INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road. Ann Arbor. MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9106869

Statistical power analysis of doctoral dissertation research in educational psychology

McKean, Kathleen Elizabeth, Ph.D.

Oklahoma State University, 1990

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

STATISTICAL POWER ANALYSIS OF DOCTORAL DISSERTATION RESEARCH IN EDUCATIONAL PSYCHOLOGY

By

KATHLEEN E. McKEAN

Bachelor of Arts
Oklahoma State University
Stillwater, Oklahoma
1976

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1979

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 1990

STATISTICAL POWER ANALYSIS OF DOCTORAL

DISSERTATION RESEARCH IN

EDUCATIONAL PSYCHOLOGY

Thesis Approved:

Norma J. Campbell
Thesis Adviser

Charles R Davis

Norman N. Durham
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to especially thank Dr. Jo Campbell for her guidance and enthusiastic support in serving as chairperson of my doctoral committee and dissertation advisor. My sincere gratitude also goes to my committee members, Dr. Dale Fuqua, Dr. Bob Davis, and Dr. Chuck Edgley for their knowledge and support.

To Dr. Sherry Maxwell for her encouragement, support, and caring for the rest of my family (Muffin, Skoshi, Mop, Dinger, and Gizmo) while I worked on this project goes my deepest appreciation.

Sincere gratitude is also extended to Mac and St. Jode, as well as Julie, Kevin, and Nancy. I also need to thank Dr. Suzie Alexander, who managed to survive a tornado, finish her own dissertation and fetch for mine.

Finally I would like to give my thanks to my sister, Susan, who always accepted me.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Most behavioral and educational research is built upon significance testing. To investigate a problem, the researcher converts hypotheses into ". . . procedures which will yield a test of significance; and he will characteristically allow the result of the test of significance to bear the essential responsibility for the conclusions which he will draw" (Bakan, 1966, p. 423). The accuracy of statistical inferences in behavioral and educational research (statistical conclusion validity) is thus a consideration of some importance.

Significance testing relies upon the laws of chance; whenever investigators draw conclusions based upon results of significance tests, there is a chance that their conclusions are in error. Two types of errors are pertinent--incorrectly concluding that a phenomenon exists (Type I error), and incorrectly concluding that a phenomenon does not exist (Type II error). As early as 1962, it was noted that Type I errors were meticulously attended to, while Type II errors were rarely referred to in the literature (Cohen, 1962). The situation has not changed. Recent surveys indicate that very few published studies make reference to the Type II error or its correlate--the power of the test of significance.

The power of a significance test is the probability of detecting a difference, or treatment effect, in the population when it exists (Cohen, 1988). It is, in effect, the probability that researchers allow themselves to verify the existence of effects which they hypothesize.

Background of the Study

The concepts of Type II error and statistical power were introduced by Neyman and Pearson in 1928, and researchers are expected to be knowledgeable regarding the concepts and their applications. Nevertheless, researchers appear to give them insufficient consideration in the design and reporting of experimental studies (Brewer, 1972; Cohen & Hyman, 1979; Hopkins, Coulter, & Hopkins, 1981; McFatter & Gollob, 1986; Rogers & Hopkins, 1988).

The probability of making a Type I error is routinely set at .05 or lower in behavioral and educational studies, and the "significance level" of test results is almost universally reported. The probability of making a Type II error is rarely reported and presumably rarely determined, either in advance of an investigation or after its completion. Literature surveys have demonstrated a "shockingly high" error rate of over 50% (Lipsey, 1990). Many behavioral studies fail to discover significant differences among sample means ". . . even when differences among corresponding population means are substantial" (Rogers & Hopkins, 1988, p. 647).

The first analysis of the power of significance testing was reported by Cohen (1962). Power analyses for published literature have been conducted in a variety of behavioral and educational fields: psychology--including counseling psychology, applied psychology, abnormal psychology, personnel selection, occupational therapy, speech and hearing, mass communications; education--including general education, science education, English education, physical education, counselor education, social work education, medical education, educational measurement; social work and social intervention research; and other fields such as market research, medicine, and geography (Cohen, 1988; Lipsey, 1990).

The published literature, however, is considered a biased sample of research in general. A publication bias toward significant results has been well-documented (Bakan, 1966; Lipsey, 1990). Due to this bias, the power of statistical tests in published studies is likely to be higher than the average power of most educational and behavioral research (Chase & Chase, 1976).

Educational and psychological journals receive many more articles than they can possibly publish. It has historically been the practice of many of these journals to use the level of significance reported as a criterion for acceptance for publication (McNemar, 1960; Mellon, 1962; Sterling, 1959). Smith (1980) reviewed meta-analyses which

included both published and unpublished research. In every comparison, average experimental effects from published studies were larger than those from unpublished research. The published effect sizes were one-third of a standard deviation higher than findings reported in theses. She concluded that, ". . . failing to represent unpublished studies in a meta-analysis may produce misleading generalizations (sic)" (p. 24). The same findings would be expected for power surveys, since power is a function of effect size. The larger effect sizes in published research would yield higher levels of statistical power.

It is not possible to determine the number of research reports which either lead to the rejection of the null or do not, or how many of each are submitted for publication. However, it has been established that the scientific community is not equally aware of all experimental results. If an investigation results in a rejection of the null hypothesis, the study is more likely to be submitted and published and less likely to be replicated (Sterling, 1959). "A generation of researchers could profitably be employed in repeating interesting studies which originally used inadequate sample sizes. Unfortunately, the ones most needing such repetition are least likely to have appeared in print" (Cohen, 1962, p. 153).

The present investigation describes the levels of statistical power in doctoral dissertation research in

educational psychology. Dissertations, as unpublished studies, may be more representative of research in general than the published literature.

Statement of the Problem

The problem to be investigated was: What is the level of power in dissertation research in selected educational and behavioral science Ph.D. degree programs? Due to the large numbers of degrees granted across the United States in any given year, the scope of the study was narrowed to doctoral theses completed in 1988 that were classified by Dissertation Abstracts International as "Educational Psychology."

Purpose of the Study

The purpose of the study was to investigate the level of power in contemporary dissertation research in educational and behavioral fields. A related purpose was to determine whether these studies, conducted under the supervision of graduate faculty members, were undertaken with consideration of Type II error and the level of statistical power.

Significance of the study

The available evidence indicates that researchers do not understand or employ the concept of statistical power. In research reports where power is a clearly relevant issue (e.g., interpretation of nonsignificant results), it is not addressed (Cohen, 1988). "If we take as evidence the

research literature, we find that statistical power is only infrequently understood and almost never determined" (Cohen, 1977, p. 1). The topic was given no attention in applied statistical texts until the 1960's (Chase & Tucker, 1976), and received only slight attention until quite recently (Cohen, 1988).

Investigators cannot draw accurate conclusions about hypotheses without adequate power. Inferences concerning the presence or absence of an effect are precluded unless the probability of making an error is known or estimated. Judd and Kenny (1981) defined statistical conclusion validity as ". . . the extent to which the research is sufficiently precise or powerful enough to enable us to detect treatment effects" (p. 29). They stressed the issues of power and Type II error for two reasons: (1) most researchers are well-trained with regard to Type I errors, and (2) Type II conclusion errors are of great importance in applied research settings. They concluded that,

> "All too often in the last 20 years, evaluations of education, rehabilitation, and social welfare programs conclude that these programs have little effect . . . Given the expense of putting together and administering these social welfare programs, it is crucial that any effects they engender be detected" (p. 29).

Estimation of power is essential to interpreting nonsignificant results. If an investigation results in failure to reject the null hypothesis and power is adequate, valid interpretations of "no effect" or "no relationship" can be made. Without knowledge of the statistical power to detect an effect, however, no interpretation is logically possible (Chase & Tucker, 1975; Tversky & Kahnman, 1971). Increased awareness of statistical power may lead to more serious interpretations of null results (Greenwald, 1975).

For example, the field of special education is currently undergoing scrutiny because of a body of research indicating that special education "pull-out" programs have little effect on increased academic achievement. The U.S. Department of Education has sponsored a "Regular Education Initiative" in response to the "negative" results of efficacy research. Since this is, in essence, an acceptance of the null hypothesis across a number of studies, the power of the research to detect effects should be investigated before it is concluded that current programs are ineffective.

Studies designed without consideration of statistical power are wasteful of research effort. Too often, research investigations are undertaken which have little chance of success (Cohen, 1962). Given the publication bias toward significant results, low power (and the resulting nonsignificant findings) may prevent important research from

appearing in the behavioral and educational literature (West, 1985).

Lack of attention to power may result in premature abandonment of useful lines of inquiry, especially in pilot studies (Chase & Tucker, 1975). Pilot studies typically have small sample sizes and low power, and failure to reject the null should be expected. The value of primary importance in pilot studies should be the effect size, not statistical significance; power analysis leads to a focus upon this factor.

Conversely, failure to compute power could result in the use of too large a sample, increasing the chances of obtaining and interpreting a meaningless effect. Computation of power forces attention to the sizes of experimental effects, which should lead to more reporting of effect sizes in the literature and a greater emphasis on the practical importance of findings. In addition to obscuring the substantial significance of effects, unnecessarily large sample sizes are not cost-effective (Luftig & Norton, 1982b; Olejnik, 1984).

