1982 Studies Compared to Cohen's (1962) Results	109
11. Percentage of Studies with Power $< .50$ and $< .80$	109
12. Average Power of 1982 Studies Compared to Other Power Survey Results	110
13. Percentage of Studies with Power $< .50$ and $< .80$	110
14. Average Power of Spontaneous Recovery Studies	137
15. Meta-Analyses Included in Study III	148
16. Power Analysis of Meta-Analyses	151
17. Power Surveys of Meta-Analyses	156

CHAPTER 1

Statistical Power

Researchers in the behavioral sciences often depend upon statistical inference in order to decide whether or not an experimental effect is present. Since, as a result of an experimental manipulation, an effect will either be present or not present, and because the results of a statistical test will either support or reject the null hypothesis of no effect, there are four possible outcomes in any experimental situation (see Table 1).

Table 1

Possible Outcomes of Statistical Test Decisions

Statistical Test Decision	State of the Population	
	Effect Absent	Effect Present
Reject Null Hypothesis	Type I Error ($p = \alpha$)	Power ($p = 1 - \beta$)
Accept Null Hypothesis	Correct Decision ($p = 1 - \alpha$)	Type II Error ($p = \beta$)

1. There is no effect, but the statistical test decision is to reject the null hypothesis. This is a type I error and it occurs with probability α .

2. There is no effect and the statistical test decision

is to accept the null hypothesis. This correct decision occurs with probability $1 - \alpha$.

3. An effect is present, but the statistical test decision is to not reject the null hypothesis. This is a type II error and it occurs with probability β .

4. An effect is present, and the statistical test decision is to reject the null hypothesis. This correct decision occurs with probability $1 - \beta$.

Closely related to the correct rejection of the null hypothesis (and often equated with it) is the concept of statistical power. The power of a statistical test is the probability that the test will yield statistically significant results. Three factors determine the statistical power of an experiment.

1. Significance level: Power increases as a researcher's willingness to accept a type I error increases. This factor includes the related issue of directional vs. nondirectional tests. While directional tests are (or were) considered somewhat controversial (e.g., Burke, 1953, 1954; Eysenck, 1960; Goldfried, 1959; Jones, 1952, 1954; Kaiser, 1960), they are nevertheless more powerful than nondirectional tests.

2. Sample size: Power increases as sample size

increases.

3. Effect size: Power increases as the magnitude of the effect being studied increases.

Since the aim of behavioral research is to discover important relationships between variables, a consideration of power might be regarded as a natural and important part of the planning and interpretation of research. Cohen (1977) has expressed the point aptly: "Since statistical significance is so earnestly sought and devoutly wished for by behavioral scientists, one would think that the a priori probability of its accomplishment would be routinely determined and well understood" (p. 1). This seems not to be the case among behavioral researchers even though a routine consideration of power may have several beneficial consequences.

1. Most directly, of course, a knowledge of the power of a statistical test gives an indication of the likelihood of obtaining statistically significant results. Presumably, most researchers would not knowingly conduct an investigation of low statistical power (e.g., $1 - \beta < .50$) unless there were absolutely no alternative (one might always get lucky, after all!). If a priori power estimates are low, the researcher may elect to increase power by manipulating one or more of the three factors listed

previously. Alternatively, the researcher may decide to abandon the proposed research altogether if the costs of increasing power are too high, or if the costs of conducting research of low power cannot be justified.

2. Knowledge of the power of a statistical test greatly facilitates the interpretation of null results. It is often stressed in statistics textbooks, though seldom adhered to in practice, that the failure to reject the null hypothesis does not require the acceptance of the null hypothesis, that is, does not constitute evidence that the null hypothesis is true. The most that can generally be claimed in this situation is that there is not sufficient evidence to reject the null hypothesis. However, if the power of the test is known to be high, then failure to reject the null hypothesis can in a sense be considered as an affirmation of the null hypothesis. This follows from the fact that if power is high, then the probability of a type II error must be low. Thus, in the same way that a statistically significant result allows a researcher to confidently reject the null hypothesis with only a small probability of error (α), high statistical power affords a researcher the luxury of accepting the null hypothesis with a relatively small probability of error

(beta). If statistical power is low, one is no worse off than before, and can furthermore assert that there was not a fair chance of rejecting the null hypothesis, so that failure to reject may not weigh so heavily against the alternative hypothesis (the "alibi" use of power). Both of the lines of reasoning put forth here have limits, of course, and these limits have to do with the concept of "effect size," which will be more fully explained in Chapter 2.

3. Beyond considerations of individual research results, attention to statistical power may provide insight concerning entire research domains. In particular, when low statistical power characterizes an entire research literature, then the veracity of even statistically significant results may be questioned. This follows from the fact that when power is low the probability of rejecting a true null hypothesis may be only slightly smaller than the probability of rejecting the null hypothesis when the alternative is true. Thus, a large proportion of published significant results may be type I errors. This point will be discussed in greater detail in Chapter 3.

This brief survey of the value of power analysis raises the question of why behavioral researchers ignore an issue

of such clear and obvious merit. Indeed, the benefits of a routine consideration of statistical power may strike some readers as almost too good to be true. Nevertheless, it is probably not an exaggeration to assert that most behavioral researchers know little more about statistical power than its definition. It will certainly surprise no one to claim that, historically, behavioral researchers have been primarily concerned with the type I error and the associated concept of the "significance" of a statistical test and have largely ignored the type II error and its concomitant, the power of a statistical test (Bakan, 1966; Chase & Tucker, 1976; Cohen, 1977; Cowles & Davis, 1982b; Hogben, 1957).

The result has been that, while the concept of statistical significance is well-known in psychology, that of statistical power is not. The latter concept has been known to mathematical statisticians for a long time, and was introduced in a famous series of papers by Neyman and Pearson (1928a, 1928b, 1933a, 1933b, 1936a, 1936b, 1938, 1967). Their aim was to develop a sound mathematical theory of statistical hypothesis testing. In so doing Neyman and Pearson introduced such now familiar concepts as power, confidence intervals, and type I and type II errors.

(It is true, however, that others had already recognized the importance of the two types of errors in making statistical decisions. Perhaps the earliest was Condorcet late in the eighteenth century (Taylor, 1959).)

One major impetus to the development of the Neyman-Pearson theory was the asymmetry in the traditional approach to statistical inference. In formulating the null hypothesis, Fisher established no specific alternative hypothesis. No particular problem results from this approach if the statistical test is significant: the researcher simply concludes that the effect being studied is not zero, that is, that the effect is present.

The asymmetry arises when the statistical test fails to reject the null hypothesis. Under these circumstances it is not possible to conclude that there is no effect (although just such a conclusion is often made--implicitly if not explicitly--in the discussion of results by researchers), only that there is no evidence to reject the null hypothesis. This follows directly from the Fisherian procedure of not stating an alternative hypothesis. From the Fisherian viewpoint, the p value associated with a statistical test represented the confidence with which one could reject the null hypothesis. Thus, the traditional interpretation of the classical significance test concerned

the validity of the null hypothesis, as revealed by the obtained results. A p value of .01 meant simply that there was only one chance in 100 that the null hypothesis was valid. This retrospective interpretation of the null hypothesis in probabilistic terms has been called the "Backward Look" (Hogben, 1957). There was no specific mention or appreciation of alpha levels or type I errors, nor of course, of type II errors (Bakan, 1966; Chase & Tucker, 1976; Cohen, 1977; Hogben, 1957; Taylor, 1959).

While Fisher seems to have generally preferred a significance level of .05, he did not maintain that a particular level of significance be specified prior to the collection of data. Although this may now seem to be an odd practice, it is not particularly surprising given that, for Fisher, the interpretation of results was retrospective. In contrast, the Neyman-Pearson approach to hypothesis testing demanded the a priori specification of significance level, that is, of alpha, since in the Neyman-Pearson approach, hypothesis testing did not constitute a retrospective evaluation of the null hypothesis, but rather was an expectation of future outcomes. No single experimental result can determine the validity of the null hypothesis. The fate of the null

depends instead on the outcome of future experiments. This was the "Forward Look" (Hogben, 1957), and though the difference between this approach and that of the "Backward Look" may appear subtle and of philosophical interest only, it was the occasion of much acrimony in the 1930's, especially among British statisticians (Bakan, 1966; Hogben, 1957; Pearson, 1970; Reid, 1982; Taylor, 1959).

The critical step taken by Neyman and Pearson was the proposal to use a specific alternative hypothesis, that is, one in which the expected departure from null is explicitly stated. This step enabled them to think in terms of both the confidence with which the null hypothesis could be rejected (which they called the type I error probability) and the confidence with which the null hypothesis could be accepted. The Neyman-Pearson approach also enabled the researcher to state in advance the probability of incorrectly accepting the null hypothesis (which they called the type II error probability) and the probability of correctly rejecting a false null hypothesis, that is, to state in advance the power of a statistical test. Thus, it is the Neyman-Pearson approach to hypothesis testing which has resulted in the modern conception of the possible outcomes of the experimental situation shown in Table 1.

The relationships shown in Table 1 are, of course, well

known to behavioral researchers, since they are included in most introductory statistics textbooks. Unfortunately, most behavioral researchers are exposed to little more than what is presented in such texts. Coverage is more extensive in graduate-level handbooks (e.g., Glass & Stanley, 1970; Hays, 1981; Keppel, 1982; Winer, 1971), though rarely runs to more than a few pages. Even so, discussion is usually limited to the power of the t test and one-way analysis of variance. Issues of power as related to more complex designs and to more sophisticated statistical techniques--such as factorial, repeated measures, and multivariate designs, and nonparametric methods--are to be found only in the technical journal literature or in textbooks of mathematical statistics, both of which are likely to be inaccessible to the majority of behavioral researchers.

Few general handbooks of statistical power are available. Odeh and Fox (1975) is appropriate for statistical tests arising from the general linear model, but their procedures are probably more complicated than most behavioral scientists would be willing to tolerate. Furthermore, their charts are designed primarily for the determination of sample size, and are not easily adapted to

estimating power. Mace (1964) was intended primarily for engineers and seems to be unknown to psychologists, although many of the techniques included in this text are used in behavioral research. Sample size calculations are its main intent, although formulas for determining power are included.

Cohen's (1969, 1977) text remains the only one devoted entirely to statistical power analysis written specifically for behavioral researchers. Extensive tables (over 60 pages devoted to the analysis of variance alone) are included for both sample size and power determination. Included are chapter length considerations of the following statistical tests: t test, Pearson r, z test for differences between correlation coefficients, sign test and the test that a proportion is .50, z test for differences between proportions, chi-square test for goodness-of-fit and for contingency tables, F test in the analysis of variance and covariance, and (in Cohen, 1977, only) F test in multiple regression/correlation analysis. Power values are given for a wide range of sample sizes and effect sizes and for three levels of alpha (.01, .05, .10). The tabled values are given to two decimal places and are accurate to about one digit in the last decimal place when compared to exact values (e.g., Tang, 1938). A technical appendix

provides computational formulas for the calculation of power and sample size for most of the statistical tests covered.

Cohen's (1962, 1965, 1969, 1977) discussions of statistical power were not the first aimed at the general community of psychological researchers (e.g., Mosteller & Bush, 1954). Nevertheless, it has been Cohen's work more than any other which has been largely responsible for what increase in interest in power among behavioral researchers has occurred over the past 20 years (e.g., Cowles, 1974; Hinkle & Oliver, 1983; Koele, 1982; Levine & Dunlap, 1982; Lipsey, 1983; Medler, Schneider, & Schneider, 1981; Sacco, 1982; Schwartz & Dalglish, 1982; Trattner & O'Leary, 1980; Westermann & Hager, 1983).

Summary

Statistical power depends on three factors: significance level, sample size and effect size. The basic concept of statistical power, as embodied in the four-fold decision matrix of test outcome (see Table 1), is well-known to behavioral researchers. Nevertheless, routine use of statistical power in the planning and interpretation of research is not a common practice among psychologists. This is unfortunate, since consideration of power not only

provides an indication of the likelihood of obtaining statistically significant results, but also greatly facilitates the interpretation of null results.

There are undoubtedly many reasons why psychological researchers do not devote more attention to issues of statistical power. Journal editors must ultimately bear some of the responsibility, but until the appearance of Cohen's (1969, 1977) handbook no convenient introduction to the topic of power was available. Coverage of statistical power in most of the graduate-level texts used by psychologists has been perfunctory at best. Detailed treatment of power has, of course, always been available in the journal literature, especially of mathematical statistics, but seems not to have "trickled down" to behavioral researchers to any great extent. The problem here probably has a great deal to do with the concept of effect size. The role of effect size, and why it has been a barrier to the use of statistical power, will be discussed in Chapter 2. Finally, one encouraging recent development may be noted: that is the inclusion in several introductory textbooks of entire chapters on power analysis (Howell, 1982; Jaccard, 1983; Minium & Clarke, 1982; Welkowitz, Ewen, & Cohen, 1982).

CHAPTER 2

The Role of Effect Size in Statistical Power Analysis

One major factor contributing to the widespread disinterest in statistical power among behavioral scientists is the concomitant disinterest in and ignorance of effect size statistics. The magnitude of the effect under investigation may be thought of as an index of the extent to which the null hypothesis is not valid. The departure from null, called the noncentrality parameter in the technical literature, is one of three factors upon which the power of a statistical test depends.

Significance level and sample size are under the direct control of the experimenter. Significance level is strongly constrained by convention at no greater than $\alpha = .05$, so that the experimenter's control here is largely illusory. The major constraint on sample size is generally one of time and money (one rarely encounters a researcher who complains of having too many subjects and too much time or money). These two factors are often restrictions on increasing the power of an experiment. However, since both factors must be addressed by researchers before their research can be initiated, neither significance level nor sample size can be barriers to the routine consideration of

experimental power.

The role of effect size is fundamentally different. This factor is not under the control of the researcher. In fact, the magnitude of the effect being studied is usually unknown to the researcher, its determination often being the primary reason for which the research is conducted (at least to the extent that the effect may be said to exist or not exist). This may seem to be a problem for power analysis, but is really only a problem when an investigator is researching a totally new phenomenon, an unlikely circumstance for behavioral research (and by some accounts, for science in general (Kuhn, 1970)).

For most behavioral research there is generally enough previous related research to enable the investigator to estimate effect size reasonably, at least for the purposes of providing a preliminary estimate of experimental power. Yet, this is not done, primarily for two reasons: (a) Estimates of effect size are rarely explicitly stated in published research (Cohen, 1965, 1977; Cowles, 1977; Craig, Eison, & Metze, 1976; Haase, Waechter, & Solomon, 1982; Katzer & Sodt, 1973); and (b) Most behavioral researchers do not know how to calculate effect size from the kinds of data which are included in published reports. Furthermore, many published reports do not include sufficient data to

enable even the motivated reader to calculate the obtained effect size (Katzner & Sadt, 1973). These two problems will be discussed in greater detail in the next two sections.

Why Effect Sizes are not Reported

Why are effect sizes not included in research reports? Slavish adherence to the significance testing routine in conjunction with ignorance of the appropriate computational procedures for determining effect size undoubtedly explains much. Particularly revealing is that many of the standard computer packages do not include effect size estimates among the default output (or even as options in some cases) for the most popular univariate statistics, such as Student's t test and the analysis of variance. For example, ω^2 --the most widely recommended measure of effect size for the analysis of variance--is not provided as output on analysis of variance programs provided by any of the three major statistical packages: BMDP (Dixon, 1981), SAS (Ray, 1982), and SPSS (SPSSX User's Guide, 1983). (SAS does provide R^2 (which is equal to η^2) as output, a reflection of its general linear model approach to the analysis of variance.)

So much has been written in the last 25 years concerning the shortcomings of using the probability value

of a significance test as an index of effect size that it seems pointless to elaborate further (Bakan, 1966; Carver, 1978; Cohen, 1965, 1977; Cook, Gruder, Hennigan, & Flay, 1979; Lykken, 1968; Meehl, 1967, 1978; Morrison & Henkel, 1970; Rozeboom, 1960; Wilson, Miller, & Lower, 1967). The most serious limitation, of course, is the dependence of significance level on sample size, so that the level of significance may be increased arbitrarily by adding observations. More likely to be forgotten is that the level of significance the researcher may claim depends on the alpha level set prior to the collection of data (Shine, 1980). At any rate, it is probably true that many researchers still equate significance level with the impact and importance of a treatment effect (Beauchamp & May, 1964; Cooper & Rosenthal, 1980; Meehl, 1978; Rosenthal & Gaito, 1963, 1964).

Even less charitable, perhaps, is the suggestion that effect size information may be suppressed in many research reports. Thus, a correlation of .45, in the context of a validity coefficient, may appear quite substantial. However, to discuss the relationship in terms of proportion of variance accounted for, and especially in terms of proportion of variance not accounted for (here about 80%) would undoubtedly diminish the magnitude of the

relationship in the minds of many readers.

Perhaps due to the ease of conversion from correlation coefficient to effect size measure (proportion of variance accounted for), those areas of behavioral research heavily invested in correlational methods seem more inclined to discuss results in terms of measures of association and seem to rely less on significance testing. Of greater concern is the widespread disavowal of correlational methods by researchers using experimental designs. Although this problem was noted many years ago in a classic paper by Cronbach (1957), little seems to have changed since. The problem is compounded here in that the popular analytic techniques (t test and analysis of variance) are not as easily transformed into an effect size estimate as are correlation coefficients. Certainly, $t(100) = 3.0$, $p < .01$ suggests a more successful experimental outcome than its equivalent, $\omega^2 = .07$, which reveals that more than 90% of the variance in the dependent variable has not been accounted for by the experimental manipulation.

The situation is charged with irony. For example, the computational effort required to determine the proportion of variance accounted for in a t test or analysis of variance is a great deal less than for the significance

test itself! For the t test, the relationship is simply

$$\omega^2 = (t^2 - 1) / (t^2 + \underline{N} - 1) , \quad (1)$$

where N is the total sample size (Hays, 1963). The relationship is even simpler for another popular proportion of variance statistic, η^2 :

$$\eta^2 = t^2 / (t^2 + df) , \quad (2)$$

where df represents the degrees of freedom associated with the significance test (Kerlinger, 1964).

Concerning the rift between the "two disciplines of scientific psychology" (Cronbach, 1957), it is perhaps sufficient to note that the analysis of variance and regression/correlation methods--far from being different techniques--both arise from the general linear model (Cohen, 1968; Cohen & Cohen, 1983). As Haggard (1958) has aptly noted, most significance tests may be thought of as ratios of controlled plus uncontrolled variance to uncontrolled variance, to test whether controlled variance is greater than would be expected by chance. Measures of association may be thought of as the ratio of controlled variance to total variance (controlled plus uncontrolled).

The similarity between significance tests and measures of association is thus greater than most behavioral researchers may recognize. The differences between these measures reflect the different questions each was designed to answer. Further irony is added when it is noted that the analysis of variance was developed specifically out of Fisher's (1924, 1925) consideration of intraclass correlation (Haggard, 1958; Rucci & Tweney, 1980; Taylor, 1959).

Perhaps most ironic of all is the concern some researchers may have that reporting measures of association will debase their findings. The root of this concern is probably the lack of a standard by which to measure the magnitude of effect size. In contrast to the situation for significance level, there are no traditional guidelines for judging an effect as large or small. Of course, to some extent this merely reflects the very real differences in the problems studied and methods used in a discipline as diverse as psychology. In laboratory fields such as psychophysics, a researcher might well be discouraged if no more than 80-90% of the variance could be accounted for. The ability of the investigator to fine-tune the parameters of the research greatly increases expectations about effect

size. In such research the significance test itself is sometimes not necessary.

The majority of research areas in psychology are not so fortunate, especially when only weak manipulations are used (where stronger manipulations might not be possible or ethically permissible), or where manipulation may not be possible at all (that hallmark of much current psychological research--gender--comes most readily to mind). That the magnitude of effect in such research is likely to be small is probably recognized by most researchers (e.g., Hyde, 1981; Maccoby & Jacklin, 1974; Rossi, 1983). But how small is a "small" effect? The irony here is that the reluctance of researchers to report measures of association has effectively masked the fact that effect sizes in most areas of psychological research are astoundingly small.

Very few surveys of the effect sizes obtained in published research have been conducted. The most extensive survey analysed 701 studies and 11044 statistical tests published between 1970 and 1979 in the Journal of Counseling Psychology (Haase, Waechter, & Solomon, 1982). The median proportion of variance accounted for (η^2) was .0830. Brown (1975) conducted a similar, but more limited survey of studies published in the American Educational

Research Journal from 1970 to 1974. Only correlation coefficients were recorded ($N = 369$); median r^2 was .051. Several other surveys have found somewhat larger average effect sizes, but all were based on very limited data, generally less than 100 statistical tests (Cooper & Findley, 1982; Craig, Eison, & Metze, 1976; Hamblin, 1971; Katzer & Sodt, 1973). These figures represent quite modest degrees of association and many researchers react with considerable surprise when informed that the average effect size in psychological research might be only of this magnitude. That this is not commonly appreciated is, of course, due to the fact that effect size is infrequently reported.

Availability of Information on Effect Size

The availability of information on effect size statistics closely resembles the situation for power analysis in general. This is not surprising given the intimate relationship between these two issues. Most introductory statistics textbooks for students in the behavioral sciences contain little or no information on the meaning or calculation of effect size statistics, except for the squared correlation coefficient r^2 . The situation has scarcely been any better in graduate level textbooks

until, with a few exceptions (e.g., Hays, 1963), quite recently (Howell, 1982; Keppel, 1982; Kirk, 1982).

This situation has developed despite the fact that many writers in the journal literature have emphasized the necessity of reporting measures of effect size as an aid to interpreting tests of significance (Bakan, 1966; Cohen, 1965; Cowles, 1977; Dunnett, 1966; Fleiss, 1969; Halderson & Glasnapp, 1972; Lykken, 1968; Vaughan & Corballis, 1969). This attention seems to have resulted in a surfeit of effect size indices, causing perhaps more confusion than anything else. The most important of these will be discussed in turn, including various proportion of variance statistics, noncentrality parameters, and several of Cohen's effect size indices. Several other infrequently utilized measures will not be discussed, including Friedman's (1968, 1969, 1982; Breen & Gaito, 1970) procedures, coefficients of utility (Bolles & Messick, 1958; Gaito, 1958; Gaito & Firth, 1973), nonparametric measures (Kramer & Andrews, 1982; Krauth, 1983; Serlin, Carr, & Marascuilo, 1982; Stavig & Acock, 1980; Strahan, 1982), and multivariate measures (Cramer & Nicewander, 1979; Huberty, 1983; Levy, 1967; Sachdeva, 1973; Serlin, 1982; Shaffer & Gillo, 1974).

Proportion of variance statistics. Measures of

association have had a long history in experimental research (Glass & Hakstian, 1969). Perhaps the most popular have been the "proportion of variance" statistics: η^2 , ϵ^2 , ω^2 , and intraclass correlation. All are ostensibly measures of the proportion of variance in the dependent variable that can be accounted for by the manipulation (or levels) of the independent variable. Alternatively, the proportion of variance accounted for may be thought of as the proportionate reduction in "uncertainty" (error variance) in the dependent variable that is afforded by a knowledge of the independent variable (Hays, 1963). Computationally, the measures are quite similar, so that for most practical situations all yield essentially the same results. The small differences between these measures reflect the different contexts within which each was developed.

η^2 is sometimes referred to as the correlation ratio, and as previously mentioned is equal to the R^2 statistic of multiple regression. Both η^2 and intraclass correlation were developed early in the century (i.e., before the development of the analysis of variance) in response to problems arising from Pearson's product-moment correlation coefficient (Haggard, 1958; Harris, 1913; Pearson, 1911).

Fisher (1924) first used them in connection with the analysis of variance, but neither was much used in behavioral research.

In the past 25 years both η^2 and intraclass correlation have been used increasingly in connection with the analysis of variance, particularly η^2 . Once used primarily as a measure of nonlinear correlation (Ghiselli, Campbell, & Zedeck, 1981; Nunnally, 1978; Winer, 1971), η^2 has become one of the most popular measures of effect size for the analysis of variance (e.g., Cohen, 1973a; Friedman, 1968; Haase, 1983; Kennedy, 1970; Kerlinger, 1964). Despite exhortations to the contrary (Haggard, 1958), intraclass correlation is rarely used except as a measure of the reliability of ratings (Bartko, 1966, 1974, 1976; Lahey, Downey, & Saal, 1983; Shrout & Fleiss, 1979). Recent textbook treatments of intraclass correlation reinforce this notion (Hays, 1981; Kerlinger, 1973; Winer, 1971).

Both ϵ^2 and ω^2 were developed specifically as measures of association within the context of the analysis of variance. Furthermore, while η^2 is a descriptive (sample) statistic, both ϵ^2 and ω^2 were developed as inferential techniques. Kelly (1935) proposed ϵ^2 as an unbiased alternative to η^2 . It was, however, not

completely successful in its stated purpose. At first popularized by Peters and van Voorhis (1940), ϵ^2 was soon criticized by Lindquist (1953) and was never generally accepted.

Perhaps the first index of effect size to gain some popularity among psychologists was ω^2 (Hays, 1963). Ω^2 and ϵ^2 are almost exactly identical computationally, although Hays (1963) seems not to have been aware of the earlier work. For the completely random effects model, ω^2 and intraclass correlation become identical.

Unlike ϵ^2 , ω^2 has received considerable attention, and detailed guidelines for its determination in various ANOVA designs (including factorial and repeated measures designs, and fixed, random, and mixed model designs) have been published in the most popular behavioral journals (e.g., Charter, 1982; Dodd & Schultz, 1973; Dwyer, 1974; Fleiss, 1969; Gaebeline, Soderquist, & Powers, 1976; Halderson & Glasnapp, 1972; Keren & Lewis, 1979; Susskind & Howland, 1980; Vaughan & Corballis, 1969). Nevertheless, coverage of ω^2 in graduate-level textbooks has been almost nonexistent until recently (Howell, 1982; Keppel, 1982; Kirk, 1982). Winer (1971), for example, does not

mention ω^2 at all. Coverage in undergraduate texts is likewise rare, but has become more frequent in recent years for both ω^2 and η^2 (e.g., Jaccard, 1983; Keppel & Saufley, 1980; Plutchik, 1983). Thus, it is not surprising that ω^2 is still reported infrequently in published research (Craig, Eison, & Metze, 1976).

More detailed comparison of all four proportion of variance statistics may be found in Camp and Maxwell (1983), Glass and Hakstian (1969), and Maxwell, Camp, and Arvey (1981).

