INFORMATION TO USERS

This manuscript has been reproduced 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.

UMI®

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

STATISTICAL POWER AND EFFECT SIZE IN THE FIELD OF HEALTH PSYCHOLOGY

BY

JASON EDWARD MADDOCK

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

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

THE UNIVERSITY OF RHODE ISLAND

1999

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

UMI Number: 9945214

UMI Microform 9945214
Copyright 1999, by UMI Company. All rights reserved.

This microform edition is protected against unauthorized copying under Title 17, United States Code.

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

DOCTOR OF PHILOSOPHY DISSERTATION

OF

JASON EDWARD MADDOCK

APPROVED:

Dissertation Committee

Major Professor

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

1999

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ABSTRACT

Statistical significance testing is one of the most pervasive techniques in psychology to examine treatment effects. Because of the ubiquitous use of these procedures, misuses have plagued the field of psychology (Harlow, 1997). Some psychologists have even suggested the banning of significance testing in psychological research (Schmidt & Hunter, 1997), though most researchers argue for an improvement of current methods (Cohen, 1994; Rossi, 1997; Mulaik, Raju & Harshman, 1997). In the first chapter, the significance test controversy is discussed in detail, and the general disregard for statistical power in most psychological research is discussed as a major contributor to this controversy (Cohen, 1988). Although statisticians have long acknowledged the problem of statistical significance testing, little improvements have occurred in the last 30 years. This section ends by discussing four methods for improving significance testing: the use of confidence intervals, testing for probable upper bounds, meta-analysis, and power analyses. Each of these methods is explored because they offer simple ways in which significance testing can be improved without great resistance.

In chapter two, power was calculated for 8,266 statistical tests in 187 journal articles published in the 1997 volumes of Health Psychology, Addictive Behaviors, and Journal of Studies on Alcohol. Power to detect small, medium and large effects was .34, .74, and .92 for Health Psychology, .34, .75, and .90 for Addictive Behaviors, and .41, .81, and .92 for the Journal of Studies on Alcohol. Mean power estimates are .36, .77, and .91, giving a good estimation for the field of health psychology. Cohen (1988) recommended that power to detect effects should be approximately .80. Using this criteria, the articles in these journals have adequate power to detect medium and large

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

effects. Comparison of these results to over 30 other power studies in fields varying from occupational therapy to teaching science indicate that health psychology journals rank among the highest in power. The three journals selected in this study also have the highest power among psychology journals. Results are encouraging for this field, although studies examining small effects are still very much underpowered. This issue is important since most intervention effects in health psychology are small.

In chapter three, a meta-analysis of interventions to reduce college student drinking was conducted. Qualitative analyses examining these interventions have produced conflicting results with some studies reporting significance and some not. Twenty-one studies met criteria to be included in the meta-analyses. This criteria included use of random assignment or statistical control for baseline differences. Results indicated that the studies when examined as a group significantly reduced drinking in college students ($p < .05$). However, cognitive-behavioral interventions ($d = .53$) produced significantly larger effects than traditional educational approaches ($d = .17$), indicating the superiority of this type of intervention. Power analysis of these articles revealed inadequate power, demonstrating the need to use higher powered studies to reduce controversial findings.

These three chapters demonstrate the need for adequate power in the field of health psychology across journals and also indicate the utility of performing meta-analyses to synthesize the findings for individual research area. Implications of improving significance testing are discussed.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ACKNOWLEDGMENT

Any time an endeavor such as this takes place, many people are responsible for the timely completion of the project. With this in mind, I would like to thank first and foremost, Dr. Joseph Rossi, my major professor. His staunch ideals for improving the way that psychologists analyze data has been an inspiration for this work. Anyone who is familiar with Joe's dissertation will quickly see the framework upon which this dissertation is based. His work has served as a foundation for this and hopefully future efforts.

Next, I would like to thank the rest of my committee. Dr. Robert Laforge, Dr. Colleen Redding, and Dr. Cynthia Willey Lessne have all provided consistent guidance and helpful insights during this project. Dr. Mark Wood and Dr. James Campbell who have served as defense chairs have also been quite helpful in shaping this dissertation.

Finally I would like to thank my social support network. My parents have always stood behind me and believed in my work. Lisa Biederman who has kept me happy and helped me get through all of this. All of my friends at the CPRC, who have made graduate school a truly enjoyable time of my life.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

PREFACE

This dissertation has been prepared in manuscript format.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE OF CONTENTS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

LIST OF TABLES

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

LIST OF FIGURES

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 1: Significance Testing as a Limitation to Cumulative Knowledge in the Social Sciences

"*Significance testing never makes a useful contribution to the development of cumulative knowledge.*" –Schmidt and Hunter, 1997, pg. 57

"*Significance tests fill an important need in answering some key research questions, and if they did not exist they would have to be invented.*" – Abelson, 1997, pg. 117.