The need for better estimation of power has increased with recent developments in research technology. Computerization has allowed researchers to use more sophisticated analyses, including multivariate techniques and meta-analyses. Misinterpretation of multivariate analyses is easily accomplished when the issues of power

and adequate sample size are ignored (Sherron, 1988). Meta-analyses conducted over the last decade have quantitatively aggregated the results of a number of studies in the behavioral and educational fields. Many of these analyses have concluded that techniques such as counseling and special education have had little effect, when the probability of Type II error was unacceptably high in the majority of investigations (Lipsey, 1990). The computer is a powerful tool for data analysis; however, it may have a negative impact upon interpretation of effects if statistical power is ignored.

Definitions

Relevant terms used in this study were defined as follows:

Power: "The power of a statistical test is the probability that it will yield statistically significant results" (Cohen, 1988, p. 1). It is the probability that the significance test will result in the conclusion that the phenomenon exists. It is the a priori probability of rejecting the null. Power is equal to 1-beta (beta = the probability of a Type II error), and can range from 0 to 1. It is often expressed as a percentage value, e.g., 80% power to detect an effect when it is present in the population.

Effect size: The degree to which the phenomenon is present in the population. Specifically, it is the discrepancy between the null and alternative hypotheses; the

degree to which the null hypothesis is false (Cohen, 1973, 1988). It is a hypothesized population value, not a sample statistic.

Educational psychology: The present study investigated doctoral theses in educational psychology. These were defined as theses for the Ph.D. degree which were coded by Dissertation Abstracts International as Subject Code Number 0525: Education, Psychology.

Assumptions

The following assumptions were necessary for estimating the power of significance tests in cases where marginally adequate data were reported:

It was assumed that the assumptions for each statistical test reported were met, since it was not practical to check these assumptions for each statistical test in each thesis.

If alpha was not explicitly stated, it was assumed to be .05.

If the directionality of a significance test was not explicitly stated, a nondirectional alternative was assumed.

Application of these assumptions resulted in the inclusion of studies that might have been eliminated due to lack of explicitly stated detail regarding tests of significance.

Limitations

The major limitation of the study was its generalizability to dissertations in other years and in other educational and behavioral fields of study. Although the sample was derived on a national basis, the comprehension and application of statistical power analysis by doctoral students at any single institution is likely highly related to the attention paid to the topic in the institution's research and statistics courses, and to the level of training in research and statistics of the faculty member directing the dissertation research. Inferences concerning the knowledge and application of power analysis at any single institution would be unwarranted.

CHAPTER II

REVIEW OF THE LITERATURE

Luftig and Norton (1982a), in a review of the literature on statistical power, concluded that researchers grossly overestimate power levels. They cited a 1971 survey by Tversky and Kahnman in which a group of APA conference attendees were asked to calculate the power of a sample research study. The median response of .85 was nearly double the actual power of .473. Haase (1974) noted that the lack of control for Type II errors in psychological and counseling research implied that investigators did not have ". . . a reasonable understanding of a principle which can only help improve the precision and clarity of their research" (p. 126).

The complaints about the nonuse and misuse of power analysis began with Cohen's (1962) study and have not abated. As recently as 1988, Brewer and Sindelar were decrying the abundance of misunderstandings and misconceptions printed in textbooks. The lack of attention to power was one of several indications that current texts minimize the planning of a study, especially in terms of determining appropriate sample size.

Greenwald (1975) surveyed Journal of Personality and Social Psychology authors and editors on the general topic

of significance testing and found interesting results regarding power. Greenwald asked authors to state the probability that they would submit hypothetical results for publication. The authors' responses reflected an overemphasis on the statistical significance of results without regard to power. If the results of a hypothetical study were identified as significant, the mean probability that authors would submit was .588; if nonsignificant findings were reported, the mean probability declined to .064.

More importantly, Greenwald (1975) asked when authors would abandon the problem. If the results of an initial test were significant, the mean probability of abandoning research was .053; but if nonsignificant results were specified, the mean probability was .314. This is an important consideration for pilot studies, which typically employ small sample sizes and consequently have inadequate power. Half the sample failed to answer a question on acceptable levels of beta; Greenwald interpreted this as ignorance of the topic. Another question asked about the probability of setting alpha and beta levels in advance --there was a .63 probability that alpha would be set, but only .17 for beta. Greenwald concluded that the responses of both authors and editors indicated a ". . . substantial lack of standard practice with regard to Type II errors" (p. 5).

Significance Testing

Description

Significance testing in the Fisherian model is analogous to trial by jury (Kraemer, 1985). The investigator's hypothesis is assumed to be false unless proven beyond a "reasonable doubt" to be true. Significance testing is based upon the rejection of the null hypothesis, a "straw man" set up to be invalidated. The "reasonable doubt" is the significance level, set conventionally as less than five chances out of one hundred that a rejection of the null hypothesis is in error.

A rejection of the null hypothesis indicates that it is very unlikely that the results were due to chance differences. If the experiment were to be repeated with a different sample of subjects, it is probable that the same findings would result. The significance test is a one-shot demonstration of a finding based upon the premise that ". . . the theoretically unusual does not happen to me" (Bakan, 1966, p. 425).

Four outcomes are possible in a test of significance:

1. The null hypothesis is true and the decision is to retain it.
2. The null hypothesis is true and the decision is to reject it.
3. The null hypothesis is false and the decision is to retain it.

4. The null hypothesis is false and the decision is to reject it.

Numbers 1 and 4 are correct decisions. Numbers 2 and 3 are incorrect decisions and have probabilities that are under the control of the researcher. Type I error (outcome 2) is the error of rejecting the null hypothesis when it is, in fact, true. The probability of this error is equal to alpha and is referred to as the level of significance of findings. Type II error (outcome 3) is the error of failing to reject the null hypothesis when it is false, or failing to detect an effect when it is present in the population. The probability of this error is equal to beta, and the value (1 - beta) is referred to as the statistical power of the test.

The acceptable level for Type I errors (alpha) is usually set before testing; by convention it is set to .05 or, less frequently, .01. Reviews of research give no indication that beta is routinely set beforehand (Haase, 1974; Kraemer, 1985; Lipsey, 1990; West, 1985). Recommended levels for beta range from .20 to .05. Cohen (1973) specified ideal power, in the absence of context-specific information, as .80. This is 4:1 ratio of Type II to Type I errors. Both McNemar (1960) and Cohen (1973) judged Type I errors to be more serious than Type II. Schmidt, Hunter, and Urry (1976) recommended the use of a .90 level of power and Lipsey (1990) recommended .95; however, .80 has become

the conventional recommendation for power when the conventional alpha = .05 is used.

Relationship Between Alpha and Beta

The relationship between alpha and beta is inverse; therefore, both cannot be set to an extremely small level. Alpha and beta are not additive to 1 because they are conditional probabilities based upon different conditions: alpha on the condition that the null hypothesis is true, and beta on the condition it is false. The truth or falsity of the hypothesis is never known with certainty; the researcher always runs the risk of making an error. It is not possible, however, to commit both errors in a single test. Setting alpha and beta is simply quantifying the probability of making each type of error (Sherron, 1988).

Alpha has definition only if the effect size is equal to zero, and power has definition only if the effect size is greater than or less than zero (Brewer & Sindelar, 1988). In planning an investigation, effect size and power may be thought of as the minimum power and the minimum expected effect size, which are estimated in advance of the collection of data to determine minimum sample size.

Significance Testing and Practical Importance

The practical importance of findings was first suggested by Edwards (1950) to take into account the meaning, in some practical sense, of the magnitude of the deviation from the null. The question, "How much of a

difference is really a difference?" has meaning only in the context of the subject matter. What is a practically important difference in the field of investigation?

> The function of statistical tests is merely to answer: Is the variation great enough for us to place some confidence in the result; or, contrarily, may the latter be merely a happenstance of the specific sample on which the test was made? This question is interesting, but it is surely secondary, auxiliary, to the main question: Does the result show a relationship which is of substantive interest because of its nature and its magnitude? (Kish, 1959, p. 336).

Effect size. The value which is used to determine substantive significance (importance) is the effect size (ES). ES is usually expressed in terms of a proportion of a standard deviation or as a proportion of variance accounted for. The effect size is a metric-free description of the magnitude of the difference or relationship.

The ES index for the difference between means may be expressed as a proportion: the difference between estimated population means divided by the estimated population standard deviation. The value is thus a proportion of a standard deviation, and does not rely upon knowledge of the original units for interpretation. Computation of this value might indicate, for example, that Group A scored

one-half standard deviation higher than Group B on a dependent measure. Again, the question is raised: Is one-half of a standard deviation an important effect? Is it large or small?

ES is a population value; it can only be estimated from sample data. Many other expressions of effect size have been described, for example, r, r^2, and n^2. These are familiar measures of the strength of relationship or proportion of variance accounted for.

Cohen's values. The meaning of an effect size is dependent upon the context of the investigation. However, Cohen (1977, 1988) presented three benchmark values for the interpretation of effect sizes in the behavioral sciences. Cohen cautioned that his values for small, medium, and large effect sizes were relative to the field of study, and presented them only as guidelines in the absence of other data. Nonetheless, they have become accepted as conventions.