Noncentrality parameters. One reason for the neglect of proportion of variance statistics by behavioral researchers may be that these measures have been ignored by mathematical statisticians as well. Fisher (1921, 1922, 1924, 1928) and others (e.g., Wishart, 1932) did much of the early work on the sampling distributions of intraclass correlation and η^2 . Research in this area is less frequent now (Carroll & Faden, 1978; Fleishman, 1980; Ranney & Thigpen, 1981). Moreover, very little is known of the sampling characteristics of ϵ^2 and ω^2 (Carroll & Nordholm, 1975; Demorest, 1983; Keselman, 1975; Lane & Dunlap, 1978).

Statisticians have preferred to think of "effect size" in terms of the noncentrality parameter. In fact,

statisticians do not think of effect size in the same way as do behavioral researchers. Following the theoretical lead of Fisher and especially of Neyman and Pearson, mathematical statisticians have developed the concepts of central and noncentral distributions for test statistics. The central distribution is that taken by a test statistic under the null hypothesis. Its parameters are its degrees of freedom. When the alternative hypothesis is true the test statistic follows a noncentral distribution. The parameters of a noncentral distribution are its degrees of freedom and its degree of "noncentrality," which characterizes the extent of the departure of a noncentral distribution from the central distribution. Thus, when a noncentrality parameter is equal to zero, equations for a noncentral distribution reduce to those for the corresponding central distribution.

The use of noncentrality parameters in statistical power analysis seems to be a barrier for many behavioral researchers. For example, tables and charts for the estimation of power for analysis of variance F tests are readily available in journal articles and general statistics texts (Dixon & Massey, 1983; Keppel, 1982; Kirk, 1982; Morrison, 1983; Overall & Dalal, 1965; Pearson &

Hartley, 1951; Rotton & Schonemann, 1978; Tang, 1938; Tiku, 1967; Winer, 1971). Yet these are rarely used by behavioral researchers. The chief difficulty is the requirement that the researcher be able to compute a noncentrality parameter so that power may be determined. Unfortunately, there are several problems associated with the use of noncentrality parameters from the point of view of behavioral researchers.

One is that the value of the noncentrality parameter in the analysis of variance (for example) is determined by several factors, including the magnitude of the treatment effect, the number of groups, sample size, and the population error variance. As a result, the noncentrality parameter does not have a clear and unambiguous operational interpretation, at least none beyond a rather general characterization as the departure from the null hypothesis. Lack of a clear conceptual definition for noncentrality parameters presents a problem for behavioral researchers who, if they are likely to think of effect size at all, are apt to think in terms of proportion of variance accounted for.

Another problem is that the calculation of noncentrality parameters requires the researcher to specify actual values for the population error variance and for the group means (or for the deviations of the group means from the population

grand mean). A noncentrality parameter may, of course, be expressed as a function of a proportion of variance statistic, a measure with which behavioral researchers are more comfortable and familiar. Unfortunately, though such relationships are not particularly complicated, they are rarely given explicitly either in texts or journal articles. Thus, the motivated researcher would have to derive the relationship through algebraic manipulation.

Yet another problem is that there are two different versions of the noncentrality parameter for the analysis of variance. These are λ and ϕ . The two are very simply related. Yet it is annoying and probably not a little discouraging for behavioral researchers attempting to determine statistical power for the first time. The noncentral χ^2 distribution (upon which the noncentral F distribution is based) was first derived by Fisher (1928), who gave β^2 as the noncentrality parameter. Others changed the notation to the now familiar λ (e.g., Neyman, 1937). Tang (1938), however, changed the operationalization of the noncentrality parameter to simplify the presentation of tabled power values for the analysis of variance. Tang (1938) called his noncentrality

parameter ϕ .

Both λ and ϕ are frequently encountered in current treatments of the power of analysis of variance F tests. Most presentations of tables and charts of power use Tang's ϕ (e.g., Dixon & Massey, 1983; Keppel, 1982; Kirk, 1982; Lehmer, 1944; Morrison, 1983; Odeh & Fox, 1975; Pearson & Hartley, 1951; Rotton & Schonemann, 1978; Tiku, 1967), although some use λ (e.g., Overall & Dalal, 1965). Technical expositions of the power of the analysis of variance usually refer to λ as the noncentrality parameter (e.g., Koele, 1982; Laubscher, 1960; Mead, Bancroft, & Han, 1975; Norton, 1983; Patnaik, 1949; Severo & Zelen, 1960), although some use ϕ (e.g., Levine & Dunlap, 1982; Nicewander & Price, 1983; Subkoviak & Levin, 1977; Tiku, 1966). Charts for sample size selection (curves of constant power) use neither λ nor ϕ , but use instead a modified ϕ called ϕ' (Feldt & Mahmoud, 1958; Lindman, 1974; Winer, 1971)!

Notational differences among authors is a commonplace in statistical work and presents little difficulty for the expert, but can be an obstacle for those new to the area. Moreover, it is important to remember that the differences between λ and ϕ (and ϕ') are not simply notational. The beginner may not even be able to overcome

these difficulties by consulting the most popular graduate-level texts. Kirk (1982) and Keppel (1982) discuss only ϕ . Lambda is used by Hays (1981), but is referred to as δ^2 ! Winer (1971) makes reference to all three versions of the noncentrality parameter. Ironically, one of the earliest discussions of statistical power employed η^2 as the index of departure from the null hypothesis (Neyman & Pearson, 1931).

Cohen's effect size indices. In his attempts to bring statistical power analysis to the behavioral sciences, Cohen (1962, 1965, 1969, 1977) has probably done more than any other writer to popularize the use of effect size statistics. Cohen recognized that the lack of a conceptually simple index of effect size was a major barrier to the use of statistical power by behavioral researchers. Cohen's (1969, 1977) power handbook has had its greatest impact through his introduction of simple indices of effect size for several statistical techniques and for his operational definitions of small, medium, and large effects.

Most influential has been Cohen's d , introduced as an index of effect size for the t test. d is simply the standardized difference between group means. It bears a

simple relationship both to δ , the noncentrality parameter of the power function of the t test (Hays, 1981; Laubscher, 1960; Winer, 1971), and to η^2 . Despite the similarity, Cohen's d has become more popular than either δ or η^2 , no doubt largely due to its simple interpretation as the difference between group means expressed in standard deviation units. Cohen's d is thus very similar to the familiar z score (and for that matter, to t itself), both in form and interpretation.

Cohen's effect size index for the analysis of variance is f , a generalized version of d . Cohen's f is the standard deviation of the k standardized population means. As with d , f is closely related to the noncentrality parameter (both λ and ϕ) and to η^2 . (In fact, f is equal to ϕ .) However, since the interpretation of f is with respect to more than two group means, proportion of variance statistics will probably continue to provide the clearest assessment of effect size for the analysis of variance.

Cohen rejected the noncentrality parameter as an index of effect magnitude opting instead for simple indices with clear operational interpretations. There is a price to be paid for simplicity, however, and that is generality. Since many of the statistical techniques covered in Cohen's

handbook arise from the general linear model, treatment of statistical power could be handled compactly if approached in terms of the parameters of the noncentral distribution, that is, degrees of freedom and noncentrality parameter. An example of this approach is the monograph by Odeh and Fox (1975), which contains about 50 pages of text and about 100 pages of nomographs.

Cohen, on the other hand, elected to tailor his treatment of statistical power to fit the statistical test under consideration. In some cases this meant introducing a new index of effect size (e.g., d for Student's t test), and in others using an already familiar measure (e.g., r for the Pearson product-moment correlation coefficient). In every case, however, the result was the decomposition of the noncentrality parameter into its component parts: sample size and an effect size index appropriate for the statistical test. (An exception was Cohen's (1977; Cohen & Cohen, 1983) treatment of multiple regression, which was more conveniently presented in terms of the noncentrality parameter λ .) Thus, Cohen required a separate chapter to "explain" the details of statistical power analysis for each statistical test covered. Furthermore, each chapter required its own tables of power and sample size, presented

in terms of the effect size index appropriate for that statistical test. Thus, conciseness had to be sacrificed: there are over 60 pages of tables for the analysis of variance alone. (By comparison, there are only 3 pages of tables for multiple regression, which retained the use of the noncentrality parameter.) There are 12 pages of power tables for the t test, even though these give exactly the same results as for the analysis of variance tables with one numerator degree of freedom. The separate tables for the t test are necessary because the effect size index for the t test differs from that used for the analysis of variance (although the relationship is simply $f = d / 2$).

The advantage of Cohen's approach is that his power tables are especially easy to use. Entry to most tables requires no more than a specification of sample size, effect size, and alpha level. What is demanded of Cohen's readers is that they become comfortable thinking in terms of his effect size indices. For those already accustomed to thinking in terms of proportion of variance statistics, conversion formulas and tables are provided.

Although Cohen's effect size indices may be artificial from a mathematical standpoint, they have been used with increasing frequency in recent treatments of power in the behavioral sciences (e.g., Barcikowski, 1973, 1981; Feldt,

1973; Hornick & Overall, 1980; Howell, 1982; Lipsey, 1983; Minium & Clarke, 1982; Welkowitz, Ewen, & Cohen, 1982). Furthermore, while no distributional work has yet been conducted on Cohen's f , research on the sampling distribution of Cohen's d has begun (Hedges, 1981, 1982, 1983; Kraemer, 1983; Rosenthal & Rubin, 1982a).

It must be admitted, however, that this work has not been motivated by considerations of statistical power analysis, but rather by the role of effect size indices in meta-analysis, a recently developed set of techniques for integrating the results of a series of independent experiments (Glass, 1976a; Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1978; see also Chapter 4). Integration is usually with respect to both significance level and effect size. While significance levels are usually reported in the articles being reviewed, effect sizes usually are not. Therefore, the reviewer wishing to conduct meta-analysis must compute the effect size for each study using information readily available in the research report. Cohen's d has thus become the preferred index of effect size for meta-analysis due to the availability of means and standard deviations in most research reports.

Although meta-analysis has been criticized by many writers, it has certainly increased among behavioral researchers an interest in and an awareness of effect size. Unfortunately, there is yet little evidence that an increased interest in and awareness of statistical power has also taken place. Any recent issue of the Social Science Citation Index will confirm that Cohen's (1969, 1977) texts are frequently cited. Ironically, most of these citations are from meta-analyses concerned with Cohen's effect size measures. Only a few sources are concerned with statistical power analysis.

Cohen's Definitions of Effect Magnitude

Nearly as influential as his introduction of simple measures of effect size were Cohen's (1969, 1977) definitions of small, medium, and large effect sizes. The .05 level has long been accepted as the standard for statistical significance in behavioral research (Cowles & Davis, 1982b). One reason behavioral researchers may be reluctant to use effect size statistics is the lack of a standard similar to that used for statistical significance to gauge the magnitude of effect size.

Since effect size is almost never reported in published research Cohen was unable to construct operational definitions of small, medium and large effect size on an

empirical basis. Thus, any such definitions would necessarily be arbitrary and probably unacceptable to many researchers. Nevertheless, Cohen felt the establishment of some standards would serve as an inducement for researchers to determine the statistical power of their research. That is, since information on effect size is generally not available in the literature, and since power cannot be determined without a specification of effect size, Cohen reasoned that even arbitrary guidelines would be better than none.

Since the index of effect size was different for each statistical test, the operational definitions of small, medium, and large effect size necessarily differed for each statistical test. The problem this situation creates is that of consistency of definition of effect magnitude across the different types of statistical tests, as we shall soon see. Note, however, that such definitions are not even possible using the noncentrality parameter as the index of effect size, since the value of the noncentrality parameter depends on sample size. In this sense, the noncentrality parameter is no better an index of the magnitude of a treatment effect than is significance level, since both depend on sample size!

Cohen (1969, 1977) gave operational definitions of small, medium, and large effects for each of the statistical tests described in his power handbook: t test, Pearson r , difference between two r 's, sign test, difference between two proportions, chi-square test, analysis of variance and covariance, and in Cohen (1977) only, F test in multiple regression/correlation analysis. A complete specification of these definitions will be given in a subsequent chapter. For now, two examples may be considered representative (Cohen, 1962).

1. For a t test of the hypothesis that the means of two independently drawn samples are equal, small, medium, and large effects were defined as mean group differences of 0.25, 0.50, and 1.00 standard deviation, respectively.

2. For the hypothesis that a correlation coefficient is zero, small, medium, and large effects were defined as correlations of .20, .40, and .60, respectively.

Although some have criticized the specific magnitudes chosen to characterize small, medium and large effects (e.g., Howell, 1982), placing effect size on a continuum does have advantages. Perhaps the most important is that it emphasizes that there is no single alternative to the null hypothesis. Rather, one may think of a series of alternative hypotheses, which differ in their degree of

departure from the null. Thus, it is an oversimplification to discuss the power of a statistical test as a singular attribute of the test. Instead, one may characterize the power of a statistical test as having a range of values against a series of alternative hypotheses. It may be especially profitable for behavioral researchers to think in these terms, since there is often little justification for selecting only a single effect size against which to compute power. Determining power for a range of possible alternatives, even if these are not necessarily those advocated by Cohen, will give the researcher a much better idea of the likelihood of obtaining significant results.

This discussion suggests a modification to the familiar decision matrix shown in Table 1, so that the state of the population effect is characterized by a range of values rather than as simply present or absent. Table 2 presents the modified version of Table 1.

Estimating power for a range of effect sizes will also greatly facilitate the interpretation of null results. For example, suppose that a researcher has obtained a nonsignificant test result and wishes to conclude that there is no correlation between two variables. Strictly

Table 2

Outcomes of Statistical Test Decisions Against a Range of Effect Sizes

Statistical Test Decision	Population Effect Size		
	Zero	Small	Medium Large
Reject Null Hypothesis	Type I Error	Low Power	Moderate Power High Power
Accept Null Hypothesis	Correct Decision	<----- Type II Error ----->	

speaking, it is not possible to "accept" the null hypothesis; all that the researcher can conclude in this situation is that there is insufficient evidence to reject the null hypothesis. However, if a power analysis is conducted, a stronger conclusion may be possible. Table 3 shows a power analysis for a sample size of 30 and for a two-tailed test ($\alpha = .05$) of the hypothesis that a population correlation is zero. The table reveals that our hypothetical researcher may feel reasonably sure that the population correlation is not as great as .50. If the researcher desires a confidence level of 95% (i.e., one equivalent to that typically used for the rejection of the null hypothesis), then only population correlations greater than .60 may be ruled out.

Table 3 also reveals that there is a substantial probability that a population correlation as great as .40 may have escaped detection. If our researcher has good reason to believe that the population correlation should have been greater than .50, or that a population correlation less than .40 would be meaningless, then the null results of the significance test are no less informative than if statistically significant results had been obtained. Only a power analysis can provide this kind of information.

Table 3

Power Analysis for a Correlation Problem

<u>r</u>	Power
.10	.071
.20	.177
.30	.358
.40	.595
.50	.816
.60	.951
.70	.994
Note. <u>n</u> = 30	
alpha = .05	

If, as is more likely to be true, our hypothetical researcher cannot assume a substantial population correlation, then the usual interpretation of null results must be invoked. In this situation the researcher may decide to (a) increase sample size, (b) increase alpha, (c) refine measurement techniques so as to reduce error and thereby increase the effective magnitude of the population correlation, or (d) abandon the research effort altogether. In any event, the results of this particular power analysis should impress upon the researcher the substantial sampling error of the correlation coefficient when sample sizes are small.

1962 vs 1977 definitions. The definitions for small,

medium, and large effect size described previously were given by Cohen (1962). With the publication of his power handbooks, Cohen (1969, 1977) revised his definitions of effect size for each statistical test. All revisions resulted in somewhat smaller definitions of effect size than originally established.

The changes in definition seem to have been made so that the magnitude of the effect implied by the various definitions would be more nearly equivalent across statistical tests (Cohen, 1969, 1970). These changes will be discussed in greater detail in a subsequent chapter. For now two examples will be given.

1. For the t test the definitions were changed from 0.25, 0.50, and 1.00 standard deviation to 0.20, 0.50, and 0.80 standard deviation for small, medium, and large effect sizes, respectively.

2. For Pearson's r the definitions were changed from .20, .40, and .60 to .10, .30, and .50 for small, medium, and large effect sizes, respectively.

For ease of comparison, several of Cohen's "1962" and "1977" effect size definitions have been converted to their equivalent proportion of variance statistics and displayed in Table 4. The proportion of variance statistic for the t

test is the point-biserial r^2 , for the analysis of variance is η^2 , and for Pearson's $\underline{r}$ is r^2 . Since the $\underline{t}$ test and the analysis of variance (necessarily) give the same results, they are listed together in Table 4.

Table 4

Effect Size Definitions in Terms of Proportion of
Accounted for Variance

Effect Size	<u>t tests/ANOVA</u>		Pearson r's	
	1962	1977	1962	1977
Small	.02	.01	.04	.01
Medium	.06	.06	.16	.09
Large	.20	.14	.36	.25

Note. Effect size definitions for 1962 and 1977 are from Cohen (1962) and (1977), respectively. The proportion of variance statistic for the $\underline{t}$ test is point-biserial r^2 , for ANOVA is η^2 , and for Pearson $\underline{r}$ is r^2 .

Two points may be noted. First, the definitions for $\underline{t}$ tests and Pearson r's are still not the same, though they are more nearly equal than in 1962. Cohen (1977) notes, however, that strict comparability is problematical, and probably not even desirable. For the $\underline{t}$ test the proportion of variance accounted for is based on the assumption that the groups belong to separate and distinct populations, and

is therefore equivalent to the point-biserial r^2 . The point-biserial r is always less than the Pearson r , which assumes a continuous distribution for both variables.

Applying a standard correction (e.g., Lindeman, Merenda, & Gold, 1980; McNemar, 1962) to the 1977 point-biserial r^2 's (and assuming equal sample sizes) yields values quite close to the 1977 Pearson r^2 's: .015, .092, and .216 for small, medium, and large effects, respectively.

The second point to be noted about Table 4 is the small size of all the definitions, even those called "large." Certainly, the proportion of variance not accounted for even by a large effect is substantial. Nevertheless, the values reported in Table 4 may well be representative of the range of effect sizes characteristic of behavioral research. As noted earlier, Haase, Waechter, and Solomon (1982) found a median η^2 of .083 for counseling psychology research, and Brown (1975) reported a median r^2 of .051 for educational research. Both figures agree well with Cohen's definition for a medium-sized effect.

Criticisms of Effect Size Measures

Psychologists have probably become accustomed to being admonished for their failure to use measures of effect size more frequently than they do. What may then come as a

surprise to many is that there exists a sizeable body of critical commentary on the use of such measures. Some comments have simply advised researchers of the limitations of effect size indicators and urged appropriately cautious interpretation (e.g., Maxwell et al., 1981; Mitchell & Hartmann, 1981; O'Grady, 1982). Other writers have maintained that effect size statistics should never be used (e.g., Dooling & Danks, 1975; Sechrest & Yeaton, 1982), while still others would permit their use only in very narrowly defined circumstances (Glass & Hakstian, 1969).

Several of the most important limitations of effect size statistics will be discussed in the following sections. Although these issues will be discussed as if they were separate problems, it should be clear that many are related. Discussion will be in the context of proportion of variance measures, since these are the most popular, have been used for the longest period of time, and have drawn the most critical commentary.

Technical problems. Both ϵ^2 and ω^2 were developed to correct estimation bias in η^2 and it has been pointed out that neither succeed in this effort (Glass & Hakstian, 1969). However, recent studies have indicated that the bias is small (Carroll & Nordholm, 1975; Demorest, 1983; Keselman, 1975). Moreover, one might simply "adopt

the view that even a biased estimator is better than none" (Vaughan & Corballis, 1969, p. 205).

Another technical issue, one debated primarily among proponents of proportion of variance statistics, concerns the computational formulas for these statistics. One aspect of this issue is how best to estimate variance components in mixed analysis of variance designs (Dodd & Schultz, 1973; Dwyer, 1974; Endler, 1966; Fleiss, 1969; Gaebelein et al., 1976; Vaughan & Corballis, 1969; Whimbey, Vaughan, & Tatsuoka, 1967), and in nonorthogonal designs (Bassler & Wolf, 1981; Keren & Lewis, 1976; LaTour, 1981a, 1981b; Wolf & Bassler, 1980). Another aspect of this problem is how to represent the total variance (Cohen, 1973a; Haase, 1983; Kennedy, 1970; Keren & Lewis, 1979). This problem will be discussed in the context of a specific issue (the "Number of Factors" problem) in a subsequent section. Critics have largely ignored these technical problems, though some have complained about the resulting variety of effect size measures (Maxwell et al., 1981; Mitchell & Hartmann, 1981; Sechrest & Yeaton, 1982).

Lack of work on the sampling distributions of ϵ^2 and ω^2 is another problem (Maxwell et al., 1981; Mitchell & Hartmann, 1981). What little work has been

conducted indicates that the sampling error for proportion of variance statistics may be substantial even for relatively large sample sizes (Carroll & Faden, 1978; Carroll & Nordholm, 1975; Demorest, 1983; Keselman, 1975; Lane & Dunlap, 1978). For example, when the proportion of variance accounted for in the population is .15, the standard error of ω^2 is .07 for the case of $n = 30$ for each of three groups (Carroll & Nordholm, 1975). Thus, even with a total sample size of 90, the 95% confidence interval for ω^2 in this case is .01 to .29!

More work on the sampling distributions of proportion of variance statistics needs to be done, especially for cases where the assumptions of the analysis of variance have not been met. Although the analysis of variance may be robust in many circumstances, proportion of variance statistics may not be. Unfortunately, the current popularity of Cohen's d for meta-analysis, and the distribution work this interest has engendered (Hedges, 1981, 1982, 1983; Kraemer, 1983; Rosenthal & Rubin, 1982a), may retard theoretical investigations of other effect size measures.

Publication bias. Even a casual perusal of the journal literature reveals that only a very small proportion of published studies fails to obtain statistically significant

results. Critics of significance tests have concluded that the paucity of nonsignificant results is not a reflection of the wonderful achievements of behavioral research, but rather is an indication of the editorial bias against the publication of null results (Bakan, 1966; McNemar, 1960; Smart, 1964; Sterling, 1959). Thus, studies reporting significant results are more likely to be published than studies not reporting such results.

One result of a publication bias against negative findings is that there may be more type I errors in the published literature than indicated by conventional alpha levels. Lane and Dunlap (1978) point out, however, that publication bias may affect estimates of effect size, as well, since larger effect sizes are more likely to be associated with significant results than are smaller effect sizes. Their simulation study revealed a substantial overestimation of effect size for a wide range of alpha levels (.001 to .20), population effect sizes (.01 to .20), and sample sizes (5 to 20 per group). As an example, suppose that the population effect size is equal to Cohen's definition for a medium-sized effect (proportion of variance accounted for is .06). Lane and Dunlap found that, when alpha was set at .05, the average proportion of

variance accounted for obtained in 5000 simulated t tests with 10 subjects per group was .26! For 20 subjects per group the overestimation was less but still substantial, with an average proportion of variance accounted for of approximately .15.

While these results are disturbing, in practice it is likely that the extent of the overestimation is less than that indicated by Lane and Dunlap (1978). Their results assume that publication decisions are based only on the obtained significance level. It is doubtful that anyone really believes that this is the case. Furthermore, their results showed a strong tendency to converge on the population effect size with increasing sample size. Unfortunately, sample sizes larger than 20 per group were not examined. Finally, since the obtained effect sizes in behavioral research are usually quite small, the extent of overestimation can not be very great.

Design type. Many critics have pointed out that proportion of variance statistics may produce misleading results because of their dependence on various design characteristics of a study. For example, whether a study is run with a between-groups or with a repeated measures design will greatly influence the proportion of variance accounted for (O'Grady, 1982). Repeated measures factors

almost always produce larger effects through the reduction of error variance. In addition, when both repeated measures and between-groups factors are included in the design, it is common for a significant repeated measures effect to account for less variance than is accounted for by a nonsignificant between-groups effect, again because of the smaller error term used for testing repeated measures factors.

Number of factors. The number of factors included in an investigation can also influence proportion of variance accounted for. Inclusion of a strong manipulation may prevent a second, but weaker, manipulation from accounting for as much variance as it would if the stronger factor were eliminated from the design. Interaction effects are frequently weak, since main effects will often "eat up" most of the reliable variance. Similarly, between-groups factors often account for a discouragingly small proportion of variance when repeated measures factors are also present in the design.

Some have proposed modifying the computational formulas for proportion of variance measures to remove the influence of one factor from another. The result has been measures such as partial η^2 (Cohen, 1973a; Haase, 1983; Kennedy,

1970), and partial ω^2 (Keren & Lewis, 1979). The variance due to a factor is compared to error variance and not to total variance, the variance due to other factors having been removed from the equation. Unfortunately, it is then possible for the sum of the proportions of variance accounted for by all effects to exceed unity as has been pointed out by several critics (Breen & Gaito, 1970; Sechrest & Yeaton, 1982).

Range of levels. The choice of specific levels for the independent variable may have a large impact on the resulting magnitude of the effect (Dooling & Danks, 1975; Fisher, 1946; Glass & Hakstian, 1969; LaTour, 1981a, 1981b; Levin, 1967; Lindquist, 1953; Maxwell et al., 1981; Mitchell & Hartmann, 1981; Norton & Lindquist, 1951; O'Grady, 1982; Sechrest & Yeaton, 1982). Widely spaced levels, often used in exploratory work, will produce a larger effect than more closely spaced levels (assuming a monotonic relationship between variables). As a restriction in range problem, only a small proportion of variance might be accounted for, although the relationship between the variables in the population might be substantial.

Alternatively, some critics fear that researchers may attempt to inflate their treatment effects through the inclusion of a group already known (or presumed) to have a

large impact. Inclusion of such a "good" group in the design guarantees that a larger proportion of variance will be accounted for, especially when treatment effects might otherwise be small. Such a "trick" would be easily discovered though. Post hoc examination of follow-up tests, or simply of group means and standard deviations, should reveal that a single group accounted for most of the reliable variance.

Fixed and random effects. The range of levels problem may also be thought of in terms of fixed and random effects. Several critics have argued that the fixed effects model is most frequently employed in behavioral research, but that only in the random effects case may the researcher freely generalize results to all levels of the independent variable (Dooling & Danks, 1975; Glass & Hakstian, 1969; Maxwell et al., 1981). These critics are concerned that researchers using fixed effects designs may generalize their obtained effect size beyond the specific levels of the independent variable. Of course, this problem applies to the analysis of variance F test as well, so that it is not clear why behavioral researchers should forget the distinction between fixed and random factors only when they begin reporting measures of effect size.