Introduction

Statistical significance testing is one of the most pervasive techniques in the social sciences to examine treatment effects. Because of the ubiquitous use of these procedures, misuses have plagued the field of psychology (Harlow, 1997). The debate over the use of significance tests has concerned psychologists for well over 30 years and shows no sign of resolution (Bakan, 1966; Morrison & Henkel, 1970; Harlow, 1997). Some psychologists have even suggested the banning of significance testing in psychological research (Schmidt & Hunter, 1997), though most researchers argue for an improvement of current methods (Cohen, 1994; Rossi, 1997; Mulaik, Raju & Harshman, 1997). As the two quotes above demonstrate, many statistically sophisticated scientists are on polar ends of this debate. This paper will outline many of the reasons behind this debate and provide four simple measures that social scientists can implement to improve the way data are currently analyzed.

History of significance testing

Null hypothesis significance testing (NHST) was developed by Fisher in the early 1920's. His influential handbook on statistical inference, Statistical Methods for Research Workers, was widely adopted and highly influenced the way that data has been

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

analyzed for the past 70 years (Fisher, 1928; 1932; Aron & Aron, 1999). However, his methodology did not include the concepts of power and type II error. Beginning in the late 1920's Neyman and Pearson recognized the importance of type II error and power in the social sciences. They subsequently introduced these concepts in a landmark series of papers over the next decade (Neyman & Pearson, 1928a, 1928b, 1933a, 1933b, 1936a, 1936b, 1938, 1967). Fisher debated furiously with Neyman and Pearson over these concepts and refused to accept the idea of type II error. It was not until after World War II that Neyman and Pearson's views became widely known and integrated into mainstream psychology. Even at this time, statistic textbooks introduced the concept without admitting that Fisher's conceptualization had been limited, in an effort to present psychology as a flawless, mechanical method of decision making (Gigerenzer & Murray, 1987). Since this time, the practice of NHST has been routinely attacked and defended (see Morrison & Henkel, 1970 for over 30 articles debating NHST). What is most shocking about the debate over NHST is the lack of progress that has been made in reforming these procedures. In the Morrison and Henkel text (1970) which consists almost entirely of reprinted articles, many from the 1950's, the salient issues are essentially the same as presented in Harlow, Mulaik, and Stieger's (1997) compilation. As Bakan (1966) states in his seminal article on the issue, "What is said in this paper is hardly original. It is, in a certain sense, what everybody knows." (pg. 231). This was written 33 years ago yet contains almost exactly the same arguments as today's writers. The next section of this paper will examine the limitations of NHST and explore reasons why it has been so resistant to change.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Limitations of NHST

Since NHST is widely disseminated and used throughout the social sciences, it is like any tool, bound to be misused. Many authors have examined the general misuse of NHST, and their texts should be examined as they annotate many common pitfalls (ie. Bakan, 1966; Cohen, 1994). However, the defenders of NHST routinely note that the technique should not be replaced because it is misused, but instead only if the technique is inherently flawed (Mulaik, Raju & Harshman, 1997). This paper is instead concerned with the limitations of NHST when used in a way that is commonly accepted by most researchers in the social sciences, not with isolated instances of misuse or faulty interpretations. As many critics demonstrate, even when used correctly, NHST is an impediment to the development of a cumulative science. Several issues will be discussed including lack of statistical power in most research designs, atheoretical use of one-tailed tests, over-reliance on p-values, and lack of information given in the typical significance test.

To explain the NHST controversy most efficiently, I will use the example of the t-test for independent means, although the logic I will use can be extended to other forms of NHST. In this scenario, two groups are (randomly) selected from a population and one group serves as the control group while the other is given a treatment and becomes the experimental group. For instance, one group of school children get a special training course to raise their IQs called the SMART program, while the other group gets no special treatment. After the treatment is conducted, the researcher wishes to examine whether the treatment group has a higher IQ than the control group, so she carefully collects the IQ scores from each of the children and finds that the treatment group has a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

mean IQ which is 5 points higher than the control group. The nature of NHST is that the researcher is interested in seeing how likely it is that the treatment group was drawn from the same population as the control group. She believes that if it was only possible 5 out of a hundred times that the groups came from the same population, she could be pretty sure that the treatment had worked. She is also only concerned if the treatment increases the IQ of the children so she conducts a one tailed t-test and finds that the mean of the treatment group exceeds the critical value and she rejoices, rejects the null hypothesis, and retains her alternative hypothesis that the treatment worked. This doesn't seem all that bad. After all she had to make a decision regarding her treatment and this gives her a good way to do it, right? Well, let's consider two other scenarios. The researcher's young, naive graduate student comes by and finds the data sitting on the computer, but no analyses. The researcher was planning on presenting the data at a national conference and wanted the graduate student to prepare the slides. The graduate student has taken the introductory ANOVA class and knows very well that he better set his alpha level at .05 so he goes ahead and conducts the significance test. However, his statistics professor always told him to conduct two-tailed tests to allow for findings in the opposite direction, so he conducts a two-tailed test and much to his dismay he finds no significant results, with exactly the same data. Luckily, the researcher sees the graduate student's data before the conference and is able to change the analysis back to the original one. She would have been quite the laughing stock after entitling her talk, "An easy way to make your child SMARTer." After all there was statistically significant evidence that demonstrated that this treatment made children smarter. So she presents her findings at the conference and 10 researchers in the field hear the talk and decide to replicate her