Cohen described a small effect as one that is not so small as to be trivial. He defined a small ES as .2 standard deviations, or, alternatively, an r of .10 (r^2 = .01). Only 1% of the total variance is accounted for by the relationship between the independent and dependent variables. This may seem trivial; however, it is equal to the mean difference in IQ between twins and nontwins, or the difference in mean height between 15- and 16-year-old

females (Cohen, 1988). Abelson (1985) analyzed batting statistics for major league baseball players, and found that the proportion of variance accounted for in a given "at bat" that was attributable to individual differences in skill was .00317, less than 1%. This illustrates the premise that the magnitude of an effect size must be evaluated in terms of the context. Millions of baseball fans (and the differential salary levels of major-league players) attest to the importance of an effect size that accounts for less than 1% of variance.

Most effects in the behavioral sciences are small, especially in univariate analyses and when research is conducted outside a controlled laboratory setting (Cohen, 1988; Lipsey, 1990). There is simply too much variation that is due to a wide range of other influences; consequently, educational and psychological measures are too imprecise to capture an effect in isolation.

Cohen defined a medium effect as .5 SD's, $\underline{r}$ = .30, or $\underline{r}^2$ = .09. A medium effect is one for which an individual would normally notice the difference. This magnitude of effect is equivalent to the difference in height between 14- and 18-year-old females, or the mean IQ difference between clerical and semiskilled workers, or between professionals and managers (Cohen, 1988).

Cohen's large effect size was defined as not so large as to preclude reasonable tests of significance. A large

effect is equal to .8 SD's, r = .50, or r^2 = .25. This size of effect reflects the IQ difference between Ph.D.'s and college freshmen, or the mean difference in height between 13- and 18-year-old females (Cohen, 1988).

Olejnik (1984) suggested that anticipated effect size may be determined from Cohen's benchmarks, or from the results of meta-analyses involving similar factors or variables. He noted that Cohen's definitions were ". . . probably the best known and widely accepted guidelines currently used by researchers" (p. 44). In addition, when deciding a priori what a meaningful effect would be, it is better to underestimate the effect size than to overestimate it, since overestimation would reduce statistical power. Olejnik referred to the general anticipation of a large effect as wishful thinking.

Power Analysis

Description

The purpose of power analysis in the design phase of a study is to ensure a reasonably high probability for the detection of effects of the anticipated magnitude when they exist in the population (Rogers & Hopkins, 1988). Power analysis may also be conducted on a post hoc basis, and power should always be computed when the null is not rejected (Fagley & McKinney, 1983).

Power analysis requires the determination or estimation of four values: sample size, alpha, effect size, and power.

When any three of the four values are fixed, the fourth is fixed and can be determined (Cohen, 1988). Cohen described four types of power analysis:

1. Power as a function of alpha, ES, and sample size. This is the relevant type of analysis for the present study.
2. Sample size as a function of ES, alpha, and power. This should be the criteria by which sample size for a study is determined a priori.
3. Effect size as a function of alpha, sample size and power. This determines the detectable ES given the other specifications. This type of analysis could be used for comparisons of results in literature reviews. Power may be defined as 1/2 or .5 by convention, and the effect sizes computed may be used to compare the sensitivity of studies.
4. Alpha as a function of sample size, power, and ES. This type of analysis is very uncommon due to research convention. The .05 and .01 alpha levels are standard in the research community. Researchers apparently are willing to accept large (unknown) beta errors rather than risking the use of unusual levels of alpha.

Factors Affecting Power

The three parameters described above directly affect the level of power: sample size, the significance

criterion, and the effect size. Other factors have an indirect effect on power, chiefly by influencing the reliability of findings or increasing the effect size.

Sample size and reliability of results. The reliability of a sample value may be dependent upon a number of factors, but it is always dependent upon the size of the sample (Cohen, 1988). Sample size is a key value, under the control of the researcher, which affects the precision of the measurement of the size of the effect. Larger sample sizes yield more accurate estimates of any parameter, including the effect size.

Sample size affects power because it is directly related to the standard error of sample statistics. The formulae for the standard error of sample statistics contain some value related to sample size, if not the sample size itself, in the denominator. As sample size increases, other things being equal, the standard error will be reduced and the statistic will be a more reliable estimate of the parameter. Increased precision of a test statistic reduces random error and thus increases the likelihood that a reliable effect will be detected.

The reliability of a sample value was defined by Cohen (1977) as ". . . the closeness with which it can be expected to approximate the population value" (p. 6). It is always directly dependent upon sample size, although other factors (e.g., the actual population value, shape of the population distribution, unit of measurement) may affect it.

Decisions on sample size are typically reached by tradition, convention, availability of data, "experience," and negotiation, ". . . rarely on the basis of a Type II error analysis, which can always be performed prior to the collection of data" (Cohen, 1962, p. 145). Sample size is often determined in a haphazard or arbitrary manner (Luftig & Norton, 1982b), and ". . . it is a rare research paper that discusses how the sample size was determined as part of the description of the sample" (Olejnik, 1984, p. 40).

Brewer and Sindelar (1988) noted that ". . . textbook authors and research reporters often treat sample size as a given rather than as a value to be determined through statements such as 'If n is large (>30), then . . .'" (p. 83). A conventional sample size of 30 (alpha = .05) does not yield adequate power unless ES is large. If the ES is small (.2 standard deviations), then power = .12. If ES = .5, power = .61. Only if ES = .8 will power be adequate (power = .86). Power analysis is clearly a rational method of determining adequate sample size for a significance test.

Any other factor which introduces random or nonrandom error reduces the reliability of the sample statistic and therefore reduces power. Reducing error variance by blocking on a variable related to the dependent variable will increase power. The use of covariates, which offers more precision than blocking, will increase power, as will

the use of correlated or matched samples (Judd & Kenny, 1981; Rogers & Hopkins, 1988). Increasing the reliability of measurement of the dependent variable or the covariate will also reduce error variance and increase power (Bonett, 1982; Rogers & Hopkins, 1988; Zimmerman & Williams, 1986).

Type I error rate. The alpha level is another mathematical determinant of power. Statistical inference involves balancing the two kinds of errors. For example, alpha may be set with no consideration of beta. If alpha is set to .01 and power = .50, then, even though the investigator did not consider power, the relative importance of the two types of errors has been defined as 50/1. The investigator is operating (usually unknowingly) on the premise that falsely finding an effect is 50 times as serious as failing to find an effect that exists in the population (Cohen, 1977).

Use of the Bonferroni procedure is recommended in many texts when testing multiple hypotheses. The emphasis given this procedure in texts demonstrates the emphasis on Type I errors to the exclusion of Type II. Use of Bonferroni procedures (distributing the acceptable experimental alpha across hypotheses) increases the number of Type II errors the researcher is risking, unless other parameters, such as sample size, are adjusted. When the experimentwise alpha is set at .05, the use of the Bonferroni procedure typically results in unacceptably low power (Keselman & Keselman, 1987; Westermann & Hager, 1983).

Directional testing has same effect as increasing alpha, if the sample result is in the specified direction. If the result lies in the opposite direction, the experimenter has no power at all to detect effects (Cohen, 1988).

Effect size. If an experiment results in no difference, or if the sample statistic is not significantly different from a hypothesized population value, then the effect size is zero. If statistical significance is achieved, then the effect size is some specific nonzero value in the population, which can be estimated (Cohen, 1988).

ES is a metric-free index. In treatment effectiveness research, effect size may be conceptualized as the difference between population means or the between-groups variance. If group 1 scored three points higher than group 2, consumers of research need to know whether three points is a meaningful increase. Three points in relation to what?

If effect size were expressed in terms of specific dependent variables, it would be impossible to make generalizations about statistical power. The most common effect size value, d, the difference between means, is general for all dependent variables. It expresses the difference in terms of units of variability, standardizing the difference between means. It is derived by dividing the difference between means by the common standard deviation.

ES is thus expressed in terms of a proportion of a standard deviation. ES is free of the original measurement units and can be easily compared across studies.

Effect size cannot be computed; it must be estimated, and is the most troublesome of the factors that determine power (Lipsey, 1990). Measures of effect size include those which indicate the proportion of variance accounted for in the dependent variable and those which indicate the difference between means.

The ES estimate d is an expression of the effect size in standard deviation units: $\underline{d} = [(\mu_1 - \mu_2)/\sigma]$. $\underline{d}$ is used to express the ES in studies which focus upon the difference between means, such as the $\underline{t}$ test.

The Pearson product-moment correlation coefficient $\underline{r}$ is another common measure of effect size. $\underline{r}$ leads to an expression of the proportion of variance accounted for in the dependent variable. Values of $\underline{r}$ can easily be converted to $\underline{d}$. Other estimates of ES are available, but all ES values can be related back to either $\underline{d}$ or $\underline{r}$. Calculation of effect sizes can be difficult, especially in multivariate designs; however, direct calculation of ES is not required for a priori power estimates. What is required is some evidence concerning the expected sizes of effects.

In the absence of other data, Cohen (1988) recommended hypothesizing a medium effect. Lipsey (1990) described three more precise methods of estimating ES: the actuarial

approach, the statistical translation approach, and the criterion group contrast approach.