Certainly, as Halderson and Glasnapp (1972) point out, any researcher interested in using effect size statistics knows better than to make this mistake.

Ω^2 in particular was developed within the context of the fixed-effects analysis of variance model (Hays, 1963). However, several writers have pointed out that it is meaningless to speak of the total variance in the dependent variable as a population estimate, since there is no population which corresponds to the specific levels of the treatment groups (Dooling & Danks, 1975; Glass & Hakstian, 1969; Maxwell et al., 1981). Hays (1963, 1973) has argued that such a population estimate can be defined by assuming that the relative sample sizes of the treatment groups are proportional to their representation in the population. This solution, however, underscores the problematical nature of the use of ω^2 in nonorthogonal designs.

Subject characteristics. Several critics have pointed out that whether subject groups are heterogeneous or homogeneous will affect the magnitude of the treatment effect (Glass & Hakstian, 1969; Golding, 1975; O'Grady, 1982; Porter, Schmidt, Floden, & Freeman, 1978; Sechrest & Yeaton, 1982). Subject heterogeneity between groups maximizes treatment variance and thus tends to increase

effect size, while subject heterogeneity within groups increases error and so decreases effect size. Similarly, subject homogeneity within groups tends to increase effect size, while homogeneity between groups decreases effect size. The problem is clearly related to the problems of range of levels and restriction in range.

Carver (1975) has noted an interesting manifestation of this problem. He notes that items on achievement tests are selected to maximize individual differences, since items which are too easy or too difficult are deleted during test construction. When such tests are used in between-groups designs, variance due to individual differences becomes part of the error variance, so that the experimental treatment or manipulation cannot account for very much of the total variance.

Measurement error. Perhaps the most important factor to be considered in the interpretation of proportion of variance statistics is reliability. All researchers must confront the problem of measurement error in the response variable, but what many researchers may not realize is that measurement error applies to the independent variable as well. Precision in the operationalization of the treatment decreases error variance and so permits a greater

proportion of variance to be accounted for (Dooling & Danks, 1975; Glass & Hakstian, 1969; Mitchell & Hartmann, 1981; O'Grady, 1982; Porter et al., 1978; Sechrest & Yeaton, 1982). Measurement error is small when the independent variable can be defined in simple, quantitative terms. A classic example of a simple, quantitative variable may be found in "mental rotation" studies, where the independent variable is the angle of orientation (measured in degrees) of a stimulus figure (Shepard & Metzler, 1971). Of course, in most areas of behavioral research such simplicity is exceptional. O'Grady (1982) notes, for example, that there are at least eight different ways to define as simple a variable as subject sex!

Complex treatment variables are more typical of behavioral research. In general, when a larger number of characteristics are needed to define or operationalize a treatment there will be more potential sources of measurement error, which will decrease the reliability of the treatment and limit the proportion of variance that may be accounted for.

In many cases different researchers will operationalize what is ostensibly the same construct in different ways. For example, in how many different ways have researchers characterized high vs low "anxiety"? Under such

circumstances it may be meaningless to compare the effect sizes obtained by different researchers. In fact, it may be difficult or impossible to meaningfully compare effect sizes for different treatments even within the same study, since the treatments may, and usually will differ in degree of precision (Dooling & Danks, 1975). The root of the problem is that behavioral researchers usually lack an independent means of assessing the effectiveness of their manipulations (Sechrest & Yeaton, 1981, 1982).

The issue of reliability is more familiar to behavioral researchers in the context of the response or dependent variable. With respect to effect size, the problem is essentially the same as for the independent variable: measurement error limits the proportion of variance that may be accounted for (LaTour, 1981a, 1981b; Mitchell & Hartmann, 1981; O'Grady, 1982; Porter et al., 1978; Sechrest & Yeaton, 1982).

The extent of the effect of measurement error on effect size is easily specified if the reliabilities of the variables are known. Simply, the proportion of variance shared by two variables cannot exceed the product of their reliabilities (Ghiselli, Campbell, & Zedeck, 1981). Thus, it is not surprising that validity coefficients are seldom

much greater than .50 and are more often closer to .30. Furthermore, failure to obtain a treatment effect produces an ambiguous situation, notwithstanding considerations of statistical power: Is there no effect because there is no relationship between the variables, or because the reliability of one or both of the variables is low? Behavioral researchers often conclude that there is no relationship without sufficient consideration of reliability.

Effect Size as an Index of Practical Significance

Many proponents of the use of effect size statistics have characterized these measures as indices of practical significance. Since statistical significance is simply a function of sample size, the use of measures of practical significance is often seen as a safeguard against the acceptance of trivial differences between groups as "significant" or important. Thus, results that are "highly" significant (e.g., $p < .0001$) may be considered unimportant if the obtained effect size is small. Large effect sizes are, of course, to be interpreted as important. Some researchers have gone so far as to recommend that effect size statistics be interpreted regardless of the results of the significance test (e.g., Carver, 1978).

Unfortunately, as was pointed out in the previous section, effect size statistics have limitations, just as significance tests do, and these limitations decrease the extent to which measures of effect size may be useful as indicators of the importance of a treatment effect (Golding, 1975; LaTour, 1981a, 1981b; O'Grady, 1982).

Too great an emphasis on the proportion of variance accounted for by an effect may well undervalue its actual contribution, especially if one is inclined to note the proportion of variance not accounted for! Yet it is incorrect to suppose that 100% of the variance could be accounted for unless measurement error could be eliminated. In practice, unreliability will severely constrain the proportion of variance that can be accounted for, especially when the constructs being studied are assessed by single, rather than by aggregated measures.

Instead of considering the variance in an experiment as divided into two parts--variance accounted for and variance not accounted for--Sechrest and Yeaton (1982) suggest that total variance be conceptualized as divided into three parts: (a) reliable variance which can be accounted for by treatment effects; (b) reliable variance which cannot be accounted for by treatment effects, but which might be

attributable to variables not included in the design of the study; and (c) error variance, which is due to measurement error and thus unreliable by definition. Measures of effect size give an estimate of variance (a), and are contrasted with the sum of variances (b) and (c). Sechrest and Yeaton (1982) suggest, however, that variance (a) be compared only to variance (b), since the sum of variances (a) and (b) represent the total amount of variance in the system that can be accounted for. Their suggestion underscores the fact that a small effect size may be more an indication of measurement error than a lack of a relationship between variables. Unfortunately, there is generally no way to separate variance (b) from variance (c).

O'Grady (1982) has also emphasized that effect size may not always be a good indication of importance. A large effect size may be trivial if the problem addressed by the research is itself inconsequential, or if it is simply a demonstration of an obvious relationship. Judgment of the importance of an effect also depends on the type of research being conducted. If the goal of the research is prediction, then a large effect size is required. However, if the goal of the research is exploratory, then even a small effect size may help researchers understand the

relationships of the variables involved, and help guide the design of future experiments.

Specific examples of effects of small, medium, and large magnitude, according to Cohen's (1977) definitions (see Table 4), may be helpful in keeping the proportion of variance issue in perspective. Cohen (1977) has given the following examples:

1. Small Effects: the difference in mean IQ between twins and nontwins; the difference in mean height between 15 and 16 year old girls;

2. Medium Effects: the difference in mean IQ between professionals and managers; the difference in mean height between 14 and 18 year old girls;

3. Large Effects: the difference in mean IQ between holders of the Ph.D. degree and college freshmen; the difference in mean height between 13 and 18 year old girls.

Whether or not one is willing to accept these definitions is a matter of personal preference, and probably depends at least to some extent on one's area of specialization. Cohen's reasoning was simply that a medium-sized effect should be "large enough to be visible to the naked eye. That is, in the course of normal experience, one would become aware of an average difference

in IQ between . . . members of professional and managerial occupational group" (Cohen, 1977, p. 26). Large effects would seem to be "grossly perceptible" while small effects would not usually be discernable without statistical aid.

One particularly dramatic transformation of proportion of variance statistics is the "binomial effect size display" (BESD) recently proposed by Rosenthal and Rubin (1979, 1982b, 1982c; Rosenthal, 1982, 1983; Huberty & Holmes, 1983, independently came to similar conclusions; for a comment on and extension of the BESD see Preece, 1983). These writers note that a treatment which improves success rate from 35% to 65% accounts for only 9% of the variance; an improvement from 25% to 75% accounts for just 25% of the variance! (This work seems to have made quite an impression on applied researchers, but it is new only in the context of the point-biserial r . Similar work for the continuous case has been around for a long time: Curtis & Alf, 1968; Levy, 1967; Symonds, 1930; Tilton, 1937).

These illustrations suggest that the fact that an effect accounts for only a small proportion of total variance does not necessarily mean that it is a trivial effect. The importance of an effect cannot be judged by a proportion of variance statistic in isolation any more than by the sole consideration of a significance test result.

The role of the effect in theory, the nature of the investigation (e.g., exploratory vs. confirmatory) and its consequences, the reliability of the measures involved, as well as many other factors, must all be considered before the magnitude of an effect can be evaluated.

One final example, taken from an entirely different research domain, should provide a powerful illustration of the occasional importance of small effects. The point in its orbit at which the planet Mercury is closest to the Sun advances by 40 arc seconds per century (about .0001 degree per year) more than can be accounted for by orbital perturbations due to the other planets. Before its solution, many theories were suggested to explain the anomaly, but one especially captures the magnitude of this phenomenon: a modest change in the inverse-square law of gravitation so that the distance dependence relation becomes $r^{-2.0000001574}$ instead of r^{-2} can account for the discrepancy (Roseveare, 1982)! Of course, the ability of Einstein's general theory of relativity to account for the anomalistic advance of Mercury's perihelion is considered one of the greatest accomplishments of modern science.

Summary

To estimate the statistical power of a proposed study the researcher must be able to anticipate the magnitude of the treatment effect. In principle, this is not difficult to do. In practice, there are several reasons why researchers do not estimate effect size, and as a consequence remain unaware of the statistical power of their research. The emphasis on significance testing procedures in most graduate and undergraduate textbooks seems to have created among behavioral researchers the impression that significance level may be used as an indicator of the size and importance of a treatment effect. It is not uncommon, for instance, to see results reported as "highly significant" accompanied by a significance level of (for instance) $p < 10^{-7}$ used to impress the reader with the importance of the findings.

Another consequence of the significance testing routine is that measures of effect size have remained generally unknown to applied researchers despite the existence of a variety of such measures. Noncentrality parameters are preferred by mathematical statisticians since power functions are most conveniently expressed in such terms. Unfortunately, noncentrality parameters do not have clear operational interpretations, and although some

statisticians have argued for a greater use of noncentrality parameters as effect size indices (Bargmann, 1970; Venables, 1975), behavioral researchers have not been inclined to do so.

Proportion of variance statistics have been the most popular measures of effect size among behavioral researchers. Several different versions exist and have been described in detail in the journal literature, although textbook coverage has been less extensive. Thus, proportion of variance statistics are still infrequently included in published research reports.

Cohen (1977) has introduced new measures of effect size for several of the most common statistical techniques. Cohen sought to develop indices with very simple conceptual interpretations so as to reduce the complexity of statistical power analysis. Unfortunately, his techniques have not yet come into wide usage except as indicators of effect size in meta-analyses. Determination of statistical power by applied researchers is even less common than the reporting of measures of effect size.

Since behavioral researchers frequently do not know what effect size to expect in their research, Cohen (1977) suggested that power be computed for a range of effect

sizes. Cohen therefore introduced standardized definitions of small, medium, and large effect size. Although these definitions were determined rationally, surveys of effect sizes obtained in behavioral research have been consistent with Cohen's definitions. Unfortunately, very few such surveys have been conducted and more are necessary to obtain a reasonable assessment of the magnitude of effects characteristic of behavioral research.

Some researchers have vigorously criticized measures of effect size, particularly proportion of variance statistics. Many of these criticisms are valid and suggest that effect size measures be interpreted cautiously. In particular, researchers should be careful in using effect size as an index of the importance of a treatment effect. The interpretive limitations on effect size statistics are not widely known among behavioral researchers. If, as expected, effect size statistics become more widely used in the next few years, it will be important for methodologists to communicate more effectively with applied researchers so that the limitations of effect size measures become more widely appreciated.

CHAPTER 3

Study I

Power Survey of Current Psychological Research

Cohen's Power Survey

Given the widespread lack of interest in statistical power exhibited by behavioral scientists, it is natural to wonder what the effect of such disinterest might be on the power of behavioral research. Cohen (1962) was the first to conduct a systematic survey of the statistical power of any area of behavioral research, and his paper has been credited with introducing the concept of statistical power into the literature of the social sciences (Chase & Tucker, 1976; but for an earlier account of power see Mosteller & Bush, 1954).

Cohen (1962) surveyed all of the articles published in the Journal of Abnormal and Social Psychology for the year 1960 (volume 61), a total of 78 papers. Eliminated from consideration were articles in which no statistical tests were conducted (e.g., case reports, factor-analytic studies, etc.), reducing the total to 70 papers. Also eliminated were statistical tests considered to be of a "peripheral" nature, that is, tests which did not address the major hypotheses under investigation. Peripheral tests typically included tests of significance for all of the

correlation coefficients in a factor analysis, tests of reliability coefficients, unhypothesized interactions in analysis of variance designs, and so on. This procedure reduced the total number of statistical tests for which power was estimated from 4829 to 2088.

To determine power for each statistical test Cohen assumed nondirectional hypothesis testing at $\alpha = .05$, and recorded the sample size on which each test was based. The major difficulty for Cohen was the estimation of the effect size on which to base the power determinations. Because authors never reported the expected population effect size, Cohen elected to estimate the power of each statistical test for a range of effect sizes: small, medium, and large. The operational definitions of small, medium, and large effect sizes necessarily differed depending on the type of statistical test under consideration. Metric-free population parameters were defined for seven different statistical tests: t test, Pearson r , z test for the difference between two correlation coefficients, sign test, z test for differences between two proportions, chi square test, and the F test in the analysis of variance and covariance. Two examples are described below.

1. For a t test of the hypothesis that the means of two independently drawn samples are equal, small, medium, and large effects were defined as mean group differences of 0.25, 0.50, and 1.00 standard deviation, respectively.

2. For the hypothesis that a correlation coefficient is zero, small, medium, and large effects were defined as correlations of .20, .40, and .60, respectively.

Admittedly arbitrary, Cohen (1962) believed these definitions were nevertheless reasonable.

Cohen's (1962) procedure resulted in three power estimates for each statistical test, based on the assumption of a small, medium, or large sized effect. Power for each article was estimated by averaging across all of the major statistical tests reported in the article. The power of abnormal-social psychological research as a whole was then estimated by averaging across the results for each of the 70 articles. The average power was .18 for small effects, .48 for medium effects, and .83 for large effects. (A separate power analysis conducted on the "peripheral" tests gave similar results.) These results demonstrated that, even for medium-sized effects, the probability that a false null hypothesis could be rejected was less than one-half.

Cohen (1962) had two recommendations on how to improve

matters. First was that increasing sample size would be the most effective means of increasing the power of abnormal-social psychological research. Manipulation of any of the other factors related to power would be less likely to be effective and more likely to be controversial (e.g., increasing alpha to .10). Cohen's second recommendation was that researchers routinely determine the power of their experimental designs before collecting their data. In the event that a priori estimates of effect size were unavailable, Cohen suggested that power be estimated for the range of effect sizes used in his paper.

Results of Other Power Surveys

Since the report of Cohen's (1962) power survey, the concept of statistical power has been gradually introduced in a variety of research areas. Issues relating to power have been discussed in the fields of special education (Hopkins, 1973), behavioral genetics (Klein, 1974), psychopharmacology (Overall, Hollister, & Dalal, 1967), social work (Crane, 1976, 1977; Schuerman, 1977), consumer research (Sawyer, 1982), medical outcome research (Berk & Chalmers, 1981; Freiman, Chalmers, Smith, & Kuebler, 1978), endocrinology (Poland, Rubin, & Weichsel, 1980), epidemiology (Whaley, Quade, & Haley, 1980), even geography (Bones, 1972)!

Several other power surveys have also been conducted, encouraged by Cohen's (1969) subsequent publication of convenient power tables (see Table 5). These more recent surveys have covered research in a number of areas, including educational research (Brewer, 1972; Brewer & Owen, 1973; Daly & Hexamer, 1983; Pennick & Brewer, 1972; Woolley & Dawson, 1983), health research (Christensen & Christensen, 1977; Jones & Brewer, 1972), speech research (Chase & Tucker, 1975; Katzer & Sodt, 1973), research in counselor education (Haase, 1974), sociological research (Spreitzer & Chase, 1974), research on speech disorders (Kroll & Chase, 1975), research on mass communication (Chase & Baran, 1976), applied psychological research (Chase & Chase, 1976), research in physical anthropology (Chase, Chase, & Tucker, 1978), gerontological research (Levenson, 1980), and marketing research (Sawyer & Ball, 1981).

All of these surveys have closely followed the procedures set by Cohen (1962). All of the studies using statistics in one or more journals were surveyed for a period of time, usually one or two years. Power values for most surveys were averaged across studies, although averaging across statistical tests was used in a few surveys (Brewer, 1972; Brewer & Owen, 1973; Jones & Brewer, 1972). For all surveys power was not computed for "peripheral" tests.

Table 5

Power Surveys of Behavioral Research

Source	Journals Included in Survey	Period Covered
Cohen (1962)	Journal of Abnormal and Social Psychology	1960, vol. 61
Brewer (1972)	American Educational Research Journal	1969 - 1971
Jones & Brewer (1972)	Research Quarterly	1969 - 1971
Pennick & Brewer (1972)	Journal of Research in Science Teaching	1969 - 1970
Brewer & Owen (1973)	a) Journal of Educational Measurement	1969 - 1971
	b) Journal of Educational Psychology	not reported
Katzer & Sodt (1973)	Journal of Communication	1971 - 1972
Haase (1974)	Counselor Education and Supervision	1968 - 1971
Sprenitzer & Chase (1974)	American Sociological Review	1972 - 1973

(cont'd)

Table 5 (Cont'd)

Power Surveys of Behavioral Research

Source	Journals Included in Survey	Period Covered
Chase & Tucker (1975)	a) American Forensic Association Journal	1973
	b) Central States Speech Journal	1973
	c) Journal of Communication	1973
	d) Quarterly Journal of Speech	1973
	e) Southern Speech Journal	1973
	f) Speech Monographs	1973
	g) Speech Teacher	1973
	h) Today's Speech	1973
	i) Western Speech	1973
Kroll & Chase (1975)	a) Journal of Communication Disorders	1973 - 1974
	b) Journal of Speech and Hearing Research	1973 - 1974

(cont'd)

Table 5

Power Surveys of Behavioral Research

Source	Journals Included in Survey	Period Covered
Chase & Baran (1976)	a) Journalism Quarterly	1974
	b) Journal of Broadcasting	1974
Chase & Chase (1976)	Journal of Applied Psychology	1974, vol. 59
Christensen (1977)	Research Quarterly	1975
Chase et al. (1978)	American Journal of Physical Anthropology	1976
Levenson (1980)	a) Gerontologist	1961 - 1977
	b) Journal of Gerontology	1946 - 1977
Sawyer & Ball (1981)	Journal of Marketing Research	1979
Daly & Heramer (1983)	Research in the Teaching of English	1978 - 1980
Woolley & Dawson (1983)	Journal of Research in Science Teaching	1977 - 1980

There is one major difference in method between the power survey conducted by Cohen (1962) and those conducted by later investigators. This difference concerns the definitions of small, medium, and large effect sizes. Changes in these definitions were suggested by Cohen (1969, 1970) himself so that the magnitude of the effect implied by the various definitions would be more nearly equivalent to each other. The difference between the "1962" and "1977" definitions of effect size was discussed in detail in Chapter 2. Two examples are repeated here for convenience. For the t test the definitions were changed from 0.25, 0.50, and 1.00 standard deviation to 0.20, 0.50, and 0.80 standard deviation for small, medium, and large effect sizes, respectively. For the Pearson correlation coefficient the definitions were changed from .20, .40, and .60 to .10, .30, and .50 for small, medium, and large effect sizes, respectively.

Since these changes resulted in somewhat smaller effect size definitions, a decrease in power might have been expected for subsequent power surveys. Instead, results were generally similar to those obtained by Cohen (1962). Table 6 shows the power estimates for all of the surveys and for the combined results. With few exceptions, most surveys produced remarkably consistent results, especially

Table 6

Results of Power Surveys of Behavioral Research

Source	Sample Size		Statistical Power Estimate		
	Articles	Tests	Small Effect	Medium Effect	Large Effect
Cohen (1962)	70	2088	.18	.48	.83
Brewer (1972)	47	373	.14	.58	.78
Jones & Brewer (1972)	106	261	.15	.54	.83
Pennick & Brewer (1972)	66	554	.22	.71	.87
Brewer & Owen (1973) (a)	13	267	.21	.72	.96
Brewer & Owen (1973) (b)	NR	469	.09	.44	.71
Katzer & Sadt (1973)	31	1671	.23	.56	.79
Haase (1974) ¹	60	206	.19	.46	.72
Spreitzer & Chase (1974)	34	1049	.55	.84	.94
Chase & Tucker (1975) ²	46	1298	.18	.52	.79
Kroll & Chase (1975) ³	62	1037	.16	.44	.73
Chase & Baran (1976) ³	48	701	.34	.76	.91
Chase & Chase (1976)	121	3373	.25	.67	.86
Christensen & Christensen (1977)	43	NR	.18	.39	.62
Chase et al. (1978)	25	3304	.38	.62	.81
Levenson (1980) (a)	26	NR	.36	.85	.94
Levenson (1980) (b)	30	NR	.37	.91	.98
Sawyer & Ball (1981)	23	475	.41	.89	.98
Daly & Hexamer (1983)	57	1233	.22	.63	.86
Woolley & Dawson (1983)	192	3556	.23	.63	.85
Totals and Means ⁴	1100	21915	.23	.61	.83
Medians ⁵	47	1037	.22	.62	.84

(cont'd)

Table 6 (Cont'd)

Results of Power Surveys of Behavioral Research

Notes: Definitions for small, medium and large effect sizes are based on Cohen (1969) for all surveys except Cohen (1962). All power values are averaged across articles except for Brewer (1972), Jones and Brewer (1972), and Brewer and Owen (1973), which are averaged across tests.

NR = not reported.

¹Power estimates are based on grouped frequency data.

²Results for the nine journals in this survey are not reported because of the small sample sizes on which each was based.

³Results for the individual journals in this survey were not recorded in the published report.

⁴Totals are given for articles and tests. Means are given for power estimates, and are weighted by the number of articles.

⁵Medians are for listed table entries.

given the diverse nature of the research fields covered. The average power estimates across the 20 power surveys is .23, .61, and .83 for small, medium, and large effects, respectively. These results are based on nearly 22000 statistical tests published in 1100 journal articles.

Another way of viewing these data is given in Table 7, which shows the percentage of studies with power less than .80 and .50 for small, medium, and large effects. These results demonstrate that, given a medium-sized effect, 66% of the studies surveyed failed to achieve power of .80, while 41% obtained a power value of less than .50. If one adopts Cohen's (1965, 1977) suggestion of .80 as the minimum desirable power value, then Table 6 shows that the power of the surveyed research is inadequate to detect all but large effects. Even so, as Table 7 indicates, nearly one-third of the surveyed studies failed to reach power of .80 even for large effects.

Power Survey of Current Psychological Research

The inadequate level of statistical power indicated by the results of power surveys has created some concern among methodologists about the health of psychological research (e.g., Bakan, 1966; Cohen, 1965; Greenwald, 1975; Smart, 1964; Tversky & Kahneman, 1971), especially with respect to the possible inflation of type I error rates.

Table 7

Percentage of Studies with Power Less than .80 and .50

Source	Power < .80			Power < .50		
	Small Effect	Medium Effect	Large Effect	Small Effect	Medium Effect	Large Effect
Cohen (1962)	100	91	31	100	57	3
Brewer (1972) ¹	99	58	29	98	43	22
Jones & Brewer (1972) ¹	99	75	36	98	50	13
Pennick & Brewer (1972)	97	64	30	95	39	14
Brewer & Owen (1973) (a) ¹	NR	NR	NR	94	14	0
Katzer & Sadt (1973)	100	73	46	87	48	10
Haase (1974)	95	83	50	90	57	27
Spreitzer & Chase (1974)	71	26	9	41	9	6
Chase & Tucker (1975) ²	100	81	46	95	55	4
Kroll & Chase (1975) ²	98	85	68	96	71	13
Chase & Baran (1976) ²	92	46	10	73	15	6
Chase & Chase (1976)	93	59	26	87	29	7
Christensen & Christensen (1977)	91	88	70	88	70	33
Chase et al. (1978)	72	60	28	68	40	12
Levenson (1980) (a)	89	23	11	NR	NR	NR
Levenson (1980) (b)	90	13	0	NR	NR	NR
Sawyer & Ball (1981)	NR	78	NR	NR	4	NR
Daly & Hexamer (1983) ¹	96	67	23	90	29	8
Woolley & Dawson (1983)	96	65	26	88	35	7
Means ³	95	66	32	89	41	11
Medians ⁴	96	66	29	90	40	9

(cont'd)

Table 7 (Cont'd)

Percentage of Studies with Power Less than .80 and .50

NR = not reported.

¹Figures are for percentages of statistical tests (not articles).

²Figures are for all journals combined.

³Means are weighted by the number of articles in each survey.

⁴Medians are for listed table entries.

One surprising aspect of these surveys is that, since the publication of Cohen's survey, only one other power survey has been conducted on psychological research (see Table 5). Chase and Chase (1976) computed the average statistical power of studies published in the Journal of Applied Psychology. They found that, although power was very low for small effect sizes (.25), power for medium-sized effects was higher (.67) than that found either by Cohen (.48) or in other power surveys. Power to detect large effects was very good (.86). Furthermore, Chase and Chase (1976) did not base their power estimates on those used by Cohen (1962), but used instead the somewhat smaller "1977" definitions of effect size.

The irony of these results is that the inadequate statistical power often attributed to psychological research is based primarily on power surveys of research in other fields and on Cohen's results, which were gathered over 20 years ago. The only recent survey of power in psychological research--that of Chase and Chase (1976)--presented a relatively less gloomy picture. On the other hand, applied psychological research may not be representative of research in other areas of psychology. In particular, the sample sizes characteristic of research reported in the Journal of Applied Psychology are

substantially larger than the sample sizes of most psychological research. Muchinsky (1979) reported average sample sizes in the Journal of Applied Psychology of approximately 350. Sample sizes for most other psychological journals are only about 50 or 60 (Holmes, 1979; Holmes, Holmes, & Fanning, 1981).