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

experiment. After all replication is the mainstay of science. They all report back to the conference the following year with much hullabaloo. It seems that 4 of them found statistically significant results, and entitled their talks "Make your kid SMART" and six did not and entitled their talks, "SMART, not!" What was causing all this controversy? All of the researchers had used one tailed tests and the same number of subjects so this wasn't the answer. What the researchers didn't know was that the power to detect this effect for the number of subjects they had was .40. So they got exactly what they should expect. A small effect existed in all of these studies and sometimes p was less than .05 and sometimes it was not. This little tale clearly outlines the problem of NHST as it is used today. It illustrates the problems that arise from not enough power, to the little direction that is given on whether to conduct one or two-tailed tests, and even to the need for effect sizes to gauge the importance of the intervention. For instance, what would have happened if all of the researchers knew about power and conducted their studies with a 1000 subjects each, and amazingly all 10 got statistically significant results. This could lead to a national implementation of the SMART treatment only to find out years later that the treatment only increased IQ's 2.5 points and cost $200 per student. Not exactly clinically significant or cost effective. As one can see from these examples, NHST in its current form is not as strongly objective as it may seem and highly susceptible to small alterations in the data. It also provides the researcher with very little information regarding the importance of the findings.

One of the main problems with significance testing is the general disregard for statistical power in most psychological research (Cohen, 1988). Lack of power in experimental research leads to controversial findings (Rossi, 1990). Although replication

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

is seen as the hallmark of good science, under-powered studies almost guarantee controversial results. The problem of type I errors, rejecting the null hypothesis when it is true, has long been acknowledged and guarded against by setting the α level to a pre-set value (usually .05 or lower; Fisher, 1949). This limits the type I rate but ignores the problem of type II error rates. Type II error rates are driven by the lack of statistical power of a test (Bakan, 1966; Chase & Tucker, 1976; Cohen 1977, 1988; Cowles & Davis, 1982; Hogben, 1957). If a test is not powerful enough to detect an effect, the effect will be rejected when it really exists. The four possible outcomes of statistical decisions are displayed in Table 1. Even though power has long been acknowledged by statisticians (Neyman & Person, 1928a, 1928b, 1933a, 1933b, 1936a, 1936b, 1938, 1967) it is still underused today (Rossi, 1997). Cohen (1988) has suggested that statistical power should be .80 or greater to detect effects. Unfortunately, repeated power analyses have revealed that the average power for medium effects among published studies in psychology is much closer to .50 (Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). No one is sure why power has not increased. The lack of interest in statistical power has contributed to difficulty or inaccessibility with standard power material, slow dispersion of new ideas, and the legacy of the Fisherian paradigm (Cohen, 1992). The lack of statistical power becomes a problem when the results of the experiment are not extreme enough to reject the null hypothesis. In low powered studies, it is not possible to conclude that the study had no effect since the type II error rate in psychology is so large. This leads to several problems including the publication of null results which lead to controversy over the existence of an effect, and when studies are not published, the file

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

drawer problem where authors are not able to publish null results. This is becoming more of a problem with the explosion of meta-analysis techniques in recent years. Since unpublished studies are more difficult to find, meta-analyses tend to draw mostly on published material. This can easily lead to an inflation in the true effect sizes given that studies that found larger effects are more likely to be significant and therefore published.

As outlined above, statistical power in psychology is frighteningly low. However, it does not appear that this necessarily has to be the case. Before 1969, standard material for calculating power was sparse and researchers had to be able to extrapolate power from the non-centrality parameter. With the publication of Cohen's (1969) text, all of this changed dramatically. This text included hundreds of tables for calculating power for all of the most common tests including the t-test, Pearson r, z test for differences between correlation coefficients, sign test, the test that a proportion is .50, z test for differences between proportions, chi-square test for goodness of fit, and the F test for the analysis of variance and covariance. In the second edition of the book (Cohen, 1977), F test in multiple regression/correlation analysis was added. However as Rossi (1984) noted 15 years ago much of the knowledge of statistical power has not 'trickled down" to the average behavioral researcher, but instead remained in the domain of the mathematically sophisticated psychologists.