The actuarial approach uses the results of other studies to estimate ES. A distribution of ES estimates in the relevant literature can be created to determine the range of probable effect sizes. The distribution may then be divided into thirds, each third representing a range of values for small, medium, and large effect sizes. The midpoint of each range could then be used as a summary value for anticipated effects. If the researcher desired enough power to detect a small effect, the power tables would be entered with the midpoint value from the bottom third of the ES distribution. Use of the actuarial method would be most feasible in fields in which meta-analyses are available.

The statistical translation approach translates ES from standard deviation units into a form that is more easily assessed, usually in terms of the dependent measure. It may be conceptualized as a "raw" ES. For example, if the dependent variable were an achievement test with a standard deviation of ten points, the researcher would define an ES of .50 as five points. This version of effect size may be more meaningful in context than the standardized ES necessary for entering power tables. Researchers could conceptualize the anticipated ES in terms of test norms and previous research using that particular dependent measure.

The criterion group contrast approach estimates ES in terms of the effect size that would be practically important in terms of the research context. A comparison is identified in which the difference between two groups is known to be practically significant. This comparison serves as a benchmark against which other effects are compared. The determination of the benchmark is achieved by identifying two groups who clearly demonstrate a difference, e.g., inpatient v. outpatient mental health clients. These two groups would be administered some measure of functioning, and the difference would establish the benchmark. An investigator researching treatment effects of counseling would then define a practically important effect as equivalent to the benchmark (one as large as the difference between inpatient and outpatient clients), and would enter power tables with the standardized ES derived from the inpatient-outpatient comparison (Lipsey, 1990).

Anything which increases the measured ES increases statistical power. ES can be increased by increasing the duration or intensity of a treatment (Rogers & Hopkins, 1988). Increasing treatment strength and maintaining treatment integrity may be a more cost-effective means of increasing statistical power than merely increasing sample size.

Interpretive Value of Power Analysis

After Cohen's (1962) seminal study, a debate on the

interpretation of statistical power was carried on in psychological and educational journals (Brewer, 1972; Brewer, 1974; Dayton, Schafer, & Rogers, 1973; Meyer, 1974). Much of the debate centered upon the interpretation of power after the results of significance tests were known.

Cohen (1973) succinctly stated the debate's resolution: No matter what the value of statistical power, the significance test itself is not affected. Once data are gathered and analyzed, power analysis fades into the background. If significant results are obtained, the issue of power is moot; a Type II error did not occur.

If, however, nonsignificant results are obtained, interpretation is dependent upon the power of the analysis. Unless power is high, one cannot conclude, even implicitly, that there is no difference. This conclusion is ". . . always strictly invalid, and is functionally invalid as well unless power is high" (Cohen, 1977, p. 16). Cohen went on to state that this conclusion has a high frequency of occurrence which ". . . can be laid squarely at the doorstep of the general neglect of attention to statistical power in the training of behavioral scientists" (p. 16).

The inference that the difference is negligible or trivial is valid if power is high (Cohen, 1977). To make this inference, one must remember that it is the effect size that is being tested, i.e., that the population value is below some specified trivial effect size. If it is, the

conclusion of negligible effect is valid; this is tantamount to accepting the null. However, it requires a very large N to establish. For example, consider the case in which $\underline{r} < .10$. For alpha = .05 and power = .80, an N of 783 is required. To establish the conclusion, power should probably be higher; for power = .90, the N required would be 1046. A sample size of over 1000 would be necessary to establish that the effect is negligible. "Accepting the null" is not a valid conclusion because the probability of wrongly doing it cannot be held constant, as it can for rejecting at a certain level of alpha.

Power Surveys

Cohen (1962) conducted the first power analysis of a body of literature, surveying 70 articles published in the Journal of Abnormal and Social Psychology. He calculated the mean a priori power of the articles, using the conventional definitions of small, medium, and large effect sizes. Cohen found that the average power of the statistical tests in the articles was .18 for small effect sizes, .48 for assumed medium effect sizes, and .83 for large effect sizes. None of the 70 articles reported statistical tests with adequate power to detect small effects. He concluded that the investigators had a poor chance of rejecting their null hypotheses, unless their effect sizes were large.

Table 1 summarizes the findings of 15 power-analytic surveys of published literature. All the cited surveys used Cohen's definitions of small, medium, and large effect sizes. The mean levels of statistical power for detecting small or medium effects were inadequate in all cases. If average effect sizes in psychology and education are indeed small, it appears that researchers are allowing themselves less than a one-in-three chance of detecting effects of this magnitude. Mean power to detect large effects was reported to be inadequate in five (one-third) of the surveys.

The data in Table 1 support Cohen's (1988) contention that the power surveys done since his initial study in 1960 have not shown an increase in statistical power in published educational and behavioral research studies. His conclusion was that the extensive literature on power since his 1962 paper has had little or no effect on actual practice. The present study employed the same framework as Cohen's for investigating levels of power in dissertation research.

Summary

Available evidence indicates that researchers in education and psychology lack an understanding of statistical power and/or ignore its application in their investigations. Three factors directly affect the power of a significance test: sample size, Type I error rate, and effect size. Of these, researchers conscientiously attend only to Type I error rate, which has a conventionally

Table 1. Levels of statistical power by effect size in research reviews of behavioral and educational literature

Research Domain	Year	Size of Effect		
		Small	Medium	Large
Abnormal Psychology	1962	.18	.48	.83
Education	1972	.13	.47	.73
Science Education	1972	.22	.71	.87
Health, Physical Education	1972	.15	.55	.81
Counselor Education	1974	.10	.36	.74
Mathematics Education	1974	.24	.62	.83
Communication	1975	.18	.52	.79
Speech Pathology	1975	.16	.44	.73
Applied Psychology	1976	.25	.67	.86
Health, Physical Education	1977	.18	.39	.62
Occupational Therapy	1982	.37	.65	.93
Science Education	1983	.23	.63	.85
English Education	1983	.22	.63	.86
Evaluation Research	1985	.28	.63	.81
Adult Education	1985	.22	.66	.88

Sources (respectively): Cohen (1962); Brewer (1972); Pennick and Brewer (1972); Jones and Brewer (1972); Haase (1974); Clark (1974) (cited in Lipsey, 1990); Chase and Tucker (1975); Kroll and Chase (1975); Chase and Chase (1976); Christiansen and Christiansen (1977); Ottenbacher (1982); Wooley and Dawson (1983); Daly and Hexamer (1983); Lipsey et al. (1985); West (1985).

accepted limit. Indeed, Fagley and McKinney (1983) noted that the Type I error rate has been (incorrectly) accepted as the probability of making an error of inference.

Surveys of the level of power in published research indicate inadequate power for detecting anything other than large effects. Significance testing for very large effects has been described as superfluous, as conclusions often meet the requirements for the Inter-Ocular Trauma Test (it hits you right between the eyes). The observation that most effects in education and psychology are small leads to the conclusion that much of the research in education and the behavioral sciences has been conducted without an adequate chance of finding the expected outcomes.

The low levels of power in published research are of greater concern because of the publication bias toward positive results. Authors of power surveys have uniformly noted that the average power of studies is overestimated due to this bias. The present study was an initial step towards investigating the accuracy of this observation by surveying power levels in unpublished dissertation research.

CHAPTER III

METHOD

The study investigated the levels of a priori power in selected dissertation research (see Appendix B for listing). A priori power estimations are less precise than post hoc power calculations, but they are essential to the determination of values critical to research planning, such as sample size. This survey, like those listed in Table 1, was concerned with the power of the research to detect effects of various sizes. The study followed the format of Cohen's (1962) original study of power levels in abnormal and social psychology as well as the power surveys listed in Table 1. The actual effect sizes, while interesting, were not relevant to the power survey, which is more general in nature.

Sample

This investigation surveyed doctoral dissertation research successfully defended in the field of educational psychology during the year 1988. The year 1988 was selected for two reasons: (1) it was the most recent year for which complete data was available in the Dissertation Abstracts International database, and (2) it was the most recent year for which it could reasonably be expected that theses would

be ready for dissemination through the interlibrary loan system.

A search of the Dissertation Abstracts On-Disk for the year 1988 in the fields of education and psychology resulted in 921 items. Sequential steps were taken to limit the sample to a definable subject area with a manageable number of dissertations (originally defined as 40-60), as follows:

1. The degree field was limited to Ph.D., resulting in 513 items.
2. The subject area was limited to Educational Psychology (code 0525), resulting in 104 items. Educational Psychology was selected because the field has a research emphasis and encompasses both educational and behavioral research.
3. Only abstracts containing the term "experimental" (including "quasi-experimental") were selected, resulting in the final sample of 69 items. This term was selected in order to eliminate descriptive studies and others for which power analysis is irrelevant from the search. The use of this term had its intended effect; all dissertations received were appropriate for power analysis.

All dissertations available through the InterLibrary Loan system were reviewed for the inclusion of adequate data. Adequate data was defined as including statistics which allowed for an a priori power analysis: a test of

significance was conducted in which the Type I error rate, sample size, and test statistic were reported (Type I error could be implied rather than reported).