The primary purpose of Study I is to provide a more current assessment of the power of psychological research than is now available. Two journals were selected for the survey: the 1982 volumes of the Journal of Personality and Social Psychology and of the Journal of Abnormal Psychology. These two journals are representative of a substantial portion of psychological research, and are quite selective, with rejection rates of 79% and 83%, respectively, for 1982 ("Summary Report," 1983). Furthermore, both are the direct "descendents" of the Journal of Abnormal and Social Psychology, the journal examined by Cohen (1962). Thus, a second purpose of this study was to provide a direct comparison to Cohen's results, to discover what change has occurred over the past 20 years in the power of abnormal and social psychological research.

Any really meaningful increase in power values is most

likely to occur at the medium effect size level. Cohen's power estimate of .48 for medium-sized effects was based on a larger definition of effect size than that used in subsequent surveys. Yet, as Table 6 shows, other surveys have reported higher power at this level of effect magnitude (averaging about .60). Thus, an increase in power here would not be surprising when Cohen's original effect size definitions are used to calculate power. However, it is probably optimistic to suppose that such an increase will be so great as to approach .80, a frequently cited criterion of adequate statistical power. Furthermore, it will be interesting to observe whether power for medium-sized effects increases at all when the more recent (and smaller) effect size definitions are used to calculate power.

Power for large and small effects should be less affected either by time or by changes in effect size definitions. Power for large effect sizes is already reasonably high, and would require substantial increases in sample size to effect an increase in power. Surveys of sample sizes used in psychological research have shown no such trends (Holmes, 1979; Holmes, Holmes, & Fanning, 1981).

Power for small effect sizes is so low that only heroic increases in sample size could result in satisfactory power

values. Even a large increase in power for small effects, while noteworthy academically, will still result in only humble odds for detecting such effects. For all practical purposes, the detection of small effects in psychological research will remain problematic for many years to come. Efforts at detecting such effects, if they are to be successful, may have to achieve greater statistical power by concentrating less on increasing sample size and by directing more resources towards increasing the magnitude of the effect, such as by employing more sensitive measurement models, and more effective experimental manipulations.

Method

Selection Procedures: Articles

All of the articles published in the Journal of Personality and Social Psychology, 1982, volume 42, and the Journal of Abnormal Psychology, 1982, volume 91, were examined and articles not reporting any statistical tests were eliminated from the study. In addition, some articles were excluded because they contained only statistical methods for which power could not be determined. The selection procedures for statistical tests are described below.

Selection Procedures: Statistical Tests

As in previous surveys, a distinction was made between "major" and "peripheral" statistical tests. Major tests are those which bear directly on the research hypotheses of the study, while peripheral tests do not. Peripheral tests were not included in the survey. Peripheral tests included all of the correlation coefficients of a factor or principal component analysis, unhypothesized higher order interactions in the analysis of variance, manipulation checks, inter-rater reliability coefficients, reliabilities of psychometric tests (internal consistency, test-retest), post hoc analysis of variance procedures (simple effects, multiple comparisons), and tests of statistical assumptions. In general, all statistical tests reported within the Method section of an article were excluded as well.

For "major" statistical tests, power determinations were limited to the those listed in Table 8. These tests were selected for several reasons. Perhaps most important, only these tests (except for multiple regression) were used in Cohen (1962). Cohen's choices were based on two factors: the frequency of a statistic's occurrence, and the availability of equations for power calculations for a statistic. Subsequent power surveys have also limited

consideration to just these statistical tests. Here, the primary limiting factor was that Cohen's (1969) power handbook contained tables only for these statistics, and it is clear that these surveys would never have been conducted had this source not been available.

Table 8

Statistical Tests Included in the Power Survey

-
1. Student's t test
 2. Pearson r
 3. z test for differences between correlation coefficients
 4. test that a proportion is .50 (sign test)
 5. z test for differences between proportions
 6. chi-square test
 7. F tests on means in the analysis of variance and covariance
 8. F tests in multiple regression/correlation analysis
-

As indicated in Table 8, multiple regression was included in this power survey, even though it was not included in Cohen's (1962) original survey. Multiple regression has also not been included in most previous power surveys, probably because Cohen's (1969) first power handbook did not include power tables for multiple regression. It should also be pointed out that in most textbook discussions of multiple regression the issue of

statistical power is treated in a most cursory fashion, if at all (e.g., Edwards, 1979; Freund & Minton, 1979; Gunst & Mason, 1980; Harris, 1975; Johnson & Wichern, 1982; Lindeman et al., 1980; Lunneborg & Abbott, 1983; Marascuilo & Levin, 1983; Pedhazur, 1982).

Cohen (1962) could ignore multiple regression because the technique was rarely used at the time (Edgington, 1964). Later researchers should have realized that, since multiple regression and the analysis of variance are identical analytical systems (Cohen, 1968; Cohen & Cohen, 1983), tables for the latter could have been used to determine power for most analyses structured as multiple regressions. Furthermore, Cohen's (1977) revised power handbook does include tables for the power of F tests in multiple regression. Since the use of this technique is no longer rare (Moore, 1981), it was decided to include multiple regression in the present power survey.

Primarily excluded from the survey are most nonparametric techniques (e.g., Mann-Whitney U , Wilcoxin, Kruskal-Wallis, rank order correlation tests, etc.), multivariate methods (e.g., discriminant analysis, canonical correlation, multivariate analysis of variance), and various other methods for which the concept of power is not relevant (e.g., factor analysis, principal component

analysis, cluster analysis, time-series analysis, multidimensional scaling, log-linear methods, causal modeling, etc.).

Surveys of the use of statistics in psychological research show that, at the time of Cohen's (1962) study, nonparametric techniques were much more popular than they are now, while the use of multivariate techniques has increased substantially in recent years (Edgington, 1964, 1974; Moore, 1975, 1981). Because the power characteristics of many nonparametric statistics were unavailable, Cohen (1962) substituted power estimates for the analogous parametric tests (e.g., t test for the Mann-Whitney U test). This procedure was not continued in the present survey for two reasons: (a) None of the power surveys conducted after Cohen (1962) have continued the practice; and (b) the infrequent use of most nonparametric statistics makes impractical the estimation of power even for those tests for which power characteristics have been investigated (e.g., Bennett & Underwood, 1970; Blair & Higgins, 1980a, 1980b; Blair, Higgins, & Smitley, 1980). Published research reports do not provide convenient computing algorithms (i.e., algorithms not requiring integrations) for most nonparametric methods. Furthermore,

those sources which provide tables summarizing power functions for nonparametric methods are usually not extensive enough to meet the needs of a power survey.

A similar situation exists for most multivariate statistical methods. In general, the power characteristics of multivariate methods are well-known, since these statistics are derived from the general linear model (e.g., Das Gupta & Perlman, 1974; Koichi, 1962; Kres, 1975/1983; Ramsey, 1982; Stevens, 1980). However, for the most part, neither convenient computer algorithms nor suitable power tables are available for most multivariate techniques. Furthermore, measures of multivariate effect size have received relatively little attention (Cramer & Nicewander, 1979; Huberty, 1983; Sachdeva, 1983; Serlin, 1982; Shaffer & Gillo, 1974; Stevens, 1980). Thus, a researcher wishing to compute power for a multivariate statistic must be able to estimate the expected variance-covariance matrix. Most behavioral researchers are not in a position to be able to do this.

Determination of Statistical Power

Cohen's (1977) tables were used to determine power for the following statistical tests: differences between correlation coefficients, sign tests, differences between proportions, and chi-square tests. The tables require

entries for sample size and effect size to retrieve power values. The tabled values are given to two decimal places and are accurate to about one digit in the last decimal place when compared to exact values. Two-tailed testing at $\alpha = .05$ was assumed.

Computer programs were written to expedite the determination of power for the remaining tests (t , r and F). This was done because of the frequency of occurrence of these statistics, and to avoid the interpolation errors inevitably encountered with the use of tables. All programs are interactive and were written in BASIC to run on an Apple II Plus microcomputer.

The computer program to determine the power of a t test was based on the normal approximation to the noncentral t distribution given by Cohen (1977). Since this formula assumes that sample sizes are equal, it was modified slightly to permit unequal n power calculations. The program requires only the input of sample size to produce power estimates for the range of effect sizes to be studied (see section below) and for two-tailed testing at $\alpha = .05$.

The determination of power for the Pearson correlation coefficient was based on the normal score approximation for

$\bar{r}$ provided by the hyperbolic arctangent transformation, plus a correction factor for small sample sizes (Cohen, 1977). This program also requires only the input of sample size to produce power estimates for the range of effect sizes to be studied and for two-tailed testing at $\alpha = .05$.

The cube root normal approximation of the noncentral F distribution was employed for the analysis of variance power program (Laubscher, 1960; Severo & Zelen, 1960). The accuracy of this formula has been found to be quite good, with errors appearing only in the 3rd or 4th decimal places for $\alpha = .05$ (Pearson, 1960; Tikku, 1966). Although more accurate approximations exist (Barton, David, & O'Neill, 1960; Han & Wang, 1983; Laubscher, 1960; Patnaik, 1949; Tikku, 1965, 1966), their additional computational complexity does not justify the gain in resolution. This program requires inputs for numerator and denominator degrees of freedom and for the associated critical (5%) F value to produce power estimates for the range of effect sizes of interest.

For all three computer programs, the normal score approximation is converted to a probability (power) value using formula 26.2.19 in Abramowitz and Stegun (1965, p. 932). The values produced by this algorithm agree with exact values given by Hawkes (1982) to five significant

figures (as many as given by Hawkes) for $z \leq 3$. Abramowitz and Stegun (1965) give the accuracy of the algorithm as $\pm 1.5 \times 10^{-7}$. This level of accuracy is more than sufficient for present purposes. The resulting power values agree well with those given in Cohen's (1977) power tables.

Range and Type of Effect Size Indices

Power determinations were made for two sets of effect size range. One set of small, medium, and large effect size definitions was based on those used by Cohen (1962) for comparison with the power results of that survey. The other set was based on the effect sizes given in Cohen (1969, 1977), for comparison with the results of power surveys conducted after Cohen (1962). The indices of effect size and magnitudes of their ranges are given in the following sections for each of the statistics listed in Table 8.

1. t test. The effect size index for student's t test is Cohen's d, the standardized difference between group means (Cohen, 1962, 1969, 1977):

$$\underline{d} = (M1 - M2) / \underline{s} , \quad (3)$$

where M_1 is the mean of the first group, M_2 is the mean of the second group, and $\underline{s}$ is the common (pooled) standard deviation. $\underline{d}$ is related to delta, the noncentrality parameter for the noncentral t distribution, as follows:

$$\text{delta} = \underline{d} * \text{sqr}(n/2) , \quad (4)$$

where $\underline{n}$ is the sample size of each group.

The definitions for small, medium and large effect sizes are (a) 0.25, 0.50, and 1.00, respectively, for 1962; and (b) 0.20, 0.50, and 0.80, respectively, for 1969.

2. Pearson r . The effect size index here is simply the correlation coefficient itself. The definitions for small, medium, and large effect sizes are (a) .20, .40, and .60, respectively, for 1962; and (b) .10, .30, and .50, respectively, for 1969.

3. Differences between correlation coefficients.

Cohen's (1962) original index was simply the difference between the two correlation coefficients. Because different correlations have different sampling variances, the detectability of equal differences is not constant throughout the range, but increases as $\underline{r}$ approaches ± 1 . Thus, for a given sample size and alpha level, it is easier to detect the difference between r 's of .6 and .8 than

between r 's of .2 and .4. Another disadvantage of this measure, in the context of a power survey, is that it requires the extraction of sample values to estimate the population correlation of the test, thus rendering the power of the test dependent on the results of the study. Cohen's (1962) procedure was to average the reported sample correlations, and to use this average as the population estimate in the power computations. Small, medium, and large effect sizes were then defined as correlation differences from this average of .10, .20, and .30, respectively.

Cohen (1969) offered a more convenient effect size indicator, g , based on the Fisher r -to- z transformation:

$$g = | z(1) - z(2) | , \quad (5)$$

where $z(1)$ and $z(2)$ are the z score equivalents of the two correlation coefficients. The z score transformation of r is given by

$$z = \ln((1 + r)/(1 - r))/2 , \quad (6)$$

or equivalently by

$$z = \operatorname{arctanh}(r) . \quad (7)$$

Small, medium, and large values of g are .10, .30, and .50, respectively. These values will be used not only for the "1977" power survey but for the "1962" survey comparison as well, because of the difficulties noted previously concerning Cohen's (1962) original effect size procedure for differences between correlation coefficients. Since the frequency of occurrence of this statistic was very low (less than 1%), the departure from Cohen's (1962) earlier procedure should not affect the outcome of the 1962 survey comparison.

4. Sign test. Here the effect size is simply defined as the departure of a proportion from .50:

$$g = | p - .50 | , \quad (8)$$

where p is the observed proportion. Small, medium, and large effect sizes were defined by Cohen (1962) as .10, .20, and .30, respectively, and by Cohen (1977) as .05, .15, and .25, respectively.

5. Differences between proportions. A problem similar to that for the test for differences between correlations confronted Cohen (1962) here. His solution was similar:

effect size was defined as the simple difference between proportions. Small, medium, and large differences were defined as .10, .20, and .30, respectively. Again, this solution is not very satisfactory, since different proportions have different sampling distributions. Constant differences between proportions are not equally detectable across the scale of proportions, but are easier to detect as the departure from .50 increases. Thus, the difference between proportions of .20 and .30 is easier to detect than the difference between .40 and .50. Although the magnitude of the differences are the same, the former difference is "larger" with respect to the statistical power function. As before, Cohen's (1962) definition is dependent on the observed sample proportions, which were averaged to estimate the level of the population proportion.

Eisenhart (1947) suggested the arcsin transformation to stabilize the variance and normalize the distribution of proportions:

$$\phi = 2 * \arcsin(\sqrt{p}) , \quad (9)$$

where p is the observed proportion. Cohen (1969) employed

this transformation to define an effect size index for differences between proportions. Since power is invariant over levels for differences in arcsins, a convenient index of effect size is

$$h = | \phi(1) - \phi(2) | , \quad (10)$$

where $\phi(1)$ and $\phi(2)$ are the arcsin transformations for the two proportions. While this index is in common use (Feigl, 1978; Overall & Hornick, 1982; Ury, 1982), an equally common approach is to assume asymptotic normality for the untransformed binomial proportions (Dixon & Massey, 1983; Fleiss, 1981; Fleiss, Tytun, & Ury, 1980; Runyon, 1969; Ury & Fleiss, 1980). Both approaches work well in power approximations for a wide range of conditions (Hornick & Overall, 1980; Ury, 1981), excepting small sample sizes or proportions near zero or one (Haseman, 1978).

The primary advantage of the arcsin approach for a power survey is that only the order of magnitude of the difference between proportions need be specified, and not the individual magnitudes. Cohen (1977) defined small, medium, and large effect sizes as differences in arcsins of .20, .50, and .80, respectively. These definitions will be

used for the 1962 power estimate comparisons as well as for the 1977 comparisons for the same reasons discussed for the z score test for differences between correlation coefficients.

6. Chi-square tests. The chi-square test for k proportions (goodness-of-fit) and for $r \times c$ contingency tables presented the greatest problems for Cohen (1962). For the goodness-of-fit test, the effect size was defined as the ratio of the largest proportion to the smallest proportion, with the assumption that the remaining $k - 2$ proportions were distributed proportionately throughout the range. Ratios for small, medium, and large effect sizes were 3:2, 2:1, and 4:1, respectively.

The definition of effect size, 1 , was given by

$$1 = \sum \{ (P(1i) - P(0i))^2 / P(0i) \}, \quad (11)$$

where $P(0i)$ is the proportion in cell i expected under the null hypothesis, and $P(1i)$ is the proportion in cell i expected under the alternative hypothesis. The sum is taken over all cells. Small, medium, and large effect sizes were based on the ratio of largest to smallest proportions, and assumed a proportional distribution throughout the range for the remaining cell proportions.

Thus, specific values of $\underline{l}$ were not defined as small, medium, or large by Cohen (1962); these values varied according to the size of the contingency table, decreasing with increasing table degrees of freedom. As with the $\underline{z}$ tests for differences in correlations and in proportions, Cohen's procedure for chi-square produces effect sizes which are dependent on the nature of the experiment, are tedious to compute (and of course must be computed for each study), and most important, are not invariant over studies. For these reasons, Cohen's (1962) procedures for the chi-square test were not employed in this study.

Cohen (1969) did not change his index of effect size, except to call it e instead of l . He did, however, provide definitions for small, medium, and large effect sizes: .05, .10, and .20, respectively. Cohen (1977) changed the effect size index again, proposing the statistic w , where

$$\underline{w} = \text{sqr}(e) = \text{sqr}(l) . \quad (12)$$

$\underline{w}$ is related to λ , the noncentrality parameter of the noncentral chi-square distribution:

$$\lambda = w^2 * \underline{N} , \quad (13)$$

where N is the total sample size.

Definitions of small, medium, and large effect size also changed, to .10, .30, and .50, respectively. Note that these values do not correspond exactly to the square roots of the 1969 definitions.

Because of their convenience, Cohen's (1977) effect size index and magnitude definitions will be used for both the 1962 and the 1977 power survey comparisons. The departure from the procedures of his earlier survey should not greatly affect the results, since the frequency of chi-square tests in the survey was low (about 5%).

7. F tests in the analysis of variance and covariance.

The effect size index here is f , the standard deviation of the k standardized population means:

$$\underline{f} = s(m) / \underline{s} , \quad (14)$$

where $\underline{s}$ is the common (pooled) standard deviation of the k groups, and $s(m)$ is the standard deviation of the means. That is,

$$s(m) = \text{sgt}(\text{sum}((m(i) - m)^2)/k) , \quad (15)$$

where $\bar{m}(i)$ is the mean of the i th group, $\bar{m}$ is the grand mean, and k is the number of groups. For the two-group case, $\underline{f}$ is related to d , the effect size index for the t test, by

$$\underline{f} = \underline{d} / 2 . \quad (16)$$

The index $\underline{f}$ is also closely related to ϕ , the noncentrality parameter of the noncentral F distribution introduced by Tang (1938):

$$\phi = \underline{f} * \text{sqr}(n) , \quad (17)$$

where $\underline{n}$ is the sample size of each of the $\underline{k}$ groups. Furthermore, $\underline{f}$ is related to λ , the noncentrality parameter used by Patnaik (1949) and others (Laubscher, 1960; Severo & Zelen, 1960), as follows:

$$\lambda = f^2 * \underline{n} * \underline{k} . \quad (18)$$

Small, medium, and large effect sizes were defined by Cohen (1962) as f 's of .125, .25, and .50, respectively. These definitions were modified by Cohen (1977) to .10, .25, and .40, respectively.

8. F tests in multiple regression/correlation analysis.

Cohen (1962, 1969) did not specifically deal with multiple regression analysis, but he did introduce the topic in his revised handbook (Cohen, 1977). For the most basic case of one variable predicted from a set of variables, Cohen (1977) suggested f^2 as a measure of effect size:

$$f^2 = R^2 / (1 - R^2) , \quad (19)$$

where R^2 is the squared multiple correlation coefficient, the proportion of variance in the dependent variable accounted for by the set of independent predictor variables. The index f^2 is the square of the effect size index f used in the analysis of variance and covariance. It is related to λ , the noncentrality parameter for the noncentral F distribution, as follows:

$$\lambda = f^2 * \underline{v} , \quad (20)$$

where

$$\underline{v} = \underline{N} - \underline{k} - 1 . \quad (21)$$

Here $\underline{v}$ is the error degrees of freedom, $\underline{N}$ is the total

sample size, and k is the number of groups.

Small, medium, and large effect sizes, in terms of f^2 , are defined as .02, .15, and .35, respectively. In terms of R^2 these values are equal to .02, .13, and .26, respectively (cf. Table 4).

Results

A total of 166 articles were examined, 58 in the Journal of Abnormal Psychology (JAP), and 108 in the Journal of Personality and Social Psychology (JPSP). Twenty-four articles were excluded, 10 in JAP and 14 in JPSP. Thirteen articles contained no statistics at all, and 11 contained only statistics for which power was not determined (primarily data reduction techniques, such as principal component analysis and causal modeling procedures).

Power was determined for statistics reported in the remaining 142 articles: 48 in JAP and 94 in JPSP. The total number of statistical tests for which power was calculated was 3452, 1135 in JAP and 2317 in JPSP. The frequency of occurrence for each test is given in Table 9. The sample is dominated by the "traditional" statistical tests: Pearson r , analysis of variance and covariance, and the t test, which constitute 90% of the reported

techniques. Because this tabulation necessarily excluded tests for which power was not determined, the listed frequencies can not fairly be compared to the results of surveys of the use of statistics in psychological research (Edgington, 1964, 1974; Hoyer, Raskind, & Abrahams, 1984; Moore, 1975, 1981).

Table 9

Frequency Distribution for Tests in Power Survey

Statistical Test	Frequency	Proportion	Cum. Prop.
Pearson χ^2	1420	.411	.411
ANOVA/ANCOVA	1083	.314	.725
t test	610	.177	.902
Chi-square	173	.050	.952
Multiple Regression	127	.037	.989
Difference between r's	21	.006	.995
Difference between p's	14	.004	.999
Sign test	4	.001	1.000

Since the number of statistical tests appearing in each study varied greatly--from 1 to 152--the study was used as the unit of analysis. Thus, each article contributed equally to the overall power assessment of the journal literature (following Cohen, 1962). The power of each study was determined by averaging across statistical tests.

Six separate power estimates were made for each statistical test. Three estimates were based on Cohen's

(1962) original definitions of small, medium, and large effect sizes. These estimates were designated Sml62, Med62, and Lrg62, respectively, for convenience. A second set of three estimates was based on Cohen's (1977) "modern" definitions of effect size. These estimates were designated Sml77, Med77, and Lrg77.

Essentially, then, two separate power surveys were conducted, one based on Cohen's original definitions of effect size, and one based on his more recent definitions. The results of these two surveys will be presented separately. Results were collapsed across journals, since the differences in average power estimates between journals was never greater than ± 0.019 . Averaging power estimates across experiments within studies, and averaging simply across all statistical tests (and ignoring studies) both also produced results very close to those which are presented below.

Power Results Based on "1962" Definitions

Power estimates based on Cohen's (1962) original effect size definitions are shown in Table 10. For comparison, the results of Cohen's (1962) survey are also displayed.

Power has increased for all levels of effect size. All increases were statistically significant: for small

effects, $t(210) = 2.78$, $p < .01$, $\omega^2 = .031$; for medium-sized effects, $t(210) = 3.18$, $p < .01$, $\omega^2 = .041$; for large effects, $t(210) = 3.04$, $p < .01$, $\omega^2 = .037$.

Table 11 shows the percentage of studies with power less than .50 and less than .80 for both the 1982 studies and for Cohen's (1962) results.

Power Results Based on "1977" Definitions

Power estimates based on Cohen's (1977) revised effect size definitions are shown in Table 12. For comparison, the results of other recent power surveys (including all surveys listed in Tables 5, 6 and 7, except for Cohen, 1962) are also displayed. Table 13 shows the percentage of studies with power less than .50 and less than .80 for both the 1982 studies and for the other recent power surveys.

Discussion

As shown in Table 10, power has increased since Cohen (1962) conducted his survey 20 years ago. Although these increases were statistically significant, the magnitude of the changes were modest (ω^2 's $< .05$). Power for small effects remains very poor (.24), while power for large effects--already very good (.83) in 1962--is now even better (.89).

Of greatest interest is the increase in power for

medium-sized effects. Here power has increased about 20%, from .48 to .59. Unfortunately, this increase--though an encouraging sign--still results in less than a 60% chance of detecting a medium-sized effect. At this rate of increase, power for medium-sized effects will not reach .80 for about 30 years!

Another way of comparing these results to those obtained by Cohen (1962) is shown in Table 11. Here the percentage of studies with power less than .50 and less than .80 are shown as a function of effect size. Again, the most noteworthy changes are for medium-sized effects. Cohen reported that 57% of all studies obtained power of less than .50; this figure has now dropped to 38%.

Of course, power of .50 should be generally unacceptable to most researchers and in most research situations. Cohen (1977) has recommended that researchers design their studies to achieve power of .80. At this level of power, the current design of research in JAP and JPSP is inadequate to detect all but large effects. Even so, 21% of all studies were unable to attain this level of power even for large effects. While there has been some improvement here for medium-sized effects, 74% of the studies surveyed failed to reach power of .80 for this

Table 10

Average Power of 1982 Studies Compared to Cohen's (1962)Results: "1962" Effect Size Definitions

Effect Size	Mean	SD	Median

1982 Studies (n = 142)			
Sml62	.240	.171	.184
Med62	.590	.251	.582
Lrg62	.893	.131	.958
Cohen's Results (n = 70)			
Sml62	.18	.08	.17
Med62	.48	.20	.46
Lrg62	.83	.16	.89

Table 11

Percentage of Studies with Power < .50 and < .80

Effect Size	1982 Studies		Cohen (1962)	
	< .50	< .80	< .50	< .80
Sml62	92	97	100	100
Med62	38	74	57	91
Lrg62	1	21	3	31

Table 12

Average Power of 1982 Studies Compared to Other PowerSurvey Results: "1977" Effect Size Definitions

Effect Size	Mean	SD	Median

1982 Studies (n = 142)			
Sml77	.153	.117	.121
Med77	.549	.242	.516
Lrg77	.814	.186	.883
Other Surveys (n = 1030)			
Sml77	.238	-	.22
Med77	.623	-	.63
Lrg77	.834	-	.85

Table 13

Percentage of Studies with Power < .50 and < .80

Effect Size	1982 Studies		Other Surveys	
	< .50	< .80	< .50	< .80
Sml77	98	99	88	94
Med77	46	84	40	65
Lrg77	8	36	12	32

level of effect size.

Power estimates based on Cohen's (1977) more recent--and smaller--effect size definitions are naturally somewhat lower than power estimates based on Cohen's (1962) original effect size definitions (see Tables 12 and 13). These results are generally consistent with the range of results obtained by other power surveys (shown in Tables 6 and 7). In fact, average power for studies published in JAP and JPSP was on the average somewhat lower than for studies published in other behavioral journals. The common tendency to generalize to psychological research the inadequate power obtained by studies in other areas of behavioral research therefore seems to have been justified.

The results of these power surveys are not encouraging, whether power is based on the "1962" effect size definitions or on the "1977" definitions. While power may be adequate if a large effect size is postulated, there is often no way to judge whether or not such an assumption is justified. Furthermore, surveys suggest that the average effect size in psychological research may be no larger than Cohen's definition of a medium-sized effect (Brown, 1975; Haase, Waechter, & Solomon, 1982), for which average power was less than .60.

It might be argued that the power of behavioral

research must be adequate, since a large proportion of all studies published in behavioral research journals have reported statistically significant results (Greenwald, 1975; Smart, 1964; Sterling, 1959). Unfortunately, not only does low power suggest that there may be a large number of type II errors in the behavioral research literature, low power also suggests the possibility of a proliferation of type I errors in the research literature.

Consider a researcher who conducts an investigation for which the null hypothesis is true. With effect size equal to zero, power will be equal to alpha, since there is no true alternative to detect. Should our unfortunate researcher (or more likely, other researchers) persist in conducting research in this area, in the long run significant results will occur at a rate of 5%, if $\alpha = .05$. Many of these significant results are likely to be published, however, while the non-significant results will for the most part not be published, but rather will remain in "file drawers" (Rosenthal, 1979), an indication of the bias against null results exhibited by journal editors (Bakan, 1966; Carver, 1978; Greenwald, 1975; Lykken, 1968; McNemar, 1960; Smart, 1964; Sterling, 1959; Tullock, 1959), reviewers (Atkinson, Furlong, & Wampold, 1982; Fagley &

McKinney, 1983; Wampold, Furlong, & Atkinson, 1983), and the research community in general (Beauchamp & May, 1964; Rosenthal & Gaito, 1963, 1964; see also the related papers by Beyth-Marom, 1982, Cowles & Davis, 1982a, Lichtenstein & Newman, 1967, Simpson, 1944, 1963, and Strahan & McGovern, 1979). Thus, 100% of all published significant results will be type I errors, despite a type I error rate of .05! In this way, low power undermines the confidence which can be placed even in statistically significant results.

While this situation is admittedly contrived, the situation with power greater than alpha but still low is not much better. When the power of research is low, the probability of rejecting a true null hypothesis may be only slightly smaller than the probability of rejecting the null hypothesis when the alternative is true. That is, the ratio of type I errors to power may be uncomfortably large, indicating that a large proportion of all significant results may be due to false rejections of valid null hypotheses. In the extreme this leads to the degenerate situation described above. Nor is this situation unknown to methodologists, many of whom have commented upon various aspects of the problem (Bakan, 1966; Cohen, 1965; Katzer & Sodt, 1973; Kuncze, Cook, & Miller, 1975; Lane & Dunlap, 1978; McNemar, 1960; Millman & Cieslak, 1970; Overall, 1969;

Rosenthal, 1966, 1979, 1980b; Selvin & Stuart, 1966; Sterling, 1959; Tullock, 1959; Walster & Cleary, 1970), some going so far as to suggest that all published significant results may be type I errors! This may well be the legacy of low statistical power.

CHAPTER 4

Study II

Statistical Power and Artifactual Controversy
in Research on Spontaneous RecoveryLimitations of Power Surveys

Most new techniques, and most new applications of existing techniques, receive their fair share of critical commentary. Surveys of statistical power have fared well in this respect, since most comments have been favorable. The only exception has been the power survey conducted by Brewer (1972), which has been extensively criticized (Cohen, 1973b; Dayton, Schafer, & Rogers, 1973; Meyer, 1974a, 1974b).

Brewer's (1972) power survey included a number of irregularities not repeated in other power surveys, which have stayed closer to the format introduced by Cohen (1962). Brewer erred in including only statistically significant results in his survey (and specifically excluding nonsignificant results), in substituting p for α in the determination of power, and most seriously in interpreting power as the probability of a valid rejection of the null hypothesis after a statistical test has been conducted. (Brewer, 1974, claims to have been misinterpreted.)

Although not often criticized, power surveys do have inherent limitations. Since reporting of effect sizes is not a common publication practice, power estimates have to be based on a range of effect sizes. The range is supposed to be representative of the magnitude of effects in behavioral research. But extensive surveys of effect size magnitude have not been conducted, so that there is no way to assess whether or not the range of effect sizes used in power surveys does, in fact, characterize behavioral research. However, relatively limited, small-scale assessments of obtained effect size generally do support Cohen's estimates, as previously mentioned (Brown, 1975; Haase, Waechter, & Solomon, 1982).

A second limiting aspect of power surveys is that all have concentrated on entire research areas, usually by analyzing all of the studies appearing in a specific journal (see Table 5). In Study I, for instance, research published in the Journal of Personality and Social Psychology and in the Journal of Abnormal Psychology was examined. The type of research published in such journals, and indeed in most journals, necessarily covers quite a broad range of topics. Thus, even if it is admitted that the range of effect sizes on which power analyses are based

is generally accurate, the results of the power analyses are nevertheless strictly applicable only to an entire area of research. The most that can be said of the results of Study I, then, is that the power of social and abnormal psychological research in general is low.

There is certainly nothing wrong with such a statement. Power surveys, as they have been conducted, have made an important (if largely ignored) contribution to the methodology of behavioral research. Yet the limitations on the interpretation of power surveys leave unanswered some interesting questions. What exactly does it mean to say that the statistical power of some very broadly defined area of research is inadequate? After all, knowing that a high type I error rate plagues a particular journal literature does not reveal which studies are type I errors and which are not! Furthermore, the impact of such knowledge is diffused across research in many specific areas, ultimately leading to the speculation that the accumulated impact of low statistical power on behavioral research may be similarly diffuse.

More interesting might be a consideration of the statistical power of a specific area of research, that is, an area for which a specific research hypothesis has been clearly defined and operationalized. What would be the

consequences of inadequate statistical power in a specific research area?

One important consequence might be the generation of a controversy concerning the existence of the phenomenon or treatment under study. This would be particularly likely to occur if, for the set of studies designed to detect the effect, the average power were approximately .50. The statistical power of such research would be inadequate to insure the detection of the effect in all studies, and yet sufficient to guarantee some significant results.

Consider the following example. A researcher discovers that a one-hour training program increases scores on the Verbal section of the Scholastic Aptitude Test (SAT-V) by 50 points ($SD = 100$). Since this is a rather startling result, several other investigators wish to confirm the effect. A total of 10 exact replications of the original study are conducted. In each replication, 50 subjects are given training, while an additional 50 subjects are used as controls. The test for statistical significance for the differences in SAT-V scores between the treatment group after training and the control group is conducted at the .01 level. Finally, assume also that the training program really does work in the manner suggested in the original

study.

What is the probability that any one of these 10 studies will yield a statistically significant result? Since we are assuming that the training program does increase SAT-V scores, one might expect all 10 studies to result in statistically significant differences. This would be true if samples perfectly representative of the two populations could be drawn. In fact, sampling error will reduce the likelihood of obtaining a statistically significant result in any particular study.

For the case described here, the resulting statistical power of a one-tailed t test at $\alpha = .01$ is .55, that is, with a total of 100 subjects, the probability of a significant test result in any one study is only .55. Thus, if all ten studies were conducted, it might be expected that only five or six of the studies would yield significant results, even assuming that the training program works to the extent described. Whether or not one obtained a significant result would be dependent on sampling error. It is easy to understand how controversy over the effectiveness of the training program might develop under these circumstances.

Although the situation described in the example is, of course, contrived, it does not seem implausible, especially

given the low power of behavioral research indicated by power surveys. Whether or not similar situations have actually arisen during the conduct of behavioral research is the question Study II was designed to answer.

The substantive topic to be considered is the "spontaneous" recovery of previously extinguished or "unlearned" verbal associations. Spontaneous recovery occupies a critical position in the interference theory of memory (Underwood & Postman, 1960), yet the issue of its existence has long "enjoyed" controversial status. The traditional approach to the study of spontaneous recovery is the standard (A-B, A-C) retroactive interference design. Within this design, spontaneous recovery is demonstrated by the significant interaction of time (retention intervals) and groups (interference vs. control group).

Analysis is accomplished through the straightforward application of a two-factor fixed-effects between-groups analysis of variance. Unfortunately, the critical interaction of groups by time has not always yielded significant results. During the period of its most intensive study (1948-1969), approximately 40 studies were published on spontaneous recovery, with roughly equal numbers of significant and nonsignificant results. The

ambiguity in the analysis of variance results has led to the conclusion in most texts and reviews that the evidence for spontaneous recovery is not convincing (e.g., Baddeley, 1976; Crowder, 1976; Keppel, 1968; Klatzky, 1980; Postman, Stark, & Fraser, 1968; Postman & Underwood, 1973). Some researchers have blamed the analysis of variance itself, and have suggested alternative analytical procedures (Brown, 1976; Coleman, 1979). The effect is often considered ephemeral, and the issue has not been resolved so much as it has been abandoned.

Lively debate on a controversial issue is often regarded as a healthy sign; in many areas of scientific research, conflicting findings generating such debate often precede major theoretical advance (Kuhn, 1970). Unfortunately, this may not always be the case in the behavioral sciences, where dependence on statistical tests to establish the existence of an effect renders the research vulnerable to the statistical power of those tests.

Study II examined the statistical power of the research which has been conducted on spontaneous recovery. Such an analysis requires the specification of the magnitude of the spontaneous recovery effect. One approach to this problem would be to estimate power for a range of effect sizes, as

in Study I and in previous power surveys. Although this was done, a more appropriate procedure is possible, one which has not yet been employed in connection with power analysis. The magnitude of the spontaneous recovery effect was estimated for each study and an average effect size was calculated. This average was then used as the effect size estimate in the power analysis. Procedures for the cumulation of effect sizes across studies--collectively known as meta-analysis--are described in the Method section.

Although this approach is quite simple, previous researchers seem not to have considered it. The most likely reason for this failure is the emphasis of previous power studies on the survey of research results, the big picture, as it were. The advantages and disadvantages of this approach have already been discussed. The scope of the research topics covered in a power survey is necessarily broad and it may well be meaningless to speak of an "average" effect size, since averages characterize heterogeneous populations poorly.

There are difficulties with the approach that will be attempted here. As with previous power surveys, the chief technical difficulty lies in the estimation of effect size.

Although many writers have called for the increased use of effect size estimates in published research reports (Bakan, 1966; Carver, 1978; Cohen, 1965, 1977; Cowles, 1977; Craig, Eison, & Metze, 1976; Dwyer, 1974; Hays, 1963, 1981; Vaughan & Corballis, 1969), the inclusion of such estimates in journal articles is still not a common practice. As the resolution of this problem is also primarily technical, its discussion will be deferred to the Method Section.

At least two additional difficulties need to be discussed. First, it might well be argued that studies on spontaneous recovery do not constitute a homogeneous set, and therefore an average effect size should not be computed. Strictly speaking, this point is correct, since the set of studies are clearly not exact replications. However, this criticism is more important from a theoretical than a methodological standpoint. For example, one obvious design feature which might affect effect size is the length of the retention interval. If there is a relationship between effect size and retention interval, then it would be clear that spontaneous recovery studies could not be considered to yield a homogeneous set of effect sizes. Such a relationship would have important theoretical implications concerning the nature of spontaneous recovery.

From the standpoint of statistical power, however, any relationship between effect size and retention interval, or any other design characteristic of spontaneous recovery research, becomes relatively unimportant. Ordinarily, even this assertion could be challenged, but for the fact that the existence of spontaneous recovery has been questioned in general. That is, the controversy surrounding spontaneous recovery is not whether it can be demonstrated at one retention interval and not another, but whether it can be demonstrated at all. Thus, it makes sense to estimate the power of spontaneous recovery research in toto, since it is the adjudged failure of spontaneous recovery research as an entire research domain which has led researchers to question the existence of the phenomenon.

The final difficulty is perhaps less subtle than the last and concerns the problem of publication bias, that is, the tendency of significant results to be published while nonsignificant results are not (Atkinson, Furlong, & Wampold, 1982; Bakan, 1966; Smart, 1964; Sterling, 1959). Cohen (1962) has pointed out that, poor as the results of his survey were, the actual statistical power of research must be even worse, since publication bias insures that

only the most powerful research (i.e., that most likely to produce significant results) will be published, and thus available for inspection.

Although some meta-analysts have essentially dismissed the problem of publication bias (e.g., Smith, 1980), others have provided some procedures for assessing the potential impact of publication bias on meta-analysis (e.g., Dawes, Landman, & Williams, 1984; Kurosawa, 1984; Orwin, 1983; Rosenthal, 1979, 1980b). As discussed previously (see Chapter 2), Lane and Dunlap (1978) have shown that publication bias may greatly distort reported effect size estimates. Although their results have limitations (most notably in that they did not consider sample sizes larger than 20), their results are nevertheless disturbing. While they do not draw from their research any lessons for meta-analysis, the implications are clear: effect sizes cumulated across studies may seriously overestimate population effect size. Curiously, the Lane and Dunlap (1978) study seems not to have made any impact on practitioners of meta-analysis: a quick survey of journal articles, conference presentations, books and book chapters on meta-analysis and effect size estimation written after 1978 revealed only a single reference (Green & Hall, 1984) to the Lane and Dunlap (1978) results--among 250 works!

Publication bias should not be a problem for Study II for two reasons: (a) It is expected that the overall effect size for spontaneous recovery research will be so small that overestimation will be inconsequential for practical purposes (i.e., the difference between 1% and 2% of accounted for variance is small even if one is 100% larger than the other). Furthermore, an inflated effect size will not materially affect the interpretation of results if the resulting estimate of statistical power is low (e.g., < .60). (b) A large proportion of spontaneous recovery studies reported nonsignificant results (hence the controversy) so that the effects of publication bias on effect size estimation are likely to be greatly diminished, though undoubtedly some publication bias is likely to be present in this literature, especially during the early years of the research.

Method

Selection of Studies

Research on spontaneous recovery began with the studies by Underwood (1948a, 1948b). Such a well-marked origin is convenient, as it provides a point of departure for a bibliographic search. Two general indices were searched for references to spontaneous recovery: Psychological

Abstracts and Social Science Citation Index. In addition, 11 journals which publish research on human learning and memory were searched exhaustively: Americal Journal of Psychology, Bulletin of the Psychonomic Society, Journal of Experimental Psychology, Journal of Psychology, Journal of Verbal Learning and Verbal Behavior, Memory & Cognition, Perceptual and Motor Skills, Psychological Record, Psychological Reports, Psychonomic Science, Quarterly Journal of Experimental Psychology.

The goal was to include as many studies as possible in the analysis. To this end, studies listed in Dissertation Abstracts and in conference proceedings were also included in this research. The primary disadvantages of these sources are that (a) abstracts frequently provided insufficient information for meta-analysis and power analysis, and (b) the sources themselves were often not readily available. Inclusion of studies from these sources should, however, help reduce the effects of publication bias.

A total of 40 spontaneous recovery studies conducted between 1948 and 1979 were discovered. These studies contained 48 separate experiments. Most of the studies were retrieved from the published research literature ($n = 35$). The remaining sources were divided between

dissertations ($n = 3$) and conference proceedings ($n = 2$). A list of the 40 studies included in the analysis is given in the Appendix.

Procedures for Meta-Analysis

Meta-analysis is a term coined by Glass (1976a, 1976b, 1977) to refer to the quantitative synthesis of results from many separate studies, usually in conjunction with an otherwise traditional literature review. Although the term is new and the technique has enjoyed great popularity in the past few years, meta-analysis is more accurately described as a collection of procedures. Most of the basic procedures have been known for many years.

Meta-analysis may be useful for answering two basic questions of a series of independent studies: (a) Is there statistical support for the research hypothesis? (b) If so, what is the magnitude of the effect? Procedures designed to answer the first question may be described generally as significance tests on the results of separate studies. The unit of analysis is the study itself, and the data analysed may be the probability level attained by the separate significance tests, or the significance test values themselves, or some transformation of these. Procedures designed to answer the second question are essentially

equivalent to averaging the effect size results of the separate studies.

Significance level. Procedures for assessing the overall level of statistical support for an effect are most useful when the results of some studies have been statistically significant while the results of other studies have not. Since this is precisely the situation for spontaneous recovery research, the first step in the meta-analysis of this literature was to summarize the statistical evidence for the effect.

Rosenthal (1978, 1980b, 1982) and Hunter, Schmidt, and Jackson (1982) have described various techniques suitable for summarizing significance levels. Perhaps the simplest and most generally useful technique is the "method of adding z scores" (Rosenthal, 1978):

$$\underline{z} = \text{sum } \{z(i)\} / \text{sqr}(N) , \quad (22)$$

where $\underline{z}$ is the normal score associated with the overall level of significance, $z(i)$ is the normal score associated with the probability level attained by the significance test for the i th experiment, and N is the number of experiments included in the analysis. Although simpler than other combining procedures, the method of adding $\underline{z}$

scores nevertheless gives results in good agreement with the results of other techniques, such as the well-known Fisher (1932; Winer, 1971) chi-square method for combining probabilities.

The method of adding z scores requires a knowledge of the exact value of the test statistic and its associated degrees of freedom so that a z score corresponding to the significance level (p) attained by the test statistic can be determined. A problem encountered here was that the value of the test statistic was not given in 13 of the experiments which reported nonsignificant results. For these studies a z score of zero was assigned, even if the direction of the results supported an interpretation of spontaneous recovery (Rosenthal, personal communication, 24 July 1980), as was frequently the case. Finally, if the results of an experiment were opposite to that predicted by the hypothesis of spontaneous recovery, a negative z score was recorded. This occurred infrequently (for three studies and for four experiments).

Average effect size. Given the overall significance of the accumulated z scores, meta-analysis proceeds by computing an average effect size for the research literature. Since effect sizes are rarely reported, they

generally have to be estimated for each experiment. The most convenient effect size measures for analysis of variance designs are the proportion of accounted for variance statistics, η^2 and ω^2 . η^2 was used in the present study since it is somewhat easier to compute for factorial designs than ω^2 and is more easily related to the noncentrality parameter (λ) of the noncentral F distribution, thus facilitating subsequent power calculations. Furthermore, both η^2 and ω^2 give very similar results when, as expected in this study, the magnitude of the effect is small.

Eta-squared is easily computed from the F statistic and degrees of freedom for the significance test of spontaneous recovery. For a two-factor design, Haase (1983) gives the following:

$$\eta^2(AB) = \frac{df(AB)F(AB)}{df(A)F(A) + df(B)F(B) + df(AB)F(AB) + df(E)}, \quad (23)$$

where $df(A)$, $df(B)$, $df(AB)$, and $df(E)$ are the degrees of freedom for factors A and B, the interaction AB, and the error term, respectively, and $F(A)$, $F(B)$, and $F(AB)$ are the F ratios obtained for factors A and B and for the interaction, respectively.

One problem frequently encountered in using equation 23 was that the value of nonsignificant F ratios were often not reported. In this case, Cohen's (1977) f was computed using the reported means and standard deviations of the groups (see equations 14 and 15). Cohen's f was then converted to η^2 using the following relationship:

$$\eta^2 = f^2 / (1 + f^2) . \quad (24)$$

In some cases no information at all was given about the spontaneous recovery effect except that it was not statistically significant. When this occurred an effect size of zero was recorded.

Power Analysis

Using the average effect size estimate as the alternative, power was calculated for each experiment based on the degrees of freedom for the significance test of spontaneous recovery which was used in the original analysis. Power calculations were based on the cube root normal approximation of the noncentral F distribution (Laubscher, 1960; Severo & Zelen, 1960), as previously described for Study I. Power was also computed for each study using Cohen's (1977) definitions of small, medium,

and large effect size ($f = .1, .25, \text{ and } .4$, respectively). These effects account for 1, 6, and 14 percent of the variance, respectively. All power computations were performed with $\alpha = .05$. Four power estimates were thus obtained for each experiment. These estimates were then averaged across experiments to obtain summary power estimates. Thus, the experiment (not the study) was used as the unit of analysis.

Results

Meta-Analysis

Significance level. Equation 22 was used to determine the overall level of significance for the spontaneous recovery effect. The z score for the combination of all 48 experiments was large ($z = 8.333, p < .001$). Because visual inspection suggested that larger studies tended to produce more nonsignificant results, the combined z was recalculated with weighting based on sample size (Rosenthal, 1978). The resulting z was only slightly smaller ($z = 8.063$). The use of three other combining methods given by Rosenthal (1978) produced similar results (testing the mean p , testing the mean z , and the Fisher chi-square procedure).

Since only 48 experiments were included in the analysis, the results may be vulnerable to the "file

drawer" problem (Rosenthal, 1979). That is, because significant results are more likely to be published than nonsignificant results (which have been filed away), significant results may be overrepresented in the calculation of equation 22. Although this is often a problem in meta-analysis, it may not be in the case of spontaneous recovery. Since 28 of the 48 experiments (58%) reported nonsignificant results, there seems to have been no hesitancy to report such findings. (Indeed, the opposite case might be made for the reporting of nonsignificant results, a "bandwagon" effect!).

Rosenthal (1979) has provided quantitative procedures for estimating the tolerance for a combined set of studies for unpublished or future null results. That is, how many null results must be filed away to reduce the level of significance indicated by equation 22 to just greater than the .05 level? This can easily be determined by rearranging terms in equation 22, solving for N , and subtracting a constant (equal to the number of studies which are known to exist). The resulting "fail-safe" number for spontaneous recovery experiments is 1184. The likelihood of there being that many unpublished null spontaneous recovery results seems remote.

Magnitude of the effect. Not surprisingly, no direct estimates of effect size were reported in any spontaneous recovery study. More disturbing, however, was that many studies did not provide sufficient data for post hoc estimation of effect size. Although the F test results for significant interaction (spontaneous recovery) effects were almost always given, F test results for main effects frequently were not given, even when these were significant, since these tests are usually of no interest. Thus, it was often not possible to use equation 23.

Use of a reduced equation, including only the interaction and error terms of equation 23, was considered but not used. This would have been equivalent to computing partial eta-squared (Cohen, 1973a). In general, this solution would be satisfactory, since the unreported F ratios are usually not significant. Thus, such F ratios would be small relative to the error degrees of freedom, and so would exert only a small influence on the denominator in equation 23. Unfortunately, this procedure was not practical in the present case since the main effects of a spontaneous recovery design (differences in recall between groups and across retention intervals), though usually of no interest, are almost always large. Thus, estimating partial eta-squared would have greatly

over-estimated the magnitude of the spontaneous recovery effect.

Most surprising was that many studies did not report even the most basic descriptive statistics. While group means were almost always reported, standard deviations frequently were not. Thus, even Cohen's f statistic could not be computed for many experiments. It is interesting to note that 75% of the spontaneous recovery studies were published in two of the most respected journals of experimental psychology: the Journal of Verbal Learning and Verbal Behavior and the Journal of Experimental Psychology.

In all, there were only 10 experiments for which effect size could be determined. Since several experiments contained more than one estimate, a total of 19 effect size estimates were obtained. These estimates were averaged within experiments first, and then averaged across experiments. The resulting average proportion of variance accounted for (η^2) was .032.

Power Analysis

All power estimates were based on 45 experiments, since there was no indication of sample size in three of the experiments. Power was computed for each experiment using the average effect size $\eta^2 = .032$ as the alternative

hypothesis. Alpha was assumed to be .05 for all studies.

Average power across experiments was .375.

Power was also computed for each experiment assuming Cohen's definitions of small, medium, and large effect sizes as the alternative hypotheses. These results are shown in Table 14 and are reported separately for experiments which obtained significant spontaneous recovery effects and for experiments in which the spontaneous recovery effect was not significant.

Table 14

Average Power of Spontaneous Recovery Studies

Significance Result	<u>n</u>	Effect Size		
		Small	Medium	Large
p < .05	18	.158	.662	.944
p > .05	27	.139	.576	.915
Total	45	.146	.611	.928

Note. n is the number of experiments in each significance category.

Discussion

Spontaneous recovery has been a controversial phenomenon for 20 years. Meta-analysis and power analysis of spontaneous recovery studies suggests, however, that

this controversy may, at least in part, be artifactual. The results of the meta-analysis show that the spontaneous recovery effect is "real" (that is, statistically significant), but that the effect is small. The average proportion of variance accounted for was .032, somewhat smaller than Cohen's (1977) definition of a medium-sized effect. Of course, Cohen was mainly speaking of research in the "softer" areas of psychology when he defined effect sizes. Thus, for a laboratory research phenomenon, the magnitude of the spontaneous recovery effect is indeed small.

The estimate of effect size for spontaneous recovery is best regarded cautiously, since it was based on data from only 10 experiments. Furthermore, only two (20%) of these studies reported significant results, compared to about 40% for the entire set of spontaneous recovery studies. Thus, it is possible that these results underestimate the size of the spontaneous recovery effect. Nevertheless, the magnitude of the effect probably does not exceed Cohen's definition for a medium-sized effect.

These results indicate that the "inconsistency" attributed to the analysis of variance interaction may be due to the emphasis reviewers have placed on the level of

significance attained by individual studies, rather than that attained by the entire series of spontaneous recovery studies. In fact, these results demonstrate that the interaction test has provided convincing support for the existence of spontaneous recovery.

It may seem surprising that such large z scores were obtained when less than half of the experiments produced statistically significant results. However, many of the nonsignificant outcomes were in the direction predicted by the hypothesis of spontaneous recovery. Brown (1976) was one of the few reviewers of the spontaneous recovery literature who concluded that there was sufficient evidence for the effect. By relying more on the direction of results rather than on actual statistical outcomes for a series of studies, Brown's conclusions more nearly resemble those of a meta-analysis than a traditional review of the literature. Brown's approach is not a typical procedure among behavioral researchers, who are more apt to be influenced by the proportions of significant and nonsignificant results in a research literature (counting noses, according to Meehl, 1978). Brown's approach, and that of meta-analysis in general, is more in the spirit of Platt's (1964) strong inference, which emphasizes converging lines of evidence. Since the evidence for

spontaneous recovery has been considered weak by most reviewers (including Rossi, 1977), the results of this analysis seem to support the findings reported by Cooper and Rosenthal (1980) that the use of statistical methods of summarizing research results will show more support for an hypothesis than will traditional reviewing methods.

Based on the obtained estimate of effect size, power was computed for each experiment. The average power for the entire set of spontaneous recovery studies was low. This is not surprising, given the small size of the effect. Moreover, the average power estimate of .375 agrees well with the actual proportion of experiments which obtained significant results: .417 (20 of 48 experiments significant at $p < .05$).

Table 14 shows the results of a "traditional" power survey of the spontaneous recovery literature. It is clear from these results that the power of spontaneous recovery studies is inadequate to detect all but large effects. Even if, as previously suggested, the obtained effect size underestimates the magnitude of the spontaneous recovery effect, it is unlikely that the effect is as large as Cohen's (1977) large effect size, which accounts for 14% of the variance. More reasonable would be to assume a

medium-sized effect for spontaneous recovery. The average power of the spontaneous recovery experiments to detect medium-sized effects was .611.