Of the 69 dissertations located in the database search, 34 were unavailable through the InterLibrary Loan system. Theses unavailable for review were from diverse colleges and universities, ranging from Carnegie-Mellon to the University of Southern California. There was no apparent difference in the size or prestige of institutions from which dissertations were received or not received. The abstracts of unavailable dissertations were reviewed for any references to power analysis or effect size; no references were found. No apparent bias was introduced into the study by the elimination of theses that could not be obtained through InterLibrary Loan.

Thirty-five dissertations were reviewed for the inclusion of data necessary for power analysis, and all 35 met these requirements. No more than four dissertations were received from any single institution. Thirty-four authors completed Ph.D. degrees in the fields of education or psychology (one author's field of study was kinesiology). One-fourth of the authors earned degrees in education, 18% were in departments of psychology, and 21% listed the degree program as educational psychology.

Although previous power surveys have limited the studies reviewed to those with the most common univariate

statistical tests (West, 1985), this investigation included all tests for which a power analytic table was available in Cohen's (1988) standard text on power analysis. With the increasing utilization of computer technology in the analysis of data, more and more research is multivariate in nature. To exclude multivariate analyses simply because previous surveys excluded them would be an exercise in stagnation. The list of tests to be included was defined as:

1. the *t* test for means
2. the significance of a product moment correlation coefficient
3. the test for differences between correlation coefficients
4. the test that a proportion is .50 and the sign test
5. the test for differences between proportions
6. the chi-square tests for goodness of fit and contingency tables
7. analysis of variance and covariance
8. multiple regression and correlation analysis
9. set correlation and multivariate methods, including MANOVA, MANCOVA, principal components and factor analysis, discriminant analysis, and hierarchical analysis.

The unit of sampling and analysis was the individual thesis (after Cohen, 1962, and others). If more than one

statistical test was performed, mean power levels for each dissertation were calculated and used as a summary statistic.

Review Procedures

The following information was recorded during review of each dissertation: number of stated research hypotheses, the type and number of statistical tests performed, the stated level of significance, and the sample size. Only tests of stated hypotheses were recorded. Secondary tests such as reliability checks, routine tests of correlation coefficients, etc., were excluded. Tests of the major hypotheses provided the best estimate of the overall power of the series of tests to provide an overall answer to the research question (Daly & Hexamer, 1983).

A recording instrument was developed to facilitate data collection. This instrument was based upon the data necessary to complete an a priori power analysis and a record of relevant background information on each thesis. A prototype research review coding sheet presented in Cooper's (1984) handbook on research reviewing was modified after a review of the methods used in previous power surveys. A limited pilot investigation was conducted to determine the completeness and utility of the instrument. The instrument included more information than was necessary for this study, and additional notes had to be made concerning values such as the number of covariates in ANCOVA analyses. Except for

minor variations, however, the instrument as designed was sufficient for data collection purposes.

The recording instrument is included in Appendix A. Space is provided for recording of the following background information: author, title, date of submission, degree area, and institution. For each research hypothesis, the following data were recorded: formal statement of hypothesis, statistical test(s) performed, results of the test (rejection/nonrejection of the null), number of groups (if applicable), number of independent and dependent variables, value of the test statistic(s), acceptable Type I error rate, directionality of alternative hypothesis, numerator and denominator degrees of freedom (if applicable) and sample size. Both the total sample size and the sample size per group were recorded. Space is also provided for recording whether or not the investigator conducted a power analysis, rationale for sample size determination, and whether or not an effect size was calculated.

Procedure for Analysis

For each statistical test, power was read directly from Cohen's (1988) tables. Power was recorded for small, medium, and large effects, using Cohen's standard definitions.

Descriptive data regarding each thesis was recorded, e.g., number and type of significance tests. The following summary data were reported and analyzed: mean number of

significance tests per thesis, frequency tables for alpha values and sample sizes, appropriate measure of central tendency for these values, number of dissertations by academic department (education v. psychology), number of each type of test for which power was calculated, number of tests for which the result was significant, and mean sample size.

Mean power values were calculated for each individual thesis. The mean power for detecting small, medium, and large effects was recorded. To facilitate comparison with previous reviews, which excluded multivariate statistical analyses, power levels were recorded separately for multivariate tests.

Summary

The present investigation is a descriptive study of the level of power in doctoral dissertations. The sample of theses examined and the restrictions for inclusion in the sample were described. An instrument for recording relevant data from each dissertation was developed based upon a review of previous power analyses and the requirements for entering tables in Cohen's (1988) standard text. This text was used for all power calculations.

The unit of analysis was the individual thesis; methods of extracting mean power for detecting small, medium, and large effects from each thesis were described. A description of the summary statistics used to describe the data was presented.

CHAPTER IV

RESULTS

Chapter IV presents the results of the power analysis, including the number and type of significance tests conducted, their results, sample sizes, and estimated a priori levels of power to detect small, medium, and large effects. Table 2 lists each thesis (see to Appendix B), its total sample size, cell sample size, the type(s) of significance tests conducted, the number of tests conducted and the results, and the mean estimated power to detect small, medium, and large effect sizes.

Relevant Descriptive Data

Table 3 summarizes the data regarding the number of significance tests conducted, Type I error rates, significant findings, and sample sizes reported for all but one of the dissertations reviewed in this study. Data from that dissertation was eliminated from this table because the study in question had a sample size of 21,337, which was extremely high in comparison to the other reported studies (the next largest sample size was 278). Therefore, inclusion of the study would have greatly skewed the mean sample size. Although that study is not included in the data reported in Table 2, the power levels for that study were included in the power analyses.

Table 2. Sample sizes, significance tests, and mean power levels for the sample of dissertations

Thesis Number[a]	Total Sample Size	Cell Sample Size	Significance Tests Conducted	Number of Tests	Number of Signif. Results	Power[b] to detect ES size: Small	Medium	Large
3	62	5	ANCOVA	15	4	8	28	60
6	32	5	ANCOVA	9	4	7	20	44
15	183	92	ANCOVA	1	1	48	99	99
25	215	9	ANCOVA	8	1	48	92	99
33	278	35	ANCOVA	10	6	36	97	99
5	39	13	ANOVA	21	2	9	36	72
9	14	7	ANOVA	27	4	7	17	36
12	45	15	ANOVA	12	10	8	29	62
14	42	21	ANOVA	10	2	10	38	73
16	122	6	ANOVA	14	8	15	68	95
1	72	18	ANOVA	25	6	17	54	78
11	60	15	ANOVA	21	4	16	53	79
22	85	9	ANOVA	15	2	10	46	87
24	84	17	ANOVA	39	24	9	32	65
27	104	26	ANOVA	5	1	10	48	86
30	173	86	ANOVA	6	3	9	75	99
32	115	38	ANOVA	3	2	23	86	99
34	48	24	ANOVA	9	6	11	50	85
35	149	12	ANOVA	23	5	14	66	98
8	24	12	ANOVA	12	3	7	23	51
26	60	15	ANOVA, t	47	11	8	24	51
17	22	11	ANOVA, r	11	5	8	26	60
19	20		chi, r	17	6	9	32	69
21	182	26	chi, r	29	8	10	53	90
20	200	25	MANOVA	5	1	14	81	99
23	119	10	MANOVA	10	7	17	82	99
28	60	15	MANOVA, ANCOVA	13	4	11	47	85
10	129	5	MANOVA	15	4	19	69	99
7	71	24	MANOVA, r	46	0	4	51	97
13	21337		M.REG	2	2	99	99	99
18	55	18	M.REG, ANOVA	9	0	27	76	98
31	60		sig. of prop.	45	23	17	68	95
2	21	7	t-tests	15	1	7	15	31
4	56	27	t-tests, r	85	45	11	45	83
29	100	50	t-tests, r	202	156	8	37	97

[a] See Appendix B.
[b] Reported as percent.

Table 3. Summary statistics of significance tests reported in the sample of dissertations[a]

Number of significance tests associated with major hypotheses:		
Range:	2-202	
Mean:	23.56	
Median:	13	
Frequencies of Type I error rates:		
Set by researcher to .05:	20	(57%)
Implied level of .05:	11	(31%)
Set by researcher to .01:	2	(6%)
Set by researcher to .001:	1	(3%)
Set by researcher to .10:	1	(3%)
Number of significance tests resulting in rejection of the null hypothesis:		
Mean number of rejections:	10.6	
Median number of rejections:	4.0	
Percent rejected:	44.4	
Sample size:		
Mean total sample size:	91.2	
Median total sample size:	60.0	
Mean sample size per cell:	21.8	
Median sample size per cell:	15.0	

[a]N = 35.

Due to a small number of studies with discrepant numbers of significance tests or sample sizes, the means and medians reported for the data in Table 3 were quite different. Since the distributions of sample sizes, number of significance tests, and number of significant results were skewed, the median is the best measure of central tendency for describing the group as a whole on these three variables.

Approximately one-half of the significance tests conducted resulted in the rejection of null hypotheses at the alpha levels set (or implied by the use of $p < .05$) by the researchers. Since the probability of rejection (power) is closely related to sample size, especially sample size per group, the average cell sample size for various kinds of significance tests was computed. The sample size per group or cell was defined as the total N divided by the number of groups (for *t*-tests and one-way ANOVAs) or cells (in factorial designs), or the number of pairs of scores (in correlation tests).

Table 4 reports the number of studies and sample size distributions for analyses of covariance (ANCOVA), univariate analyses of variance (ANOVA), multivariate tests (MV), tests of correlation coefficients, and *t* tests. (Tests which were conducted in fewer than three theses are included in the "Other" column of Table 4). One-half of the dissertations employed univariate ANOVA's; the remainder

Table 4. Average sample sizes by type of test reported in the sample of dissertations

	ANCOVA	ANOVA	MV[a]	r[b]	t[c]	Other[d]
Mean n:[e]	21.4	20.8	15.2	76.0	28.0	78.4
Median n:	60.0	15.0	15.0	56.0	27.0	60.0
Mean N:[f]	154.0	75.1	126.3	76.0	57.0	87.3

[a] Multivariate tests of significance.
[b] Tests of correlation coefficients.
[c] t-tests for the difference between means.
[d] e.g., chi-square, tests of proportions, multiple regression.
[e] Cell sizes.
[f] Total sample sizes.

were distributed as follows: tests of correlation coefficients, 17%; ANCOVA, 17%, MANOVA, 14%; t-tests, 9%; chi-square, 6%; multiple regression, 6%, and the test for significance of a proportion, 3%.

The correlational analyses and "other" statistical analyses had approximately three times as many subjects per group as the analyses that tested for the differences between means. Tests of means were utilized in 89% of the dissertations in the sample.

Power Analysis

Table 5 summarizes the results of the power analyses for the group of dissertations as a whole, and for multivariate analyses. Cohen's standard values for small (d = .20), medium (d = .50), and large (d = .80) effect sizes were used, and power was read directly from Cohen's (1988) tables. The unit of analysis was the thesis; power computations for all major tests in each thesis were averaged to yield a single value for each anticipated effect size (small, medium, and large).

The mean power of statistical tests in the sample to detect small, medium, and large effects was larger for the multivariate analyses than for the sample as a whole. The mean power estimates are reported in Table 5. The median power levels were also computed for the total sample. The median power to detect small effects was .10; for medium effects, .465; and for large effects, .855. Since the mean

Table 5. Mean power estimates to detect significant effects for total sample of dissertations and multivariate analyses

Effect Size	Total Sample[a]	Multivariate Only[b]
Small	.169	.150
Medium	.541	.720
Large	.796	.965

[a]N = 30.
[b]N = 5.

and median power levels were of similar magnitude, mean power levels are the focus of discussion in this section.

To determine whether extremely large effect sizes or chance factors accounted for significant results in studies with extremely low power (less than an 80% chance to detect even large effects), the percentage of significant results was calculated. One-third of the significance tests run in these studies resulted in rejection of the null hypothesis. Nearly all had experimentwise alpha rates much higher than the .05 level; the number of significance tests conducted in these studies ranged from 1 to 202. Review of each study indicated that some "significant" results were probably spurious (e.g., 1 significant finding out of 15 tests run at a .05 alpha level, power = .31), and some were likely due to very large effect sizes (e.g., 10 significant results out of 12 tests at .05 alpha, power = .62). The proportion of "significant" results ranged from .00 to 1.00. The median proportion was .26.

The power levels for multivariate analyses were slightly higher, especially for detecting large effect sizes. Most of the multivariate analyses (MANOVAs) were one-way analyses, leading to a larger sample size per cell than the univariate ANOVAs. Most (78.6%) of the ANOVA designs were factorial; half were three-factor or more complex designs. Both the ANOVA and MANOVA investigations had a median cell sample size of 15; however, the complex

univariate analyses spent valuable degrees of freedom on interaction terms, lowering the overall power of the study.

Only three researchers reported a power analysis. Two (from the same university) were a priori analyses. The other was a post hoc analysis that was conducted after the investigator failed to find more than two significant results from a total of 21 significance tests.

With the exception of the two a priori power analyses, none of the studies reported any rationale for the a priori determination of appropriate sample size. With the exception of a single study (which reported effect sizes following a chi-square analysis), the only effect sizes reported were associated with tests of correlation coefficients and multiple correlations, where the sample statistic being tested was itself a measure of effect size. Effect sizes were not anticipated before conducting the research, nor were they computed following tests of significance.

Summary

Thirty-five dissertations comprised the final sample. The median number of significance tests conducted per thesis was 13; slightly less than one-third of the null hypotheses were rejected.

The mean power to detect a small effect was .169, for medium effects the mean power was .541, and the mean power to detect large effects was .796. Power was slightly higher

for multivariate tests of significance; in general, these were not as complex as the univariate analyses. Only three of the dissertations reviewed reported a power analysis.

CHAPTER V

DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

The present investigation sampled Ph.D. level doctoral dissertations completed in 1988 in the area of educational psychology. The purpose of the study was to investigate the a priori levels of statistical power; that is, what were the a priori chances of rejecting the null hypothesis, given that an effect was present in the population? The unit of analysis was the individual thesis. The levels of power to detect small, medium, and large effects in unpublished dissertation research reviewed in this study were unsatisfactorily low.

The levels of power in these studies were similar to those reported in reviews of the published literature in the fields of education and behavioral science. The findings of this investigation supported the conclusion, initially established through a review of the relevant literature, that researchers in education and psychology lack an understanding of power, and ignore its application in their investigations.

The publication bias towards significant results has been well-established, and it has been hypothesized that recent power surveys of the published literature were

overestimates of the power of research-in-general. The present investigation reviewed one source of unpublished research, doctoral dissertations, and found that the power levels were quite similar to those of power surveys for refereed journals.

Discussion

The levels of power in dissertation research in educational psychology were similar to the levels reported in Cohen's (1962) original study, and to studies of the levels of power in published research that were enumerated in Table 1. Levels of power in unpublished dissertation research were no better than, and no worse than, levels of power in research journals.

The theses reviewed had, on the average, only a 17% chance of finding an existing "small" effect. Only one study (with an N of over 21,000) had 80% power to detect small effects. Evidence that most effect sizes in behavioral and educational research are small was presented in Chapter II; the results of the present study indicate that investigators had very little chance of finding effects of this magnitude.

Only 20% of the sampled studies had 80% power to detect effects of medium size. The dissertations reviewed in this study had a 50-50 chance of detecting a medium effect. The researchers would have had the same chance if they had set up their hypotheses, collected data, and flipped coins. The

data in Table 3 indicated that the investigators found significant results for approximately one-half of the statistical tests they conducted. It would be incorrect to assume, however, that they were dealing primarily with effects of medium size. If the median, rather than the mean, number of significant results is compared to the median number of tests, the researchers rejected less than one-fourth of their null hypotheses.

The investigations did have marginally adequate power to detect large effects (.796). Sixty percent of the studies had power greater than .80; however, a substantial number of the dissertations did not have adequate power to find large effects that existed in the populations. Since all but two of the reviewed studies had at least one significant result, two conclusions are possible: (1) the actual effect sizes were extremely large for behavioral and educational research, or (2) the investigators were, in a moderate percentage of cases, interpreting chance effects.

Cohen (1988), in describing "large" effect sizes, stated that his conventional values were defined so that the running of significance tests was not superfluous. Differences larger than .8 standard deviations should be interpretable without testing for significant differences. Since most of the sampled studies did not report effect sizes, it is assumed that the investigators routinely conducted significance tests, whether they were needed or

not. Indeed, the study with an N of 21,000 reported results of tests of significance, when this information was clearly unnecessary. The chance that the study in question made a Type II error was much less than 1 in 100, consequently, "significance" was assured.

Cohen (1988), in the preface to the third revision of his text on statistical power, summarized the literature on the power of research in the behavioral sciences and concluded that his 1962 paper and subsequent publications have had little effect on actual practice: He wrote, ". . . it is clear that power analysis has not had the impact on behavioral research that I (and other right-thinking methodologists) had expected" (p. xiv). The present study supports Cohen's conclusion.

Conclusions

This power survey, in conjunction with those conducted for published journal literature, indicated inadequate power for detecting anything other than large effect sizes. The utility of conducting studies that can reasonably be expected to find only large effects can reasonably be questioned, not only by research methodologists, but also by individuals and organizations that fund research studies.

The low levels of power in doctoral-level dissertation research is surprising, given the concern of most Ph.D. candidates for finding "significant" results. The power levels are also disappointing, in that they indicate that

current doctoral programs are not adequately training researchers, or consumers of research, in the issues relevant to power analysis. These issues include the concept of the effect size, which is critical to gauging the practical importance of results. Only one of the reviewed theses addressed the concept of effect size in the interpretation of results.

Adequate power is essential to the interpretation of nonsignificant findings. Conducting an investigation without adequate power is analogous to target-shooting -- if the target is not consistently available, the shooter cannot possibly hit it. If I count the number of "hits" and "misses" with no reference to the number of times the target was actually available, I would misinterpret the accuracy of the shooter. In a similar fashion, nonsignificant results have been interpreted, in both the journal literature and in the dissertations reviewed in this study, as though no effect exists--when the researchers' "targets" were not consistently available.