Although the power of spontaneous recovery research is low, it is adequate to insure that at least some studies will obtain significant results. It is easy to see how controversy over the existence of the effect might develop under these circumstances. This is especially so for spontaneous recovery, since it is unlikely that researchers in this field ever seriously considered the issue of statistical power, either a priori or in the evaluation of null results. As a laboratory endeavor, it was probably assumed that sufficient control of the experimental situation would be all that was necessary to elucidate the effect. Small effects are unusual in laboratory studies of human learning and memory, so that the issue of statistical power is generally not raised.

The average sample size of research in spontaneous recovery was about 80, somewhat larger than research in other areas of psychological research (Holmes, 1979; Holmes et al., 1981). If the magnitude of the spontaneous recovery effect is as small as suggested by the results of this study, then even larger sample sizes would be required in any future research. Of course, it is important to

remember that smaller sample sizes would also have avoided the controversy, since the effect would rarely have been detected. This point is important, since the null hypothesis is never exactly true. Thus, researchers should be able to compute the power of their studies, not only to detect the effect they seek, but also to avoid detecting a trivially small effect.

The critical question for spontaneous recovery researchers is, How small an effect should be considered unimportant? What constitutes a small effect is admittedly something of an arbitrary decision. If the goal of research is prediction or classification, then effects even as large as Cohen's large effect size may contribute very little information (Hyde, 1981; Rossi, 1983). If the goal of the research is theoretical then even a small effect, in Cohen's sense, may be an important starting point for future research. As the research continues, improvements in understanding the phenomenon under study should lead to more sensitive experimental tests, thus increasing the effective magnitude of the effect. This process has worked well in the physical sciences, where the investigation of small effects is a routine business (see Chapter 6).

Since spontaneous recovery occupied a critical position

in the interference theory of forgetting (Underwood & Postman, 1960), it may be that even a small effect would have been worth pursuing. Instead, the issue of spontaneous recovery seems to have faded away slowly during the 1970s, a fate not uncommon to theories in the "soft" areas of psychological research (Meehl, 1978), but unusual in laboratory research. In current textbooks on human learning and memory the interference theory of forgetting is still a popular topic. However, that aspect of the theory dealing with spontaneous recovery is rarely mentioned.

CHAPTER 5

Study III

Inadequate Statistical Power and Artifactual
Controversy in Psychological Research

The purpose of Study III was to generalize the findings of Study II. Since spontaneous recovery is not the only research literature which is plagued with inconsistent findings, the role of statistical power in other research domains was investigated in Study III. Meehl (1978) has noted that many theories in "soft" psychology fade away when researchers lose their enthusiasm as inconsistent and unreplicable results accumulate. That this outcome is not limited to social and clinical research is evident from the fate of spontaneous recovery.

One problem with conducting many power analyses such as that done in Study II is the necessity of also performing the prerequisite meta-analyses to determine the average effect size. Such meta-analyses are very time-consuming, since the literature must be searched exhaustively, and effect sizes then computed for every study. Furthermore, such an endeavor almost always requires the participation of a subject matter expert.

Fortunately, a great number of meta-analyses have been

published in recent years. Many, though by no means all, of these have been conducted specifically because of the inconsistent state of a particular research literature. Use of already published meta-analyses is particularly advantageous in that the literature searches have already been conducted and the effect size calculations have already been made.

Method

Selection of Meta-Analyses

The search for meta-analyses was conducted using the subject indices of Psychological Abstracts and Social Science Citation Index. In addition, several journals which typically publish review articles were examined exhaustively, including the American Educational Research Journal, American Psychologist, Journal of Consulting and Clinical Psychology, Journal of Educational Psychology, Journal of Personality and Social Psychology, Personality and Social Psychology Bulletin, Psychological Bulletin, Review of Educational Research. Journal searches commenced with the 1978 volumes. The reference lists of recently published books on meta-analysis were also examined (Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1980a).

A total of 95 meta-analyses published between 1977 and

1984 were discovered. The distribution of sources was as follows: 81 journal publications; 5 book chapters; 4 conferenced presentations; 4 unpublished manuscripts obtained from the authors; and 1 book. The most popular publication outlets were Psychological Bulletin and the Review of Educational Research.

To be suitable for power analysis, a meta-analysis must provide certain minimum information. In particular, sample size and effect size information must be listed for every primary study included in the meta-analysis. These data are required for power to be estimated for each study. Unfortunately, most of the meta-analyses ($n = 89 = 94\%$) did not give both sample size and effect size for each study included in the analysis. In fact, most of the meta-analyses did not report any individual study results. Instead, only average effect size and its correlation with suspected moderator variables was usually reported (e.g., Landman & Dawes, 1982; Smith & Glass, 1977; Willson & Putnam, 1982). This situation seems somewhat ironic, since meta-analyses are themselves dependent on full reporting of data in the primary studies.

The remaining six meta-analyses included five published in research journals and one unpublished manuscript. The

results of 17 separate meta-analyses were given in the six reports, since several reports included the results of more than one meta-analytic review. Of these 17 meta-analyses, four were excluded because they were conducted on research areas of minimal controversy: more than 80% of their constituent primary studies reached similar conclusions (i.e., more than 80% or fewer than 20% of the primary reports obtained statistically significant results). Finally, one additional meta-analysis was deleted because it included data from only four primary studies. The remaining 12 meta-analyses were based on a total of 215 studies and 160122 subjects. A list of these meta-analyses and the substantive research area reviewed is shown in Table 15.

Procedures for Meta-analysis

All 12 meta-analyses used Cohen's (1977) d as the index of effect size. Average effect size reported in each meta-analysis was not used as the basis for power estimates. This was not done because several authors used questionable techniques for averaging effect sizes. In particular, several authors simply omitted from the effect size calculations primary studies for which Cohen's d could not specifically be computed.

For the present study, Cohen's d was estimated based on

Table 15

Meta-Analyses Included in Study III

Source	Area Reviewed
Arkin et al. (1980) #1	Self-serving attribution bias: Teacher-student paradigm
Arkin et al. (1980) #2	Self-serving attribution bias: Client-therapist paradigm
Arkin et al. (1980) #3	Self-serving attribution bias: Miscellaneous paradigms
Burger (1981)	Defensive-attribution bias
Cooper (1979) #1	Gender differences in conformity: Asch-type experiments
Cooper (1979) #2	Gender differences in conformity: Fictitious norm groups
Hall (1978) #1	Gender differences in decoding nonverbal cues: Visual
Hall (1978) #2	Gender differences in decoding nonverbal cues: Auditory
Hasenfeld (1983)	Gender differences in aggression
Hyde (1981) #1	Gender differences in verbal ability
Hyde (1981) #2	Gender differences in quantitative ability
Hyde (1981) #3	Gender differences in field articulation

the reported significance level obtained in the primary study. For example, if Cohen's d was not reported, but the study was listed as reporting no significant results, a d score of zero was assigned. If instead, a significance level was reported, then the lowest d value which would have resulted in that significance level, based on the reported sample size of the study, was assigned. In some cases, both Cohen's d and another index of effect size (such as ω^2) were reported in the meta-analysis. When Cohen's d was not given but the other measure was, then Cohen's d was determined by deriving equations relating the two measures of effect size. Recomputation of effect size also provided a check on the calculations in the reported meta-analysis results. Thus, the average effect size reported for meta-analyses here might differ from those reported by the original authors.

Power Analysis

Using the average effect size as the alternative to detect, power was computed for each study in each meta-analysis. These power estimates were then averaged for each meta-analysis.

A "traditional" power survey of each meta-analysis was also conducted. For each study power was estimated using Cohen's (1977) definitions of small, medium, and large

effect size as the alternative hypotheses. (Small, medium, and large effect sizes were defined by Cohen, 1977, as $d = 0.2$, 0.5 , and 0.8 , respectively, accounting for 1%, 6%, and 14% of the variance (η^2), respectively.) Averages were then obtained for each level of effect size by averaging across studies within meta-analyses. Thus, four power estimates were obtained for each primary study, three based on Cohen's definitions of small, medium, and large effect sizes, and one based on the average effect size found for the particular meta-analysis in which the primary study was reported. All power estimates assumed $\alpha = .05$.

Results and Discussion

The power analysis results for each of the 12 meta-analyses is given in Table 16. Also shown for each meta-analysis is the average effect size and the proportion of primary studies which obtained statistically significant results ($p < .05$). If the conflicting findings reported in a meta-analysis are due to inadequate statistical power, then it may be expected that the average power for the meta-analysis should be a reasonably good estimate of the proportion of studies which resulted in significant effects. These results are given in the last two columns of Table 16.

Table 16

Power Analysis of Meta-Analyses

Source	N ¹	Sample Size ²		Significance Rate	Power
		Total	Median		
Arkin et al. (1980) #1	8	709	88	.625	.754
Arkin et al. (1980) #2	7	553	80	.714	.895
Arkin et al. (1980) #3	8	549	68	.500	.522
Burger (1981)	22	4404	160	.273	.420
Cooper (1979) #1	16	2274	116	.438	.353
Cooper (1979) #2	8	992	100	.250	.082
Hall (1978) #1	55	12253	115	.236	.214
Hall (1978) #2	13	1896	119	.385	.161
Hasenfeld (1982)	11	768	80	.420	.394
Hyde (1981) #1	27	68899	528	.444	.502
Hyde (1981) #2	16	65193	620	.688	.878
Hyde (1981) #3	24	1632	54	.458	.358
Totals	215	160122	115	.400	.397

¹N is the number of studies included in each meta-analysis.

²Sample size refers to the total number of subjects in each meta-analysis and to the median number of subjects included in each study within the meta-analysis.

³d is the average effect size for each meta-analysis, in standard deviation units. Average effect size for all meta-analyses was weighted by N.

⁴Significance rate is the proportion of all studies in each meta-analysis reporting statistically significant effects ($p < .05$).

⁵The power of each primary study in each meta-analysis was computed based on the average effect size for that meta-analysis, on the sample size of the study, and assuming $\alpha = .05$. Average power for all meta-analyses was weighted by N.

Although an indication of the concordance of these two measures for the entire set of meta-analyses is really of only secondary importance, it is interesting to note that the correlation is fairly good ($r = .904$). The correlation seems especially good since the ends of the distribution were eliminated, that is, the low controversy studies. For these studies power and significance rate were either both high or both low, and their inclusion would have inflated the correlation.

Another way to assess the relationship between power and significance rate for the entire set of meta-analyses is given by a comparison of the average power to the overall proportion of significant results. These figures were very similar: the average power for all 215 studies was .397, while the proportion of all primary studies obtaining statistically significant results was .400.

Although these results are suggestive, global assessments of the relationship between power and significance rate suffer the same disadvantages as traditional power surveys: results are collapsed across disparate research areas, and thus may not be readily generalizable. More appropriate is a consideration of the power results for each meta-analysis separately, for this will give an indication of the extent to which the case of

spontaneous recovery is an isolated one. For six of the meta-analyses the correspondence between the power estimate and the significance rate seems reasonably good (Arkin, Cooper, & Kolditz, 1980, #3; Cooper, 1979, #1; Hall, 1978, #1; Hasenfeld, 1983; Hyde, 1981, #1 & #3). For two of the remaining meta-analyses power was substantially lower than significance rate (Cooper, 1979, #2; Hall, 1978, #2). This outcome is not too surprising, since significant results are more likely to be published than nonsignificant results.

For the four remaining meta-analyses power was greater than significance rate (Arkin et al., 1980, #1 & #2; Burger, 1981; Hyde, 1981, #2). Thus, for these four areas of research inadequate statistical power cannot completely account for the existence of conflicting findings. This is especially true for three of these meta-analyses, for which power ranged from .754 to .895. These are substantial levels of power and should be satisfactory for most areas of behavioral research. Controversy in these areas may be due to any of a number of possible sources: different methodologies or measuring instruments, differential failure to control extraneous sources of error, incompatible operational definitions, conflicting

theoretical orientations, etc. It should be noted, however, that for one of the four meta-analyses (Burger, 1981) power, although much higher than significance rate, was still quite low (.420), so that the influence of inadequate power in this area of research cannot be ruled out.

Perhaps the most disappointing aspect of these results is that for only three of the 12 areas of research can inadequate power be ruled out as a source of the conflicting findings. For the remaining areas, inadequate power appears to be at least partially responsible for the conflict, although only further research within each specific area can determine whether there are yet additional sources contributing to the existence of the conflicting findings.

Primarily responsible for this poor state of affairs are the magnitudes of the effects being investigated. As shown in Table 16, average effect size across all meta-analyses was only .278 (Cohen's d). The average proportion of variance accounted for is .019! These are small effects indeed, and it should not be surprising that there may be difficulties in demonstrating significant relationships in these research areas. In no case, however, was statistical power mentioned in these

meta-analyses as a potential source of the conflict. This is in spite of the fact that meta-analysts are, in general, more aware of statistical issues than most other behavioral researchers. In fact, increased statistical power is often given as one of the primary reasons for cumulating results across studies.

The small effect sizes characteristic of these meta-analyses is emphasized by the results of the "traditional" power survey shown in Table 17. Here it can be seen that the power of the primary studies in these meta-analyses is higher than the power obtained in previously conducted power surveys. Especially noteworthy is that the average power for detecting medium-sized effects is .707, compared to .61 for other power surveys (see Table 6). The results shown in Table 17 might suggest that the power of the studies in these meta-analyses is good. However, as has already been pointed out, the actual average power for the primary studies is much lower, only .397! This is a good indication of the limitations inherent in traditional power surveys: the range of effect sizes suggested by Cohen (1977) are simply not very likely to be characteristic of the actual effect sizes in any particular research literature.

Table 17

Power Surveys of Meta-Analyses

Meta-Analysis	Effect Size		
	Small	Medium	Large
Arkin et al. (1980) #1	.152	.631	.941
Arkin et al. (1980) #2	.139	.589	.930
Arkin et al. (1980) #3	.126	.509	.836
Burger (1981)	.273	.792	.960
Cooper (1979) #1	.216	.744	.952
Cooper (1979) #2	.194	.703	.942
Hall (1978) #1	.299	.707	.871
Hall (1978) #2	.221	.731	.943
Hasenfeld (1983)	.130	.523	.848
Hyde (1981) #1	.567	.852	.959
Hyde (1981) #2	.666	.938	.989
Hyde (1981) #3	.128	.482	.792
Averages	.298	.707	.906

In fact, Table 17 shows that the power of two meta-analyses was excellent. Hyde's (1981) #1 and #2 meta-analyses had power values of .567 and .666, respectively, to detect small effect sizes. Median sample sizes for the primary studies in these meta-analyses were 528 and 620, respectively. Thus, it is understandable why inadequate statistical power was never suggested as a problem. Yet, the effect size for Hyde's first meta-analysis was so small ($d = .166$, $\eta^2 = .007$) that average power was still quite low (.502).

The results of Study III should not be interpreted to mean that all, or even most, conflicting findings in behavioral research are due to inadequate statistical power. Certainly there may be many reasons why the results of various studies do not agree: variations in subject characteristics, or in test administrations, for example. It is even for such reasons that the effect size under investigation may be small in the first place--lost in the noise of uncontrolled extraneous factors. Nevertheless, in many areas of scientific research, conflicting findings generate lively debate, and often precede major theoretical advance (Kuhn, 1970). Unfortunately, this does not seem to be the case in the behavioral sciences, where dependence on statistical tests to establish the existence of an effect

renders the researcher vulnerable to the power of those tests, a consideration frequently ignored. Some of the reasons for the differences between the behavioral sciences and the physical sciences are explored in the next chapter.

CHAPTER 6

Behavioral and Physical Sciences Compared

It has been said that true science does not need statistics. It is doubtful that many scientists believe this statement. Certainly even a casual reading of such major journals as Science and Nature reveals that the behavioral sciences are not the only disciplines which use statistical tests and statistical inference.

In fact, it is evident that the behavioral sciences do not even have a monopoly on small effect sizes. One example is the anomalous advance of the orbit of Mercury, discussed at the end of Chapter 2. Another, perhaps quintessential example is that the total amount of energy from outside the solar system ever received by all the radio telescopes in the world is less than the energy of a single falling snowflake (Sagan, 1980)! The following examples are from recent issues of Science.

1. Raup and Sepkoski (1982) reported a decline in the extinction rate of families of marine life over the past 600 million years. Their regression r^2 was .22 after major extinctions (outliers) were removed. Quinn (1983) reanalyzed their data and reported a regression r^2 (for all data points) of .07 if a normal model for extinction rates

was assumed and an r^2 of .16 if a log-normal model was assumed.

2. Harvey, Prabhakaran, Mao, and McCammon (1984) conducted a computer simulation study of the lengths of hydrogen bonds in phenylalanine transfer RNA. They stated that the lengths of these bonds "should be highly correlated (emphasis added) as the bases move away from one another. This is confirmed by the positive coefficients for bonds in adenine-uracil (AU) pairs and for neighboring bonds in guanine-cytosine (GC) pairs" (p. 1190). The correlations (not r^2 's) for the seven AU pairs ranged from -.23 to .35, with an average correlation of .17 (SD = .21). For the 25 adjacent GU pairs correlations ranged from .07 to .70, with an average correlation of .36 (SD = .19). The average correlations were not reported by the authors.

3. Cranston et al. (1984) have recently reported that the parasite P. falciparum reduces the ability of red cells to recover their normal shape after deformation by shear fluid stress. Time to recover for 70 normal cells was 121 msec (SD = 29) and for 70 infected cell was 137 msec (SD = 36). This difference was statistically significant: $t(138) = 2.896$, $p < .005$. The authors did not report the proportion of variance accounted for by their results: $\omega^2 = .05$.

Examples of the use of statistics in physics--the hardest of the "hard" sciences--are a bit more difficult to come across. No recent issues of Science, Nature, or the Proceedings of the National Academy of Sciences included any physics reports which used statistical hypothesis testing procedures of the kind that would be familiar to behavioral scientists. However, an interesting example of hypothesis testing in physics was recently discussed by Franklin (1981).

In the 1960s physicists conducted an intensive search for new elementary particles. The criterion for the establishment of a new particle was three standard deviations (i.e., $p < .0027$) above the background noise. Several particles reported in the literature could not be confirmed by subsequent research. The problem of "false alarms" was solved when the required significance level was increased to four standard deviations ($p < .000064$). The behavioral scientist will recognize the problem as that of the inflated type I error rate following multiple comparisons!

Significance testing of the sort familiar to behavioral researchers is rare in the physical sciences. One is much more likely to encounter simple descriptive statistics,

such as the mean and standard error, in most physical science journals. Why is this? The answer has to do with the different roles played by theory in the physical and behavioral sciences.

In the behavioral sciences theories are usually weak, and so are only able to make simple predictions: an effect is predicted to exist or not to exist. In contrast, theories in the physical sciences are typically so powerful that they are able to predict the magnitude of the effect that is being sought. In the behavioral sciences it is often sufficient to establish one's hypothesis simply by rejecting the null hypothesis. However, since no precise alternative is predicted, rejecting the null hypothesis is consistent with an infinite number of alternative hypotheses! Furthermore, as has frequently been pointed out by methodologists, the null is a straw-man hypothesis, since the a priori probability of its being exactly true is vanishingly small (Bakan, 1966; Meehl, 1967, 1978).

In the physical sciences it is not sufficient to reject a null hypothesis: results must be consistent with (i.e., not significantly different from) theoretical predictions. A recent series of experiments in quantum mechanics is particularly illustrative. Since its development in the 1920s quantum mechanics has been very successful in

predicting the behavior of matter at the atomic and subatomic levels. Unfortunately, certain predictions of quantum theory conflict with expectations based on two cherished assumptions of physical theory: (a) locality--the premise of the special theory of relativity that information cannot be transmitted faster than the speed of light, and (b) realism--the idea that physical systems have well-defined properties whether they are measured or not, that is, that such properties are independent of human observation.

To avoid the uncomfortable philosophical implications of quantum mechanics, physicists have proposed a class of theories known as local hidden variables theories. These theories postulate that the probabilistic character of quantum mechanics is not due to the indeterminateness of nature, but instead is simply an emergent feature of the operation of unknown (hidden) variables. If these variables could be revealed, physical theory would once again become deterministic.

Aspect, Grangier, and Roger (1981) tested certain predictions of local hidden variables theory against those of quantum mechanics. The quantity Δ was predicted to be less than zero by hidden variables theory and to be

0.058 ± 0.002 by quantum mechanical theory. The observed result was 0.0572 ± 0.0043 , violating the prediction of hidden variables theory by more than 13 standard deviations, but in good agreement with the quantum mechanical prediction. Similarly, the quantity $\underline{S}$ was predicted by hidden variables theory to be between -1 and 0, and by quantum theory to be 0.118 ± 0.005 . The obtained result was $\underline{S} = 0.126 \pm 0.014$.

Subsequent attempts to confirm these results using more refined techniques were successful. In one such experiment the quantity $\underline{S}$ was predicted by local hidden variables theory to be between -2 and 2 and by quantum mechanics to be 2.70 ± 0.05 . The observed result was 2.697 ± 0.015 (Aspect, Grangier, & Roger, 1982). The violation of hidden variables theory was 46 standard deviations! The obtained results agreed well with the quantum mechanical prediction. In yet a third series of experiments $\underline{S}$ was again predicted to be between -1 and 0 by hidden variables theory (Aspect, Dalibard, & Roger, 1982). The prediction of quantum mechanics was 0.112. The obtained result was $\underline{S} = 0.101 \pm 0.020$.

What is particularly remarkable about these experiments--from the standpoint of the behavioral scientist--is not only the extent to which the predictions

of one theory are rejected, but also the degree to which the data agree with the predictions of the alternative theory. So good are these results that the graph captions included in these papers must point out that the displayed functions are predicted and have not been fit to the plotted data points! Not surprisingly, these studies are considered to be responsible for the death of local hidden variables theories (Rohrlich, 1983). (The reader interested in the philosophical implications of these results should consult d'Espagnat, 1979, Robinson, 1982, 1983, and Rohrlich, 1983.)

It is the degree of accuracy and precision in measurement, perhaps more than anything else, that gives physical science its enormous advantage and success, compared to the other sciences, in revealing small effects (Pipkin & Ritter, 1983). The example given at the end of the last section--the advance of the perihelion of Mercury--provides a good illustration. The magnitude of the effect is very small, but the deviation from the inverse square law that was proposed to account for the anomalous advance had to be rejected because the revision produced an unacceptably large deviation when applied to the known orbit of the Moon (Roseveare, 1982). Thus, it

was the precision with which the Moon's orbit was known that allowed a prediction about the orbit of Mercury to be rejected.

Although the remarkable precision of the physical sciences may seem out of reach for the behavioral sciences, it should be encouraging to note that the statistical procedures used to confirm and reject predictions in the physical sciences are really no different, conceptually, from those used in the behavioral sciences. These procedures are essentially the acceptance and rejection of null and alternative hypotheses (although one might well quibble concerning which is which). The real difference in procedure between the physical and behavioral sciences lies in the insistence by physical scientists that one prediction be confirmed while the other is rejected. Again, it is the difference in the role played by theory that gives the physical scientist the advantage.

This advantage can be exploited in ways that the behavioral scientist can not normally be expected to follow. Physical theory has resulted in certain predictions which may seem, superficially, to be strange: for example, that the force of gravity has been weakening, or that the solar constant has been decreasing. There is no evidence for these speculations, other than that, if

true, certain other anomalous conditions may be accounted for. The respect for physical theory is strong enough so that such predictions are not dismissed without some consideration and, if possible, test.

In some cases, physical theory is stretched to the limit, unable to predict a magnitude for the size of an effect. Two of the strangest predictions of quantum theory are of this sort: (a) that the photon does not have a rest mass of zero, as is usually assumed, but has instead a very small, though finite, mass; and (b) that the proton is not stable, but has an extremely long, though again finite, half-life. In both cases, specific predictions of mass and half-life, respectively, can not be made. However, the regard for quantum theory is great enough so that considerable time and money have been expended to detect the effects. All such attempts have so far failed. Rather than conclude that these effects do not exist, however, physical scientists have simply concluded that, if they do in fact exist, these effects must be smaller than a certain magnitude, that magnitude being determined by the error of measurement in the relevant experiments.

In both cases the limits are quite impressive. For the rest mass of the photon the upper bound is 8×10^{-48} gram

(Davis, Goldhaber, & Nieto, 1975; Goldhaber & Nieto, 1976). For the proton, the upper limit on half-life is more than 10^{30} years (Weinberg, 1981). Ironically, the situation is not so dissimilar to one which frequently occurs to the hypothesis-testing behavioral scientist: failure to reject the null hypothesis! The behavioral scientist in this situation often asserts, incorrectly, that there is no effect. The physical scientist simply assigns an upper bound on the magnitude of the effect, and leaves a decision on the existence of the effect to future research.

The behavioral scientist who has estimated statistical power for a range of possible effect sizes is in a position somewhat similar to that of the physical scientist. In this situation, as illustrated in Table 3, the behavioral scientist has had to consider the possibility that the sensitivity of the planned experiment may not be great enough to detect the effect. One of the most important results of a consideration of statistical power, then, may be that attention is focused on effect size as a kind of signal-to-noise ratio. A decrease in noise results in a more effective signal. Thus, effect size need not be thought of as a fixed and immutable quantity that we strive to detect. Instead, behavioral scientists should realize that effect size may be effectively increased through

increased measurement accuracy (noise reduction). The notion of a "population effect size" is thus really meaningless, though it may serve as a convenient fiction as an aid in the determination of statistical power.

The emphasis on effect size as a signal-to-noise ratio may be the single most important lesson that the behavioral sciences can learn from the physical sciences, or for that matter, from statistical power analysis. The physical sciences have achieved an impressive record in discerning small effects by combining strong theory with highly precise measurement techniques. The real success of the physical sciences is in effectively increasing the signal-to-noise ratio, essentially making effect sizes larger. Among the behavioral sciences, only in the area of psychophysics has this idea been well established, though largely through a route independent of statistical power analysis (i.e., signal detection theory).

Many of the previously discussed examples of small effects in the physical sciences are not really small at all, at least not in the sense usually held by behavioral scientists. Superficially, of course, such effects are small, in much the same way that a virus is small. However, such effects as the anomalous advance of Mercury's

orbit are easily discernable when sufficiently precise measurement techniques are applied to the problem.