The problem of low power is especially relevant in program evaluation and treatment-effectiveness research. In recent years, a number of "social" programs and interventions have come under fire due to published reports of "efficacy research." Adequate power in studies that conclude "no difference" or "no effect" is crucial to proper interpretation of results, but power is rarely computed or

reported. If the findings of this study and the power analyses reported in the literature are accurate, conclusions of "no effect" are being drawn on the basis of faulty premises.

Research design texts often refer to the "error of misplaced precision." This type of inferential error could well be expanded to include the misplaced precision regarding Type I and Type II errors. Educational and behavioral researchers are quite consistent in reporting alpha errors, at least in the case of the individual significance test. (The Type I errors attributable to experimentwise alpha levels is, however, another area of concern.) We are careful to avoid concluding that an effect exists when, in fact, it does not. The results of this power analysis, in conjunction with those reviewed in Chapter II, indicate that we are not nearly so careful about missing an effect that is present in the population.

Educational and behavioral researchers must deal with the question of whether or not "small" effects are important. Most research in education and psychology is dependent upon the test of significance, but most tests of significance are likely to miss small effects. Some would argue that this is a good thing, that detecting small effects is wasteful and misleading.

The question of how large an effect must be before we regard it as meaningful can only be answered in the context

of the individual investigation. The emphasis on the test of significance leads researchers away from the issue of effect size. Power analysis, on the other hand, forces consideration on the hypothesized or actual size of the effect.

The most disturbing aspect of the results of this investigation was the lack of attention to effect size. Only one study reported and interpreted effect sizes. It appears that the next generation of researchers is destined to repeat the mistakes of the past, focusing all the attention upon a single-shot demonstration of a "significant" finding. Until more attention is paid to effect sizes, "significance" will continue to stand for substance. Perhaps the problem is one of semantics. Perhaps the fault lies with the individual who first declared that a theoretically unusual occurrence was "significant," instead of rare.

Since it is usually the goal of researchers to find and interpret significant effects, it is reasonable to examine the question of why so few researchers conduct power analyses prior to (or following) the specification of an adequate sample size and the collection of data. It is the conclusion of this researcher, after computing power for hundreds of significance tests, that power analyses are not difficult to do, but that learning how to conduct them for a host of different types of significance tests is a

time-consuming task that may well be beyond the scope of most graduate research courses. However, the concept of effect size is not.

The problem with effect sizes is that there is no common metric. Students must learn different effect-size calculations for each type of significance test. They may learn to calculate them, but the lack of a common metric renders them virtually uninterpretable to most students. Emphasis in training programs should not be concentrated on the calculation of effect sizes, but on interpretation of them. Computerized statistical analysis programs should include both effect sizes and power on the printouts for any test of significance. Until these routines are included in computer statistical package programs, and power parameters routinely included on printouts, effect sizes and the power of significance tests will probably continue to be ignored.

Lack of technical skill may account for a great part of the problem, but "lock-step" thinking regarding significance testing is a problem for both graduate faculty and journal editors (including field readers). Educational and behavioral research is built upon the interpretation of significance tests. Researchers who may take great pains with all threats to the internal validity of a study pay little or no attention to the validity of the statistical conclusion, and even less attention to its practical value.

It is possible that students (and faculty) do not understand the concepts of effect size and power because they are not taught well. Educational and behavioral researchers, especially those who train other researchers, need to find new ways to teach these concepts, or technology will make the problem worse. We have made it easy to run highly complex analyses, and we are no longer satisfied with simple ones. We do not see the trade-offs. How does one interpret the effect size of a complex interaction? How does one conceive of it?

The computer cannot stop an individual from running a multivariate analysis of 15 variables with 50 subjects. It cannot stop a faculty member from demanding complex statistical analyses in all dissertations. Until we get beyond the idea that "more is better," the problem will remain unsolved.

Recommendations

The present study was limited in scope to Ph.D. dissertations in a specific year and one area of study. Since the sample was both quite recent and national in scope; and since it supports the findings concerning published research that have accumulated over the past 28 years, generalizations about the current state-of-the-art do not appear unwarranted.

The area of educational psychology was selected because it is a research-oriented field; it may be that power levels

are different in degree programs in fields such as counseling that have a greater orientation toward developing practitioners. A study of power levels for students completing the Ed.D. or Ph.D. degrees may also find a difference in power levels.

It might be expected that power levels would be lower in practitioner-oriented programs. It has been established that low-powered research is a problem, and this investigator is not at all sure that much more would be gained by further description of the problem. Frankly, it was anticipated that power in dissertation research in educational psychology would be higher than that reported in power analyses of published research. The recency of the research and training of the investigators were two variables that, it was thought, might lead to higher power in the material reviewed for this study. Further investigation should probably not address whether or not the problem of low power exists, but rather the factors in the training of researchers which contribute to the problem.

Delineation of the typical training program regarding significance testing and power analysis would help in understanding the problem. A survey of universities granting graduate degrees in the educational and behavioral sciences might be an initial step. A survey of graduate departments in education and the behavioral sciences requesting information on the scope and sequence of required

graduate research courses for various degree programs would provide a list of courses. Appropriate course descriptions should then be obtained from university catalogues or directly from instructors.

A similar survey could be made of the texts most commonly used in graduate research and statistics courses. The problem could then be defined in terms of the availability of information on power analysis (and effect size) and the emphasis it is given in courses and textbooks.

The problem of statistical analyses having low power should not be of concern only to research specialists since publication of nonsignificant findings and the incorrect conclusions drawn can have a dramatic effect on social policy. The U.S. Department of Education's Regular Education Initiative in lieu of special education programs is only one example. We need to find better ways to teach the concepts related to power analysis and devise easier methods of calculating or estimating power. Journal editors and readers should not accept research articles that lack adequate power due to the potential for misinterpretation. No simple solution to the problem is evident; it has taken over twenty-five years to define its scope and significance. The slow processes of technological advancement and development of new methods of teaching likely hold the keys to the solution.

REFERENCES

Abelson, R.P. (1985). A variance explanation paradox: When a little is a lot. Psychological Bulletin, 97, 129-133.

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

Bonett, D.G. (1982). On post-hoc blocking. Educational and Psychological Measurement, 42, 35-39.

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., & Sindelar, P.T. (1988). Adequate sample size: A priori and post hoc considerations. Journal of Special Education, 21(4), 74-84.

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.

Christiansen, J.E., & Christiansen, 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. (1973). Statistical power analysis and research results. American Educational Research Journal, 10, 225-229.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (revised edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, S.A., & Hyman, J.S. (1979). How come so many hypotheses in educational research are supported? (A modest proposal). Educational Researcher, 8(11), 12-16.

Cooper, H.M. (1984). The integrative research review: A systematic approach. Beverly Hills, CA: Sage Publications.

Daly, J.A., & Hexamer, A. (1983). Statistical power in research in English education. Research in the Teaching of English, 17, 157-164.

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.

Edwards, A.L. (1950). Experimental design in psychological research. New York: Rinehart.

Fagley, N.S., & McKinney, I.J. (1983). Reviewer bias for statistically significant results. Journal of Counseling Psychology, 30, 298-300.

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

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

Hopkins, K.D., Coulter, D.K., & Hopkins, B.R. (1981). Tables for quick power estimates when comparing means. Journal of Special Education, 15, 389-394.

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.

Judd, C.M., & Kenny, D.A. (1981). Estimating the effects of social interventions. Cambridge, England: Cambridge University Press.

Keselman, J.C., & Keselman, H.J. (1987). Detecting treatment effects in educational research. Educational and Psychological Measurement, 47, 903-910.

Kish, L. (1959). Some statistical problems in research design. American Sociological Review, 24, 328-338.

Kraemer, H.C. (1985). A strategy to teach the concept and application of power and statistical tests. Journal of Educational Statistics, 10, 173-195.

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

Lipsey, M.W. (1990). Design sensitivity: Statistical power for experimental research. Newberry Park, CA: Sage Publications.

Lipsey, M.W., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of science. New Directions for Program Evaluation, 27, 7-28.

Luftig, J.T., & Norton, W.P. (1982a). Improving your hypothesis testing: Type II error and power. Journal of Studies in Technical Careers, 4, 1-12.

Luftig, J.T., & Norton, W.P. (1982b). Improving your hypothesis testing: Determining sample sizes. Journal of Studies in Technical Careers, 4, 107-115.

McFatter, R.M., & Gollob, H.F. (1986). The power of hypothesis tests for comparisons. Educational and Psychological Measurement, 46, 883-886.

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

Mellon, A.W. (1962). Editorial. Journal of Experimental Psychology, 64, 553-557.

Meyer, D.L. (1974). Statistical tests and surveys of power: A critique. American Educational Research Journal, 11, 179-188, 193-194.

Neyman, J., & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika, 20A, 175-240, 263-294.

Olejnik, S.F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

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

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.

Rogers, W.T., & Hopkins, K.D. (1988). Power estimates in the presence of a covariate and measurement error. Educational and Psychological Measurement, 48, 647-656.

Schmidt, F.L., Hunter, J.E., & Urry, V.W. (1976). Statistical power in criterion-related validity studies. Journal of Applied Psychology, 61, 473-485.

Sherron, R.H. (1988). Power analysis: The other half of the coin. Community and Junior College Quarterly, 12, 169-175.

Smith, M.L. (1980). Publication bias and meta-analysis. Evaluation in Education, 4, 22-24.