The degree of measurement accuracy that physical science can bring to bear on a problem is many orders of magnitude greater than that available to behavioral scientists. Perhaps this difference in technical ability is responsible for the different roles played by theory in the physical and behavioral sciences. These roles have been described briefly already. But there is a difference in these roles that is especially indicative of the approaches to science developed by physical and behavioral scientists.

Although theories are weak in the behavioral sciences, there is considerable emphasis in behavioral research to fit research results into some sort of theoretical framework. The pursuit of purely empirical research is frowned upon and such work has trouble getting published.

Empiricism is highly regarded in the physical sciences, wherein it is generally recognized that one of the basic problems of science is to make things measurable (Lenzen, 1938). It may seem paradoxical, but in the physical sciences theory is not nearly so highly regarded as in the behavioral sciences, even though theories in the physical sciences are much more powerful and have been much more

successful than theories in the behavioral sciences.

The reason for this is that physical scientists regard theories with skepticism, as candidates for rejection. Theories are made specific and thus easy to falsify. Empiricism is highly regarded because measurement provides the raw data with which to put theory to the test. Theories are not (or should not be) regarded too highly, since the ultimate fate of all theories is to be discarded.

The situation is clearly different in the behavioral sciences, as Meehl (1967, 1978) has noted. Good theories are difficult to construct, and there is a natural reluctance to let go in the face of "mere facts." Thus, one finds much ad hoc explanation and reconstruction to salvage theory. Successful theories in the physical sciences must undergo increasingly difficult tests to survive (Popper, 1959). The opposite seems generally to be true in the behavioral sciences. The result in the physical sciences is that successful theories become extremely strong and precise as they survive repeated attempts at falsification. Unsuccessful theories are continually replaced by stronger candidates. In the behavioral sciences successful theories discourage any attempt at falsification by incorporating whatever aspects

of their competitors that are necessary to survive. Such theories are able, eventually, to tolerate almost any outcome of experimental test, especially given the nature of the statistical hypothesis testing procedure. Their fate is to slowly fade away, without clear-cut rejection, or even replacement, as researchers lose interest in the problem and move on to other matters (Meehl, 1978).

The lessons for psychology are clear. In the face of low statistical power for the effects of interest, it is not sufficient to increase sample sizes. An attempt to make effect sizes larger is also necessary. This can be done by increasing measurement accuracy through the use of more effective measurement models, more sensitive experimental designs, and more sophisticated statistical techniques (e.g., multivariate methods, causal modeling, time-series analysis, etc.). Construction of theories which make specific predictions concerning the magnitudes of effects would also be desirable. This will not be easy to do, since it will require a fundamental change in the nature of psychological theorizing. It is a goal psychologists should seriously contemplate.

CHAPTER 7

Summary of Findings

The statistical power of psychological research was examined in three studies. Study I provided a general assessment of statistical power in psychological research. All of the studies using statistical tests in the 1982 issues of the Journal of Abnormal Psychology and the Journal of Personality and Social Psychology ($N = 142$) were evaluated and compared to the results of a previous survey (Cohen, 1962). Power was computed against small, medium, and large effect size alternatives. Results indicated that the power of psychological research has increased over the past 20 years. Power to detect small effects has increased from .18 to .24. Power to detect medium-sized effects has increased from .48 to .59, and power to detect large effects has increased from .83 to .89. Although all increases were statistically significant ($p < .05$), the magnitudes of the changes were relatively small, and did not affect the characterization of statistical power in psychological research. Power for small effects is still very low, and power for large effects is still very good. Power to detect medium-sized effects showed the largest increase. Although this is an encouraging trend, power for

medium-sized effects is still not particularly good.

The results of Study I are limited because accurate estimates of effect size, on which power analysis must be based, are not given in most research reports. Although Cohen's (1962, 1977) suggested definitions of small, medium, and large effect sizes are useful, there is very little evidence to suggest that these definitions are characteristic of psychological research. Furthermore, the results of power surveys are based on heterogeneous sets of research articles, since all of the studies appearing in entire journal issues are included in the analyses.

In Studies II and III recently devised techniques of meta-analysis were used to provide estimates of effect size for subsequent power analyses. Use of these procedures permitted power analyses to be conducted using precise estimates of effect size, rather than a wide range of effect sizes, as in previous power surveys.

In Study II, a meta-analysis of the research on the spontaneous recovery of verbal learning was conducted. A total of 40 studies (including 48 experiments) conducted between 1948 and 1979 were included in the analysis. This area of research was selected because previous studies of spontaneous recovery have produced conflicting results concerning the existence of the phenomenon. Results

indicated that the evidence for spontaneous recovery is statistically significant, but that the magnitude of the effect is small ($\eta^2 = .032$). A power analysis was then performed on the spontaneous recovery studies using the effect size information obtained by the meta-analysis. Results indicated that the power of research on spontaneous recovery was low. Moreover, the average power estimate of .375 agreed well with the actual proportion of experiments which obtained significant results: .417 (20 of 48 experiments significant at $p < .05$).

The results of Study II suggest that the inconsistent findings reported in the spontaneous recovery literature may be an artifact of the low statistical power that characterizes this research area. This implication was examined further in Study III. Power analyses were conducted on each of 12 published meta-analyses, all of which had been performed on research literatures characterized by inconsistent results. The results of Study II were generally confirmed. For six of the meta-analyses statistical power was similar to the proportion of significant results. In only three of the meta-analyses was power relatively high ($> .75$), while for the remaining meta-analyses power was very low ($< .55$).

The low power of these research areas is due to the small size of the effects that are being investigated. Average effect size (η^2) across all meta-analyses was only .019. For effects of this magnitude only very large sample sizes will yield acceptable levels of statistical power (e.g., .80). It is recommended, therefore, that psychologists increase power by making effect sizes larger, that is by increasing the signal-to-noise ratio of research designs. Ultimately desirable is the construction of "strong" theories, which would enable researchers to make specific predictions concerning the magnitudes of the effects they study.

REFERENCES

- Abramowitz, M., & Stegun, I. A. (1965). Handbook of mathematical functions with formulas, graphs, and mathematical tables. New York: Dover.
- Arkin, R., Cooper, H., & Kolditz, T. (1980). A statistical review of the literature concerning the self-serving attribution bias in interpersonal influence situations. Journal of Personality, 48, 435-448.
- Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. Physical Review Letters, 49, 1804-1807.
- Aspect, A., Grangier, P., & Roger, G. (1981). Experimental tests of realistic local theories via Bell's theorem. Physical Review Letters, 47, 460-463.
- Aspect, A., Grangier, P., & Roger, G. (1982). Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A new violation of Bell's inequalities. Physical Review Letters, 49, 91-94.
- Atkinson, D. R., Furlong, 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.
- Baddeley, A. D. (1976). The psychology of memory. New York:

Basic Books.

- Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66, 423-437.
- Barcikowski, R. S. (1973). Optimum sample size and number of levels in a one-way random-effects analysis of variance. Journal of Experimental Education, 41(4), 10-16.
- Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. Journal of Educational Statistics, 6, 267-285.
- Bargmann, R. E. (1970). Interpretation and use of a generalized discriminant function. In R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, & K. J. C. Smith (Eds.), Essays in probability and statistics (pp. 35-60). Chapel Hill, NC: University of North Carolina Press.
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.
- Bartko, J. J. (1974). A note on the intraclass correlation as a measure of reliability. Psychological Reports, 34, 418.
- Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83,

762-765.

Barton, D. E., David, F. N., & O'Neill, A. F. (1960). Some properties of the distribution of the logarithm of non-central F. Biometrika, 47, 417-429.

Bassler, J., & Wolf, G. (1981). The importance of effect size: A contrasting view. Decision Sciences, 12, 142-149.

Beauchamp, K. L., & May, R. B. (1964). Replication report: Interpretation of levels of significance by psychological researchers. Psychological Reports, 14, 272.

Bennett, B. M., & Underwood, R. E. (1970). On McNemar's test for the 2 X 2 table and its power function. Biometrics, 26, 339-343.

Berk, A. A., & Chalmers, T. C. (1981). Cost and efficacy of the substitution of ambulatory for inpatient care. New England Journal of Medicine, 304, 393-397.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. Journal of Forecasting, 1, 257-269.

Blair, R. C., & Higgins, J. J. (1980a). A comparison of the power of Wilcoxin's rank-sum statistic to that of Student's t statistic under various nonnormal distributions. Journal of Educational Statistics, 5, 309-335.

Blair, R. C., & Higgins, J. J. (1980b). The power of t and

Wilcoxin statistics: A comparison. Evaluation Review, 4, 645-656.

Blair, R. C., Higgins, J. J., & Smitley, W. D. S. (1980). On the relative power of the U and t tests. British Journal of Mathematical and Statistical Psychology, 33, 114-120.

Bolles, R., & Messick, S. (1958). Statistical utility in experimental inference. Psychological Reports, 4, 223-227.

Bones, J. (1972). Statistical power analysis and geography. Professional Geographer, 24, p29-232.

Breen, L., & Gaito, J. (1970). Comments on Friedman's F_m procedure. Psychological Bulletin, 73, 309-310.

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

Brewer, J. K. (1974). Issues of power: Clarification. American Educational Research Journal, 11, 189-192.

Brewer, J. K., & Owen, P. W. (1973). A note on the power of statistical tests in the Journal of Educational Measurement. Journal of Educational Measurement, 10, 71-74.

Brown, A. S. (1976). Spontaneous recovery in human learning. Psychological Bulletin, 83, 321-338.

- Brown, D. J. (1975). Down with the linear model. American Educational Research Journal, 12, 491-505.
- Burger, J. M. (1981). Motivational biases in the attribution of responsibility for an accident: A meta-analysis of the defensive-attribution hypothesis. Psychological Bulletin, 90, 496-512.
- Burke, C. J. (1953). A brief note on one-tailed tests. Psychological Bulletin, 50, 384-387.
- Burke, C. J. (1954). Further remarks on one-tailed tests. Psychological Bulletin, 51, 587-590.
- Camp, C. J., & Maxwell, S. E. (1983). A comparison of various strength of association measures commonly used in gerontological research. Journal of Gerontology, 38, 3-7.
- Carroll, R. M., & Faden, V. B. (1978). Some sampling characteristics of three estimators of the intraclass correlation. Educational and Psychological Measurement, 38, 855-863.
- Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelley's ϵ^2 and Hays' ω^2 . Educational and Psychological Measurement, 35, 541-554.
- Carver, R. P. (1975). The Coleman Report: Using inappropriately designed achievement tests. American Educational Research Journal, 12, 77-86.
- Carver, R. P. (1978). The case against statistical

significance testing. Harvard Educational Review, 48, 378-399.

Charter, R. A. (1982). Practical formulas of strength of association measures. Educational and Psychological Measurement, 42, 969-974.

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

Chase, L. J., Chase, L. R., & Tucker, R. K. (1978). Statistical power in physical anthropology: A technical report. American Journal of Physical Anthropology, 49, 133-138.

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. Psychological Record, 26, 473-486.

Christensen, J. E., & Christensen, C. E. (1977). Statistical power analysis of health, physical education, and

recreation research. Research Quarterly, 48, 204-208.

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. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

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. (1973a). Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educational and Psychological Measurement, 33, 107-112.

Cohen, J. (1973b). Statistical power analysis and research results. American Educational Research Journal, 10, 225-230.

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

- Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Coleman, E. B. (1979). The Solzhenitsyn Finger Test: A significance test for spontaneous recovery. Psychological Bulletin, 86, 148-150.
- Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. Psychological Bulletin, 86, 662-679.
- Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 37, 131-146.
- Cooper, H., & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. Personality and Social Psychology Bulletin, 8, 168-173.
- Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442-449.
- Cowles, M. P. (1974). $N = 35$: A rule of thumb for psychological research. Perceptual and Motor Skills, 38,

1135-1138.

Cowles, M. P. (1977). Note on measures of association and why they are not reported. Perceptual and Motor Skills, 44, 799-802.

Cowles, M. P., & Davis, C. (1982a). Is the .05 level subjectively reasonable? Canadian Journal of Behavioral Science, 14, 248-252.

Cowles, M. P., & Davis, C. (1982b). On the origins of the .05 level of statistical significance. American Psychologist, 37, 553-558.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and ω^2 . Bulletin of the Psychonomic Society, 7, 280-282.

Cramer, E. M., & Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. Psychometrika, 44, 43-54.

Crane, J. A. (1976). The power of social intervention experiments to discriminate differences between experimental and control groups. Social Service Review, 50, 224-242.

Crane, J. A. (1977). Author's reply. Social Service Review, 51, 526-529.

Cranston, H. A., Boylan, C. W., Carroll, G. L., Sutura, S.

- P., Williamson, J. R., Gluzman, I. Y., & Krogstad, D. J. (1984). Plasmodium falciparum maturation abolishes physiologic red cell deformability. Science, 223, 400-403.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671-684.
- Crowder, R. G. (1976). Principles of learning and memory. Hillsdale, NJ: Lawrence Erlbaum.
- Curtis, E. W., & Alf, E. F. (1968). A correlational approach for measuring overlap of distributions and proportions of misclassification. Psychological Bulletin, 70, 626-630.
- Daly, J. A., & Hexamer, A. (1983). Statistical power in research in English Education. Research in the Teaching of English, 17, 157-164.
- Das Gupta, S., & Perlman, M. D. (1974). Power of the noncentral F-test: Effect of additional variates on Hotelling's T^2 -test. Journal of the American Statistical Association, 69, 174-180.
- Davis, L., Goldhaber, A. S., & Nieto, M. M. (1975). Limit on the photon mass deduced from Pioneer-10 observations of Jupiter's magnetic field. Physical Review Letters, 35, 1402-1405.
- Dawes, R. M., Landman, J., & Williams, M. (1984). Reply to

- Kurosawa. American Psychologist, 39, 74-75.
- Dayton, C. M., Schafer, W. D., & Rogers, B. G. (1973). On appropriate uses and interpretations of power analysis: A comment. American Educational Research Journal, 10, 231-234.
- Demorest, M. E. (1983, April). Bias and sampling error in the estimation of ω^2 . Paper presented at the 55th annual meeting of the Eastern Psychological Association, Philadelphia, PA.
- d'Espagnat, B. (1979). The quantum theory and reality. Scientific American, 241(5), 158-181.
- Dixon, W. J. (Ed.). (1981). BMDP statistical software: 1981. Berkeley, CA: University of California Press.
- Dixon, W. J., & Massey, F. J. (1983). Introduction to statistical analysis (4th ed.). New York: McGraw-Hill.
- Dodd, D. H., & Schultz, R. F. (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. Psychological Bulletin, 79, 391-395.
- Dooling, D. J., & Danks, J. H. (1975). Going beyond tests of significance: Is psychology ready? Bulletin of the Psychonomic Society, 5, 15-17.
- Dunnette, M. D. (1966). Fads, fashions, and folderol in psychology. American Psychologist, 21, 343-352.

- Dwyer, J. H. (1974). Analysis of variance and the magnitude of effects: A general approach. Psychological Bulletin, 81, 731-737.
- Edgington, E. S. (1964). A tabulation of inferential statistics used in psychology journals. American Psychologist, 19, 202-203.
- Edgington, E. S. (1974). A new tabulation of statistical procedures used in APA journals. American Psychologist, 29, 25-26.
- Edwards, A. L. (1979). Multiple regression and the analysis of variance and covariance. San Francisco: Freeman.
- Eisenhart, C. (1947). Inverse sine transformation of proportions. In C. Eisenhart, M. W. Hastay, & W. A. Wallis, (Eds.), Selected techniques of statistical analysis for scientific and industrial research production and management engineering (pp. 395-416). New York: McGraw-Hill.
- Endler, N. S. (1966). Estimating variance components from mean squares for random and mixed effects analysis of variance models. Perceptual and Motor Skills, 22, 559-570.
- Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests.

Psychological Review, 67, 269-271.

Fagley, N. S., & McKinney, I. J. (1983). Reviewer bias for statistically significant results: A reexamination.

Journal of Counseling Psychology, 30, 298-300.

Feigl, P. (1978). A graphical aid for determining sample size when comparing two independent proportions.

Biometrics, 34, 111-122.

Feldt, L. S. (1973). What size samples for methods/materials experiments? Journal of Educational Measurement, 10, 221-226.

Feldt, L. S., & Mahmoud, M. W. (1958). Power function charts for specification of sample size in analysis of variance. Psychometrika, 23, 201-210.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. Metron, 1, 3-32.

Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. Journal of the Royal Statistical Society, 85, 597-612.

Fisher, R. A. (1924). On a distribution yielding the error functions of several well known statistics. Proceedings of the International Congress of Mathematics, Toronto, 2, 805-813.

- Fisher, R. A. (1925). Statistical methods for research workers. London: Oliver & Boyd.
- Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. Proceedings of the Royal Society of London, Series A, 121, 654-673.
- Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). London: Oliver & Boyd.
- Fisher, R. A. (1946). Statistical methods for research workers (10th ed.). London: Oliver & Boyd.
- Fleishman, A. I. (1980). Confidence intervals for correlation ratios. Educational and Psychological Measurement, 40, 659-670.
- Fleiss, J. L. (1969). Estimating the magnitude of experimental effects. Psychological Bulletin, 72, 273-276.
- Fleiss, J. L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: Wiley.
- Fleiss, J. L., Tytun, A., & Ury, H. K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics, 36, 343-346.
- Franklin, A. D. (1981). What makes a 'good' experiment? British Journal of the Philosophy of Science, 32, 367-374.

- Freiman, J. A., Chalmers, T. C., Smith, H., & 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. New England Journal of Medicine, 299, 690-694.
- Freund, R. J., & Minton, P. D. (1979). Regression methods: A tool for data analysis. New York: Marcel Dekker.
- Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. Psychological Bulletin, 70, 245-251.
- Friedman, H. (1969). A simplified table for the estimation of magnitude of experimental effect. Psychonomic Science, 14, 193, 195.
- Friedman, H. (1982). Simplified determinations of statistical power, magnitude of effect and research sample sizes. Educational and Psychological Measurement, 42, 521-526.
- Gaebelin, J. W., Soderquist, D. R., & Powers, W. A. (1976). A note on variance explained in the mixed analysis of variance model. Psychological Bulletin, 83, 1110-1112.
- Gaito, J. (1958). The Bolles-Messick coefficient of utility. Psychological Reports, 4, 595-598.
- Gaito, J., & Firth, J. (1973). Procedures for estimating magnitude of effects. Journal of Psychology, 83, 151-161.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981).

Measurement theory for the behavioral sciences. San Francisco: Freeman.

Glass, G. V. (1976a). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Glass, G. V. (1976b, April). Primary, secondary, and meta-analysis of research. Presidential address of the American Educational Research Association, San Francisco.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-379.

Glass, G. V., & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. American Educational Research Journal, 6, 403-414.

Glass, G. V., McGaw, B., & Smith, M. L. (1981).

Meta-analysis in social research. Beverly Hills, CA: Sage.

Glass, G. V., & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Goldfried, M. R. (1959). One tailed tests and "unexpected" results. Psychological Review, 66, 79-80.

Goldhaber, A. S., & Nieto, M. M. (1976). The mass of the

photon. Scientific American, 234 (5), 86-91, 94-96.

Golding, S. L. (1975). Flies in the ointment: Methodological problems in the analysis of the percentage of variance due to persons and situations. Psychological Bulletin, 82, 278-288.

Green, B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. Annual Review of Psychology, 35, 37-53.

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

Gunst, R. F., & Mason, R. L. (1980). Regression analysis and its application: A data-oriented approach. New York: Marcel Dekker.

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

Haase, R. F. (1983). Classical and partial eta square in multifactor ANOVA designs. Educational and Psychological Measurement, 43, 35-39.

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.

Haggard, E. A. (1958). Intraclass correlation and the

analysis of variance. New York: Dryden Press.

- Halderson, J. S., & Glasnapp, D. R. (1972). Generalized rules for calculating the magnitude of an effect in factorial and repeated measures ANOVA designs. American Educational Research Journal, 9, 301-310.
- Hall, J. A. (1978). Gender effects in decoding nonverbal cues. Psychological Bulletin, 85, 845-857.
- Hamblin, R. L. (1971). Mathematical experimentation and sociological theory: A critical analysis. Sociometry, 34, 423-452.
- Han, C.-P., & Wang, J. L. G. (1983). Computation of noncentral F distributions with even denominator degrees of freedom. Communications in Statistics: Simulation and Computation, 12, 1-9.
- Harris, J. A. (1913). On the calculation of intraclass and interclass coefficients from class moments when the number of possible combinations is large. Biometrika, 9, 446-472.
- Harris, R. J. (1975). A primer of multivariate statistics. New York: Academic.
- Harvey, S. C., Prabhakaran, M., Mao, B., McCammon, J. A. (1984). Phenylalanine transfer RNA: Molecular dynamics simulation. Science, 223, 1189-1191.

- Haseman, J. K. (1978). Exact sample sizes for use with the Fisher-Irwin test for 2 X 2 tables. Biometrics, 34, 106-109.
- Hasenfeld, R. (1983). Empathy and justification: Two contextual cues related to gender differences in aggressive behavior. Unpublished manuscript, University of Rhode Island, Department of Psychology, Kingston.
- Hawkes, A. G. (1982). Approximating the normal tail. Statistician, 31, 231-236.
- Hays, W. L. (1963). Statistics. New York: Holt, Rinehart & Winston.
- Hays, W. L. (1973). Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart & Winston.
- Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart & Winston.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.
- Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490-499.
- Hedges, L. V. (1983). A random effects model for effect sizes. Psychological Bulletin, 93, 388-395.
- Hinkle, D. E., & Oliver, J. D. (1983). How large should the

sample be? A question with no simple answer? Or

Educational and Psychological Measurement, 43, 1051-1060.

Hogben, L. (1957). Statistical theory: The relationship of probability, credibility, and error. An examination of the contemporary crisis in statistical theory from a behaviourist viewpoint. London: Allen & Unwin.

Holmes, C. B. (1979). Sample size in psychological research. Perceptual and Motor Skills, 49, 283-288.

Holmes, C. B., Holmes, J. R., & Fanning, J. J. (1981). Sample size in non-APA journals. Journal of Psychology, 108, 263-266.

Hopkins, K. D. (1973). Research and analysis clinic: Preventing the number-one misinterpretation of behavioral research, or how to increase statistical power. Journal of Special Education, 7, 103-107.

Hornick, C. W., & Overall, J. E. (1980). Evaluation of three sample size formulae for 2 X 2 contingency tables. Journal of Educational Statistics, 5, 351-362.

Howell, D. C. (1982). Statistical methods for psychology. Boston: Duxbury.

Hoyer, W. J., Raskind, C. L., & Abrahams, J. P. (1984). Research practices in the psychology of aging: A survey of research published in the Journal of Gerontology,

1975-1982. Journal of Gerontology, 39, 44-48.

Huberty, C. J. (1983). Some univariate-multivariate generalizations. Educational and Psychological Measurement, 43, 705-721.

Huberty, C. J., & Holmes, S. E. (1983). Two-group comparisons and univariate classification. Educational and Psychological Measurement, 43, 15-26.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using ω^2 and d. American Psychologist, 36, 892-901.

Jaccard, J. (1983). Statistics for the behavioral sciences. Belmont, CA: Wadsworth.

Johnson, R. A., & Wichern, D. W. (1982). Applied multivariate statistical analysis. Englewood Cliffs, NJ: Prentice-Hall.

Jones, B. J., & Brewer, J. K. (1972). An analysis of the power of statistical tests reported in the Research Quarterly. Research Quarterly, 43, 23-30.

Jones, L. V. (1952). Tests of hypotheses: One-sided and two-sided alternatives. Psychological Bulletin, 49, 43-46.

- Jones, L. V. (1954). A rejoinder on one-tailed tests.
Psychological Bulletin, 51, 585-586.
- Kaiser, H. F. (1960). Directional statistical decisions.
Psychological Review, 67, 160-167.
- Katzer, J, & Sodt, J. (1973). An analysis of the use of
 statistical testing in communication research. The
Journal of Communication, 1973, 23, 251-265.
- Kelly, T. L. (1935). An unbiased correlation ratio measure.
Proceedings of the National Academy of Sciences, 21,
 554-559.
- Kennedy, J. J. (1970). The eta coefficient in complex ANOVA
 designs. Educational and Psychological Measurement, 30,
 885-889.
- Keppel, G. (1968). Retroactive and proactive inhibition. In
 T. R. Dixon & D. L. Horton (Eds.), Verbal behavior and
general behavior theory (pp. 172-213). Englewood Cliffs,
 NJ: Prentice-Hall.
- Keppel, G. (1982). Design and analysis: A researcher's
handbook (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Keppel, G., & Saufley, W. H. (1980). Introduction to design
and analysis: A student's handbook. San Francisco:
 Freeman.
- Keren, G., & Lewis, C. (1976). Nonorthogonal designs: Sample

versus population. Psychological Bulletin, 83, 817-826.

Keren, G., & Lewis, C. (1979). Partial omega squared for ANOVA designs. Educational and Psychological Measurement, 39, 119-128.

Kerlinger, F. N. (1964). Foundations of behavioral research. New York: Holt, Rinehart, & Winston.

Kerlinger, F. N. (1973). Foundations of behavioral research (2nd ed). New York: Holt, Rinehart, & Winston.

Keselman, H. J. (1975). A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. Canadian Psychological Review, 16, 44-48.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Belmont, CA: Wadsworth.

Klatzky, R. L. (1980). Human memory: Structures and processes (2nd ed.). San Francisco: Freeman.

Klein, T. W. (1974). Heritability and genetic correlation: Statistical power, population comparisons, and sample size. Behavior Genetics, 4, 171-189.

Koele, P. (1982). Calculating power in analysis of variance. Psychological Bulletin, 92, 513-516.

Koichi, I. (1962). A comparison of the powers of two multivariate analysis of variance tests. Eiometrika, 49, 455-462.

- Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. Journal of Educational Statistics, 8, 93-101.
- Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. Psychological Bulletin, 91, 404-412.
- Krauth, J. (1983). Nonparametric effect size estimation: A comment on Kraemer and Andrews. Psychological Bulletin, 94, 190-192.
- Kres, H. (1983). Statistical tables for multivariate analysis: A handbook with references to applications (P. Wadsack, Trans.). New York: Springer-Verlag. (Original work published 1975)
- Kroll, R. M., & Chase, L. J. (1975). Communication disorders: A power analytic assessment of recent research. Journal of Communication Disorders, 8, 237-247.
- Kuhn, T. S. (1970). The structure of scientific revolutions (2nd ed.). Chicago: University of Chicago Press.
- Kunce, J. T., Cook, D. W., & Miller, D. E. (1975). Random variables and correlational overkill. Educational and Psychological Measurement, 35, 529-534.
- Kurosawa, K. (1984). Meta-analysis and selective publication bias. American Psychologist, 39, 73-74.

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983).

Intraclass correlation: There's more there than meets the eye. Psychological Bulletin, 93, 586-595.

Landman, J. T., & Dawes, R. M. (1982). Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. American Psychologist, 37, 504-516.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. British Journal of Mathematical and Statistical Psychology, 31, 107-112.

LaTour, S. A. (1981a). Effect size estimation: A commentary on Wolf and Bassler. Decision Sciences, 12, 136-141.

LaTour, S. A. (1981b). Variance explained: It measures neither importance nor effect size. Decision Sciences, 12, 150-160.

Laubscher, N. F. (1960). Normalizing the noncentral t and F distributions. Annals of Mathematical Statistics, 31, 1105-1112.

Lehmer, E. (1944). Inverse tables of probabilities of errors of the second kind. Annals of Mathematical Statistics, 15, 388-398.

Lenzen, V. F. (1938). Procedures of empirical science. Chicago: University of Chicago Press.

Levenson, R. L. (1980). Statistical power analysis:

Implications for researchers, planners, and practitioners in Gerontology. Gerontologist, 20, 494-498.

Levin, J. R. (1967). Misinterpreting the significance of "explained variation." American Psychologist, 22, 675-676.

Levine, D. W., & Dunlap, W. P. (1982). Power of the F test with skewed data: Should one transform or not? Psychological Bulletin, 92, 272-280.

Levy, P. (1967). Substantive significance of significant differences between two groups. Psychological Bulletin, 67, 37-40.

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. Psychonomic Science, 9, 563-564.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). Introduction to bivariate and multivariate analysis. Glenview, IL: Scott, Foresman.

Lindman, H. R. (1974). Analysis of variance in complex experimental designs. San Francisco: Freeman.

Lindquist, E. F. (1953). Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin.

Lipsey, M. W. (1983). A scheme for assessing measurement sensitivity in program evaluation and other applied

research. Psychological Bulletin, 94, 152-165.

Lunneborg, C. E., & Abbott, R. D. (1983). Elementary multivariate analysis for the behavioral sciences: Applications of basic structure. New York: North-Holland.

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

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.

Mace, A. E. (1964). Sample-size determination. New York: Reinhold.

Marascuilo, L. A., & Levin, J. R. (1983). Multivariate statistics in the social sciences: A researcher's guide. Monterey, CA: Brooks/Cole.

Maxwell, S. E., Camp, C. J., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66, 525-534.

McNemar, Q. (1960). At random: Sense and nonsense. American Psychologist, 15, 295-300.

McNemar, Q. (1962). Psychological statistics (3rd ed.). New York: Wiley.

Mead, R., Bancroft, T. A., & Han, C.-P. (1975). Power of analysis of variance test procedures for incompletely specified fixed models. Annals of Statistics, 3, 797-808.

- Medler, J. F., Schneider, P. R., & Schneider, A. L. (1981). Statistical power analysis and experimental field research. Evaluation Review, 5, 834-850.
- Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.
- Meyer, D. L. (1974a). Issues of power: Rejoinder. American Educational Research Journal, 11, 193-194.
- Meyer, D. L. (1974b). Statistical tests and surveys of power: A criticism. American Educational Research Journal, 11, 179-188.
- Millman, J., & Cieslak, P. (1970). Letter to the editor. American Statistician, 24(4), 49.
- Minium, E. W., & Clarke, R. B. (1982). Elements of statistical reasoning. New York: Wiley & Sons.
- Mitchell, C., & Hartmann, D. P. (1981). A cautionary note on the use of omega squared to evaluate the effectiveness of behavioral treatments. Behavioral Assessment, 3, 93-100.
- Moore, M. (1975). The use of statistical techniques by

psychological journal articles. American Statistician, 29, 68-69.

Moore, M. (1981). Empirical statistics: III. Changes in use of statistics in psychology. Psychological Reports, 48, 751-753.

Morrison, D. E., & Henkel, R. E. (1970). The significance test controversy: A reader. Chicago: Aldine.

Morrison, D. F. (1983). Applied linear statistical methods. Englewood Cliffs, NJ: Prentice-Hall.

Mosteller, F., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), Handbook of social psychology: Vol. 1. Theory and method (pp. 289-334). Reading, MA: Addison-Wesley.

Muchinsky, P. M. (1979). Some changes in the characteristics of articles published in the Journal of Applied Psychology over the past 20 years. Journal of Applied Psychology, 64, 455-459.

Neyman, J. (1937). 'Smooth' test for goodness of fit. Skandinavisk Aktuarietidskrift, 20, 149-199.

Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika, 20A, 175-240.

Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of

statistical inference: Part II. Biometrika, 20A, 263-294.

Neyman, J., & Pearson, E. S. (1931). On the problem of k samples. Bulletin of the Polish Academy of Science and Letters, Series A, 460-481.

Neyman, J., & Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London, Series A, 231, 289-337.

Neyman, J., & Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. Proceedings of the Cambridge Philosophical Society, 29, 492-510.

Neyman, J., & Pearson, E. S. (1936a). Contributions to the theory of testing statistical hypotheses. Statistical Research Memoirs, 1, 1-37.

Neyman, J., & Pearson, E. S. (1936b). Sufficient statistics and uniformly most powerful tests of statistical hypotheses. Statistical Research Memoirs, 1, 113-137.

Neyman, J., & Pearson, E. S. (1938). Contributions to the theory of testing statistical hypotheses. Statistical Research Memoirs, 2, 25-57.

Neyman, J., & Pearson, E. S. (1967). Joint statistical papers. Berkeley, CA: University of California Press.

- Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests: Some new results. Psychological Bulletin, 94, 524-533.
- Norton, D. W., & Lindquist, E. F. (1951). Applications of experimental design and analysis. Review of Educational Research, 21, 350-367.
- Norton, V. (1983). A simple algorithm for computing the non-central F distribution. Journal of the Royal Statistical Society, Series C, (Applied Statistics), 32, 84-85.
- Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.
- Odeh, R. E., & Fox, M. (1975). Sample size choice: Charts for experiments with linear models. New York: Marcel Dekker.
- O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. Psychological Bulletin, 92, 766-777.
- Orwin, R. G. (1983). A fail safe N for effect size in meta-analysis. Journal of Educational Statistics, 8, 157-159.
- Overall, J. E. (1969). Classical statistical hypothesis testing within the context of Bayesian theory. Psychological Bulletin, 71, 285-292.

Overall, J. E., & Dalal, S. N. (1965). Design of experiments to maximize power relative to cost. Psychological Bulletin, 64, 339-350.

Overall, J. E., Hollister, L. E., & Dalal, S. N. (1967). Psychiatric drug research: Sample size requirements for one vs two raters. Archives of General Psychiatry, 16, 152-161.

Overall, J. E., & Hornick, C. W. (1982). An evaluation of power and sample-size requirements for the continuity-corrected Fisher exact test. Perceptual and Motor Skills, 54, 83-86.

Patnaik, P. B. (1949). The non-central χ^2 - and F-distributions and their applications. Biometrika, 36, 203-232.

Pearson, E. S. (1960). Editorial note. Biometrika, 47, 430-431.

Pearson, E. S. (1970). The Neyman-Pearson story: 1926-34: Historical sidelights on an episode in Anglo-Polish collaboration. In E. S. Pearson & M. G. Kendall (Eds.), Studies in the history of statistics and probability (pp. 455-477). London: Griffin.

Pearson, E. S., & Hartley, H. O. (1951). Charts of the power function for analysis of variance tests, derived from the

non-central F-distribution. Biometrika, 38, 112-130.

Pearson, K. (1911). On a correction to be made to the correlation ratio η . Biometrika, 8, 254-256.

Pedhazur, E. J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart, & Winston.

Pennick, J. E., & Brewer, J. K. (1972). The power of statistical tests in science teaching research. Journal of Research in Science Teaching, 9, 377-381.

Peters, C. C., & van Voorhis, W. R. (1940). Statistical procedures and their mathematical bases. New York: McGraw-Hill.

Pipkin, F. M., & Ritter, R. C. (1983). Precision measurements and fundamental constants. Science, 219, 913-921.

Platt, J. R. (1964). Strong inference. Science, 146, 347-353.

Plutchik, R. (1983). Foundations of experimental research (3rd ed.). New York: Harper & Row.

Poland, R. E., Rubin, R. T., & Weichsel, M. E. (1980). Circadian patterns of rat anterior pituitary and target gland hormones in serum: Determination of the appropriate sample size by statistical power analysis. Psychoneuroendocrinology, 5, 209-224.

Popper, K. R. (1959). The logic of scientific discovery. New

York: Basic Books.

- Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. (1978). Practical significance in program evaluation. American Educational Research Journal, 15, 529-539.
- Postman, L., Stark, K., & Fraser, J. (1968). Temporal changes in interference. Journal of Verbal Learning and Verbal Behavior, 7, 672-694.
- Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. Memory & Cognition, 1, 19-40.
- Preece, P. F. W. (1983). A measure of experimental effect size based on success rates. Educational and Psychological Measurement, 43, 763-766.
- Quinn, J. F. (1983). Mass extinctions in the fossil record. Science, 219, 1239-1240.
- Ramsey, P. H. (1982). Empirical power of procedures for comparing two groups on p variables. Journal of Educational Statistics, 7, 139-156.
- Ranney, G. B., & Thigpen, C. C. (1981). The sample coefficient of determination in simple linear regression. American Statistician, 35, 152-153.
- Raup, D. M., & Sepkoski, J. J. (1982). Mass extinctions in the marine fossil record. Science, 215, 1501-1503.
- Ray, A. A. (Ed.). (1982). SAS user's guide: Statistics, 1982

edition. Cary, NC: SAS Institute.

Reid, C. (1982). Neyman--from life. New York: Springer-Verlag.

Robinson, A. L. (1982). Quantum mechanics passes another test. Science, 217, 435-436.

Robinson, A. L. (1983). Loophole closed in quantum mechanics. Science, 219, 40-41.

Rohrlich, F. (1983). Facing quantum mechanical reality. Science, 221, 1251-1255.

Rosenthal, R. (1966). Experimenter effects in behavioral research. New York: Appleton-Century-Crofts.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

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

Rosenthal, R. (Ed.). (1980a). Quantitative assessment of research domains: New directions for methodology of social and behavioral science (Number 5). San Francisco: Jossey-Bass.

Rosenthal, R. (1980b). Summarizing significance levels. In R. Rosenthal (Ed.), Quantitative assessment of research domains: New directions for methodology of social and behavioral science (Number 5) (pp. 33-46). San Francisco:

Jossey-Bass.

Rosenthal, R. (1982). Valid Interpretations of quantitative research results. In D. Brinberg & L. H. Kidder (Eds.), Forms of validity in research: New directions for methodology of social and behavioral science (Number 12) (pp. 59-75). San Francisco: Jossey-Bass.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. Journal of Consulting and Clinical Psychology, 51, 4-13.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. Journal of Psychology, 55, 33-38.

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. Psychological Reports, 15, 570.

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. Journal of Applied Social Psychology, 9, 395-396.

Rosenthal, R., & Rubin, D. B. (1982a). Comparing effect sizes of independent studies. Psychological Bulletin, 92, 500-504.

Rosenthal, R., & Rubin, D. B. (1982b). Further meta-analytic

procedures for assessing cognitive gender differences.

Journal of Educational Psychology, 74, 708-712.

Rosenthal, R., & Rubin, D. B. (1982c). A simple, general purpose display of magnitude of experimental effect.

Journal of Educational Psychology, 74, 166-169.

Roseveare, N. T. (1982). Mercury's perihelion: From LeVerrier to Einstein. New York: Oxford.

Rossi, J. S. (1977). On the generality of the laws of unlearning: Spontaneous recovery? Unpublished manuscript, University of Rhode Island, Department of Psychology, Kingston.

Rossi, J. S. (1983). Ratios exaggerate gender differences in mathematical ability. American Psychologist, 38, 348.

Rotton, J., & Schonemann, P. H. (1978). Power tables for analysis of variance. Educational and Psychological Measurement, 38, 213-229.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

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.

Runyon, R. (1969). Minimum sample size required to achieve power $1 - \beta$ in testing for the significance of a difference in independent proportions. Perceptual and

Motor Skills, 28, 247-250.

Sacco, W. P. (1982). Statistical power considerations in the use of cost-effectiveness analysis. Professional Psychology, 13, 752-758.

Sachdeva, D. (1973). Estimating strength of relationship in multivariate analysis of variance. Educational and Psychological Measurement, 33, 627-631.

Sagan, C. (1980). Cosmos. New York: Random House.

Sawyer, A. G. (1982). Statistical power and effect size in consumer research. In A. A. Mitchell (Ed.), Advances in consumer research: Vol. 9. Ann Arbor: Association for Consumer Research.

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

Schuerman, J. R. (1977). Comments on statistical power in evaluative research design. Social Service Review, 51, 524-526.

Schwartz, S., & Dalglish, L. (1982). Statistical inference in personality research. Journal of Research in Personality, 16, 290-302.

Sechrest, L., & Yeaton, W. H. (1981). Empirical bases for estimating effect size. In R. F. Boruch, P. M. Wortman, &

D. S. Cordray (Eds.), Reanalyzing program evaluations: Policies and practices for secondary analysis of social and educational programs (pp. 212-224). San Francisco: Jossey-Bass.

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. Evaluation Review, 6, 579-600.

Selvin, H. C., & Stuart, A. (1966). Data-dredging procedures in survey research. The American Statistician, 20(3), 20-23.

Serlin, R. C. (1982). A multivariate measure of association based on the Pillai-Bartlett procedure. Psychological Bulletin, 91, 413-417.

Serlin, R. C., Carr, J., & Marascuilo, L. A. (1982). A measure of association for selected nonparametric procedures. Psychological Bulletin, 92, 786-790.

Severo, N. C., & Zelen, M. (1960). Normal approximation to the chi-square and non-central F probability functions. Biometrika, 47, 411-416.

Shaffer, J. P., & Gillo, M. W. (1974). A multivariate extension of the correlation ratio. Educational and Psychological Measurement, 34, 521-524.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.

- Shine, L. C. (1980). The fallacy of replacing an a priori significance level with an a posteriori significance level. Educational and Psychological Measurement, 40, 331-335.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.
- Simpson, R. H. (1944). The specific meanings of terms indicating differing degrees of frequency. Quarterly Journal of Speech, 30, 328-330.
- Simpson, R. H. (1963). Stability in meaning for quantitative terms. A comparison over 20 years. Quarterly Journal of Speech, 49, 146-151.
- Smart, R. (1964). The importance of negative results in psychological research. Canadian Psychologist, 5a, 225-232.
- Smith, M. L. (1980). Publication bias and meta-analysis. Evaluation in Education: An International Review Series, 4, 22-24.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.
- Spreitzer, E., & Chase, L. J. (1974). Statistical power in

- sociological research: An examination of data-analytic strategies. Unpublished manuscript, Bowling Green State University, Department of Sociology, Bowling Green, OH.
- SPSSX user's guide. (1983). New York: McGraw-Hill.
- Stavig, G. R., & Acock, A. C. (1980). Coefficients of association analogous to Pearson's r for nonparametric statistics. Educational and Psychological Measurement, 40, 679-685.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance--or vice versa. Journal of the American Statistical Association, 54, 30-34.
- Stevens, J. P. (1980). Power of the multivariate analysis of variance tests. Psychological Bulletin, 88, 728-737.
- Strahan, R. P. (1982). Assessing magnitude of effect from rank-order correlation coefficients. Educational and Psychological Measurement, 42, 763-765.
- Strahan, R. P., & McGovern, P. G. (1979). What's in a number? Verbal equivalents of numerical p-values. Proceedings of the Section on Statistical Education of the American Statistical Association, 91-94.
- Subkoviak, M. J., & Levin, J. R. (1977). Fallibility of measurement and the power of a statistical test. Journal of Educational Measurement, 14, 47-52.

Summary report of journal operations: 1982. (1983). American Psychologist, 38, 739.

Susskind, E. C., & Howland, E. W. (1980). Measuring effect magnitude in repeated measures ANOVA designs: Implications for gerontological research. Journal of Gerontology, 35, 867-876.

Symonds, P. M. (1930). A comparison of statistical measures of overlapping with charts for estimating the value of bi-serial r . Journal of Educational Psychology, 21, 586-596.

Tang, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. Statistical Research Memoirs, 2, 126-149.

Taylor, F. B. (1959). Development of the testing of statistical hypotheses. Unpublished doctoral dissertation, Columbia University, New York.

Tiku, M. L. (1965). Laguerre series forms of non-central X^2 and F distributions. Biometrika, 52, 415-427.

Tiku, M. L. (1966). A note on approximating to the non-central F distribution. Biometrika, 53, 606-610.

Tiku, M. L. (1967). Tables of the power of the F -test. Journal of the American Statistical Association, 62, 525-539.

- Tilton, J. W. (1937). The measurement of overlapping. Journal of Educational Psychology, 28, 656-662.
- Trattner, M. H., & O'Leary, B. S. (1980). Sample sizes for specified statistical power in testing for differential validity. Journal of Applied Psychology, 65, 127-134.
- Tullock, G. (1959). Publication decisions and tests of significance--a comment. Journal of the American Statistical Association, 54, 593.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.
- Underwood, B. J. (1948a). Retroactive and proactive inhibition after five and forty-eight hours. Journal of Experimental Psychology, 38, 29-38.
- Underwood, B. J. (1948b). "Spontaneous recovery" of verbal associations. Journal of Experimental Psychology, 38, 429-439.
- Underwood, B. J., & Postman, L. (1960). Extra-experimental sources of interference in forgetting. Psychological Review, 67, 73-95.
- Ury, H. K. (1981). Continuity-corrected approximations to sample size or power when comparing two proportions: Chi-squared or arc sine? Statistician, 30, 199-203.
- Ury, H. K. (1982). Comparing two proportions: Finding p_2 when p_1 , n , α , and β are specified. Statistician, 31,

245-250.

- Ury, H. K., & Fleiss, J. L. (1980). On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. Biometrics, 36, 347-351.
- Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. Psychological Bulletin, 72, 204-213.
- Venables, W. (1975). Calculation of confidence intervals for noncentrality parameters. Journal of the Royal Statistical Society, Series B, 37, 406-412.
- Walster, G. W., & Cleary, T. A. (1970). A proposal for a new editorial policy in the social sciences. American Statistician, 24(2), 16-19.
- Wampold, B. E., Furlong, M. J., & Atkinson, D. R. (1983). Statistical significance, power, and effect size: A response to the reexamination of reviewer bias. Journal of Counseling Psychology, 30, 459-463.
- Weinberg, S. (1981). The decay of the proton. Scientific American, 244(6), 64-75.
- Welkowitz, J., Ewen, R. B., & Cohen, J. (1982) Introductory statistics for the behavioral sciences (3rd ed.). New York: Academic Press.
- Westermann, R., & Hager, W. (1983). The relative importance

of low significance level and high power in multiple tests of significance. Perceptual and Motor Skills, 56, 407-413.

Whaley, F. S., Quade, D., & Haley, R. W. (1980). Effects of method error on the power of a statistical test: Implications of imperfect sensitivity and specificity in retrospective chart review. American Journal of Epidemiology, 111, 534-542.

Whimbey, A., Vaughan, G. M., & Tatsunoka, M. M. (1967). Fixed effects vs random effects: Estimating variance components from mean squares. Perceptual and Motor Skills, 25, 668.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. American Educational Research Journal, 19, 249-258.

Wilson, W., Miller, H. L., & Lower, J. S. (1967). Much ado about the null hypothesis. Psychological Bulletin, 67, 188-196.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Wishart, J. (1932). A note on the distribution of the correlation ratio. Biometrika, 24, 441-456.

Wolf, G., & Bassler, J. (1980). Importance of hypotheses: Group comparison analysis. Decision Sciences, 11, 27-41.

Woolley, T. W., & Dawson, G. O. (1983). A follow-up power

analysis of the statistical tests used in the Journal of Research in Science Teaching. Journal of Research in Science Teaching, 20, 673-681.

APPENDIX

Studies Included in the Meta-Analysis

- Abra, J. C. (1967). Time changes in the strength of forward and backward associations. Journal of Verbal Learning and Verbal Behavioral, 6, 640-645.
- Abra, J. C. (1968). Time changes in the strength of extinguished context and specific associations. Journal of Experimental Psychology, 77, 684-686.
- Abra, J. C. (1969). List-1 unlearning and recovery as a function of the point of interpolated learning. Journal of Verbal Learning and Verbal Behavior, 8, 494-500.
- Adams, S. (1962). Temporal changes in the strength of competing verbal associates. Unpublished doctoral dissertation, University of California, Berkeley. (Cited in Brown, 1976).
- Archer, E. J., & Underwood, B. J. (1951). Retroactive inhibition of verbal associations as a multiple function of temporal point of interpolation and degree of interpolated learning. Journal of Experimental Psychology, 42, 283-290.
- Birnbaum, I. M. (1965). Long-term retention of first-list associations in the A-B, A-C paradigm. Journal of Verbal Learning and Verbal Behavior, 4, 515-520.
- Briggs, G. E. (1954). Acquisition, extinction, and recovery

functions in retroactive inhibition. Journal of Experimental Psychology, 47, 285-293.

Briggs, G. E., Thompson, R. F., & Brogden, W. J. (1954).

Retention functions and reproductive inhibition. Journal of Experimental Psychology, 48, 419-423.

Brown, A. S. (1974). Spontaneous recovery in a within-list A-B, A-D design. Dissertation Abstracts International, 35, 3050B-3051B.

Ceraso, J., & Henderson, A. (1965). Unavailability and associative loss in RI and PI. Journal of Experimental Psychology, 70, 300-303.

Ceraso, J., & Henderson, A. (1966). Unavailability and associative loss in RI and PI: Second try. Journal of Experimental Psychology, 72, 314-316.

Ceraso, J., Schiffman, D., & Becker, B. (1965). Recall interference in retroactive inhibition. Journal of Psychology, 59, 259-265.

Dallett, K. M. (1964). Proactive and retroactive inhibition in the A-B, A-B' paradigm. Journal of Experimental Psychology, 68, 190-200.

Forrester, W. E. (1970). Retroactive inhibition and spontaneous recovery in the A-B, D-C paradigm. Journal of Verbal Learning and Verbal Behavior, 9, 525-528.

- Goggin, J. (1966). Retroactive and proactive inhibition in the short-term retention of paired associates. Journal of Verbal Learning and Verbal Behavior, 5, 526-535.
- Houston, J. P. (1966). First-list retention and time and method of recall. Journal of Experimental Psychology, 71, 839-843.
- Houston, J. P. (1967). Unlearning of specific associations in the A-B, C-D paradigm. Journal of Experimental Psychology, 74, 254-258.
- Houston, J. P. (1968). Unlearning and spontaneous recovery in short-term memory. Journal of Verbal Learning and Verbal Behavior, 7, 251-253.
- Howe, T. S. (1967). Unlearning and competition in List-1 recall. Journal of Experimental Psychology, 75, 559-565.
- Ishihara, I. (1951). The process of retroactive inhibition in retention. Japanese Journal of Psychology, 75, 559-565. (Cited in Brown, 1976).
- Jones, W. P. (1953). A comparison of retroactive and proactive inhibition as functions of the time interval between original learning and the measurement of retention. Dissertation Abstracts, 13, 1249.
- Kammen, R., & Melton, A. W. (1967). Absolute recovery of first-list responses from unlearning during 26 minutes filled with an easy or difficult information processing

task. Proceedings of the 75th Annual Convention of the American Psychological Association, 2, 63-64.

Koppelaar, R. J. (1963). Time changes in strength of A-B, A-C lists: Spontaneous recovery? Journal of Verbal Learning and Verbal Behavior, 2, 310-319.

Lehr, D. J., & Duncan, C. P. (1970). Effects of priming on spontaneous recovery of verbal lists. Journal of Verbal Learning and Verbal Behavior, 9, 106-110.

Lehr, D. J., Frank, R. C., & Mattison, D. W. (1972). Retroactive inhibition, spontaneous recovery, and type of interpolated learning. Journal of Experimental Psychology, 92, 232-236.

Lehr, D. J., & Netti, J. F. (1973). Retroactive inhibition and spontaneous recovery in a contingent verbal discrimination task. Psychological Reports, 32, 443-450.

Martin, E., & Mackay, S. A. (1970). A test of the list-differentiation hypothesis. American Journal of Psychology, 83, 311-321.

Newton, J. M., & Wickens, D. D. (1956). Retroactive inhibition as a function of the temporal position of interpolated learning. Journal of Experimental Psychology, 51, 149-154.

Postman, L., & Gray, W. D. (1978). Response recall and

retroactive inhibition. American Journal of Psychology, 91, 3-22.

Postman, L., Stark, K., & Fraser, J. (1968). Temporal changes interference. Journal of Verbal Learning and Verbal Behavior, 7, 672-694.

Postman, L., Stark, K., & Henschel, D. (1969). Conditions of recovery after unlearning. Journal of Experimental Psychology Monograph, 82 (1, Pt. 2).

Rothkopf, E. Z. (1957). A deduction from an excitation-inhibition account of retroactive inhibition. Journal of Experimental Psychology, 53, 207-213.

Saltz, E., & Hamilton, H. (1967). Spontaneous recovery of list 1 responses in the A-B, A'-C paradigm. Journal of Experimental Psychology, 75, 267-273.

Shulman, H. G., & Martin, E. (1970). Effects of response-set similarity on unlearning and spontaneous recovery. Journal of Experimental Psychology, 86, 230-235.

Silverstein, A. (1967). Unlearning, spontaneous recovery, and the partial-reinforcement effect in paired-associate learning. Journal of Experimental Psychology, 73, 15-21.

Silverstein, A., & Rossi, J. S. (1979, April). Retroactive interference at the level of single associations. Paper presented at the 50th Annual Meeting of the Eastern

Psychological Association, Philadelphia, PA.

Slamecka, N. J. (1966). A search for spontaneous recovery of verbal associations. Journal of Verbal Learning and Verbal Behavior, 5, 205-207.

Underwood, B. J. (1948). Retroactive and proactive inhibition after five and forty-eight hours. Journal of Experimental Psychology, 38, 29-38.

Underwood, B. J. (1948). "Spontaneous recovery" of verbal associations. Journal of Experimental Psychology, 38, 429-439.

Wickens, D. D., & Dalezman, J. (1974). Spontaneous recovery and clustering of first-list responses. Journal of Experimental Psychology, 103, 1067-1073.