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance--and vice versa. American Statistical Association Journal, 54, 30-34.

Tversky, A., and Kahnman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

West, R.F. (1985). A power analytic investigation of research in adult education: 1970-1982. Adult Education Quarterly, 35(3), 131-141.

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.

Wooley, T.W., & Dawson, G.O. (1983). A follow-up power analysis of the tests used in Journal of Research in Science Teaching. Journal of Research in Science Teaching, 20, 673-681.

Zimmerman, D.W., & Williams, R.H. (1986). Note on the reliability of experimental measures and the power of significance tests. Psychological Bulletin, 100, 123-124.

APPENDICES

APPENDIX A

RECORDING INSTRUMENT

RECORDING FORM

Date of Completion: YEAR: ________ MONTH: ______

Department: __________ Institution: ______________

Author: ____

Title: ____

____

Power Analysis Conducted? ______ ES Calculated? ______

Rationale for Sample Size: _______________________

Number of Hypotheses: _____

HYPOTHESIS # ___

Statement: _____________________________________

____

Number of Tests: _____

Test # _____

Type of Test: ____________ Result: ____________
Test Statistic: __________ # Groups: ___________
IND Variables: _________ # DEP Variables: _______
Alpha: _________________ Direction: __________
N: _________ n: _________ ES: ______________
Numerator df: ___________ Denominator df: _______

Test # _____

Type of Test: ____________ Result: ____________
Test Statistic: __________ # Groups: ___________
IND Variables: _________ # DEP Variables: _______
Alpha: _________________ Direction: __________
N: _________ n: _________ ES: ______________
Numerator df: ___________ Denominator df: _______

Test # _____

Type of Test: ____________ Result: ____________
Test Statistic: __________ # Groups: ___________
IND Variables: _________ # DEP Variables: _______
Alpha: _________________ Direction: __________
N: _________ n: _________ ES: ______________
Numerator df: ___________ Denominator df: _______

APPENDIX B

DISSERTATIONS INCLUDED IN THE SAMPLE

1. Adams, J.G. (1988). The relationship between children's diabetic control and self concept, internal-external locus of control and trait anxiety. Unpublished Ph.D. dissertation, State University of New York at Buffalo.

2. Adams, L.W. (1988). An experimental study of the effectiveness of the use of personalized and semi-personalized musical recordings as a means of improving the self esteem of young children. Unpublished Ph.D. dissertation, The Union for Experimenting Colleges and Universities.

3. Alexander, N.J. (1988). The effects of grading on the number of skills students master in Algebra II. Unpublished Ph.D. dissertation, University of Denver.

4. Archer, J.A. (1988). Feedback effects on achievement, attitude, and group dynamics of adolescents in interdependent cooperative groups for beginning second language and culture study (Volumes I and II). Unpublished Ph.D. dissertation, University of Minnesota.

5. Bhattacharya, B.B. (1988). The effects of a self-monitored jogging program on anxiety and personality. Unpublished Ph.D. dissertation, University of Illinois at Urbana-Champaign.

6. Bockoven, J.N. (1988). Towards an effective and efficient treatment for minimally and moderately distressed couples: A comparison of modeling and couples enrichment. Unpublished Ph.D. dissertation, University of Oregon.

7. Boyce, T.N. (1988). Psychological screening for high-risk police specialization. Unpublished Ph.D. dissertation, Georgia State University.

8. Burke, K.L. (1988). The effect of a perceptual cognitive training program on attention/concentration style and performance of the tennis service. Unpublished Ph.D. dissertation, Florida State University.

9. Clements, L.M.S. (1988). An investigation of training within the supervisory dyad. Unpublished Ph.D. dissertation, University of Missouri-Columbia.

10. Coviello, D.M. (1988). Differential effects of elaboration techniques on individual differences in learning from instruction. Unpublished Ph.D. dissertation, State University of New York at Buffalo.

11. D'Ambrosio, A. (1988). The effects of levels of encoding and organizational processes on children's recall of words. Unpublished Ph.D. dissertation, Fordham University.

12. Gajria, M.L. (1988). Direct instruction of a summarization strategy: Effect on text comprehension and recall in learning-disabled students. Unpublished Ph.D. dissertation, Pennsylvania State University.

13. Grissom, J.B. (1988). Structural equation modeling of retention and overage effects on dropping out of school. Unpublished Ph.D. dissertation, University of Colorado at Boulder.

14. Heerman, J.A. (1988). Problem-solving instruction: Examining its effect on the control of childhood asthma. Unpublished Ph.D. dissertation, University of Nebraska-Lincoln.

15. Kilmer, J.R. (1988). Relationship of caning to internal-external locus-of-control among selected African secondary and college students (corporal punishment). Unpublished Ph.D. dissertation, Andrews University.

16. Lancaster, J.W. (1988). The effects of single versus multiple stress management techniques and life stress on the production of creative responses in intermediate-level school children. Unpublished Ph.D. dissertation, Mississippi State University.

17. Leroux, M.D. (1988). Global beliefs and types of assertiveness training as predictors of assertiveness. Unpublished Ph.D. dissertation, University of Houston.

18. Li, R. (1988). The effects of activating schemata at different structural levels on high-schools students' retention and comprehension of a narrative passage. Unpublished Ph.D. dissertation, Florida State University.

19. Meinhold, P.M. (1988). Relating measures of maternal responsibility to selected aspects of infant learning and affect. Unpublished Ph.D. dissertation, University of North Carolina at Greensboro.

20. Naumann, J.J. (1988). The effects of teacher expressiveness, physical attractiveness and self-disclosure on student ratings of teaching. Unpublished Ph.D. dissertation, Fordham University.

21. Neville, M.E. (1988). The effect of information encoding on analogical problem solving. Unpublished Ph.D. dissertation, University of Georgia.

22. Nichols, T.M. (1988). Effects of problem solving strategies on different ability levels. Unpublished Ph.D. dissertation, University of Alabama.

23. Reekie, F.A. (1988). A quasi-experimental investigation of the functional curriculum and learning strategies approaches with low-achieving and learning-disabled early adolescents. Unpublished Ph.D. dissertation, University of Saskatchewan, Canada.

24. Reese, C.M. (1988). The effect of grade level and text type on comprehension monitoring. Unpublished Ph.D. dissertation, Pennsylvania State University.

25. Reimer, W.L. (1988). The effects of the Tribes program on self-esteem and academic achievement. Unpublished Ph.D. dissertation, University of Toledo.

26. Rothenberg, J.L. (1988). Preschoolers' reaction to the birth of a sibling. Unpublished Ph.D. dissertation, University of Minnesota.

27. Salas, S.B. (1988). The effects of three different corrective procedures on concept learning using computer assisted instruction. Unpublished Ph.D. dissertation, University of Tennessee.

28. Santiago, A.D. (1988). A study of the effect of jogging on inferred self-concept in a group of college-age return-migrant and non-migrant females. Unpublished Ph.D. dissertation, Pennsylvania State University.

29. Schmidt, G.D. (1988). Concurrent validity of the scale of adolescent adaptive behavior. Unpublished Ph.D. dissertation, Pennsylvania State University.

30. Stephenson, S.U. (1988). The effect of a cognitive-behavioral course in assertiveness training procedures on internalized shame in college students. Unpublished Ph.D. dissertation, Oregon State University.

31. Vandyne, E.E. (1988). A study exploring the use of the class "Increasing Your Life Satisfaction" with community elderly people. Unpublished Ph.D. dissertation, Ohio University.

32. Walker, R.L. (1988). Transfer of a lecture notetaking skill. Unpublished Ph.D. dissertation, Texas A&M University.

33. Yamin, S.B. (1988). Frequency of testing and its effects on achievement, test anxiety, and attitudes towards science of students at University Technology of Malaysia. Unpublished Ph.D. dissertation, Oregon State University.

34. Youn, G. (1988). The magnitude of the Mueller-Lyer illusion as a function of hue, saturation, fundus pigmentation, and simulated aging. Unpublished Ph.D. dissertation, University of Georgia.

35. Zikmund, A.B. (1988). The effect of grade level, gender, and learning style on responses to conservation type rhythmic and melodic patterns. Unpublished Ph.D. dissertation, University of Nebraska-Lincoln.

VITA

Kathleen E. McKean

Candidate for the Degree of

Doctor of Philosophy

Thesis: STATISTICAL POWER ANALYSIS OF DOCTORAL DISSERTATION RESEARCH IN EDUCATIONAL PSYCHOLOGY

Major Field: Applied Behavioral Studies

Biographical:

Personal Data: Born in Indianapolis, Indiana, February 24, 1954, daughter of C. R. and Joan C. McKean.

Education: Graduated from Pryor High School, Pryor, Oklahoma, in May, 1972; received Bachelor of Arts degree (Honors) from Oklahoma State University in May, 1976; received Master of Science degree from Oklahoma State University in July, 1979; completed requirements for the Doctor of Philosophy degree at Oklahoma State University in July, 1990.

Professional Experience: School Psychologist, Oklahoma Child Service Demonstration Center, Cushing, Oklahoma, 1979-85; Assistant Director, Project Director, Grant Writer, Program Evaluator, Oklahoma Child Service Demonstration Center, 1985 to present.