In more recent years though, computer programs, such as Power and Precision (Borenstien, Cohen, & Rothstein, 1997) have been developed. These programs are Windows based and feature easy to use pull down menus and instructions. With the development of these texts and programs, power calculations are easily accessible to any social scientists who is reasonably knowledgeable about quantitative methods. An

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

informal survey of undergraduate statistics textbooks revealed that everyone at least included some mention of statistical power (Aron & Aron, 1999; Gravetter & Wallnau, 1996; Minium, Clarke & Coadarci, 1999; Spence, Cotton, Underwood & Duncan, 1990) Although only one (Aron & Aron, 1999) provided methods for computing power for all of the statistical tests presented. The abundance of statistical power material indicates that knowledge about statistical power or the ability to calculate power should not be an impediment to performing high powered studies. Since knowledge of statistical power is probably not the cause of low statistical power, it is best to examine what the possible impediments for designing high powered studies are.

Power consists of only three variables: significance criterion, sample size, and effect size (Cohen, 1992). All of these areas are under direct control of the researcher.

The significance criteria or alpha level is set a priori to the study and is traditionally .05 or less. Psychologists tend to regard this number with magical significance. Beware the researcher who violates this proscription. However, as Cowles and Davis (1982) point out the arrival at this convention is quite arbitrary and has little mathematical significance. Unfortunately, since this convention is unlikely to change, this dissertation will instead examine the problem of using even smaller alpha conventions. Researchers are often so worried about type I error that they are willing to reduce the alpha level to .001. As Cohen (1988) points out this can reduce power of .80 with alpha at .05, to power of .10 at alpha .001. The ratio of beta to alpha then becomes 900 to 1. It is difficult to think of an area where rejecting the null hypothesis when it is true is 900 times more important than accepting the null hypothesis when it is false! Cohen (1988) recommends an alpha level of .05 and power of .80 which results in a 4 to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1 ratio, a much more realistic assumption. Examination of the significance criterion for power indicates that little flexibility is given to the researcher due to long standing conventions regarding the alpha level.

Sample size is the variable that is most under the control of the researcher. The larger the sample size is, the smaller the error and the greater the reliability of the sample. The smaller the error of the sample is, the greater the power is. While some populations are difficult to recruit into studies it is often necessary to increase sample size as much as possible to ensure adequate power to detect effects. Increased sample size also increases the reliability of sample results. The reliability of the sample relates to how well the sample can approximate the relevant population value. Hence, the more reliable the sample the greater the power, since differences between the populations are more easily detected. Reliability can be controlled in two ways: through experimental design and manipulation of sample size. Experimental design is often an under-examined area of power analysis. This is true because it is not quantifiable a priori. The effect of experimental design is simple, whatever reduces within-group variation will increase power. However, this is often not desirable since highly homogeneous groups are often not generalizable to larger populations. The recommendation for this area is to use the strongest experimental design that is applicable for the research question.

The final variable entered into the power equation is effect size. "Effect size is the degree to which a phenomenon is present in a population" (Cohen, 1988, p. 9). The null hypothesis posits an effect size of 0. Deviations from 0 indicate the effect size in the population. The greater the effect size the greater the power is. Also larger effect sizes indicate the need for smaller sample sizes without loss of power. Effect size is often the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

stopping point for many would be power calculators. How does one know the effect size before a study is conducted? Simply, one doesn't. However it can often be estimated from theory and other studies in closely related areas. If this is not possible, Cohen (1988) has constructed a series of effect sizes for each of the statistical tests in this book. For each of these tests he has constructed effect sizes that correspond to "small", "medium" and "large" effect sizes. While these are just conventions, they do allow the researcher some guidelines when conducting power analyses. While effect sizes are not generally alterable, the researcher is at liberty to decide which effects to examine. If the researcher knows that the effect is expected to be small, a design can be selected that will be powerful enough to detect these size effects. Conversely, in an area where a large effect is expected, scarce resources can be saved by employing the appropriate number of subjects.

While effect sizes are essential to psychological research, the size of the effect differs in importance depending on the aims of the study. The researcher also has to examine the clinical significance of the finding. A small effect in medicine, that is inexpensive and saves many lives, such as the use of aspirin to prevent heart attacks may be important while a larger effect that is expensive to implement may not be justified (Kazis et al., 1989; Rosenthal, 1990). A study using a large population may have a statistically significant effect, but may not be clinically significant. Conversely, a result may be found to be not statistically significant, but the effect size might justify the need for a larger study to analyze the possibility of an important finding (Deyo & Patrick, 1995). Also, as Prentice and Miller (1992) have pointed out, small effects are impressive when minimal interventions are used or when it seems unlikely that a dependent variable

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

will be influenced by an independent variable. It is essential for the researcher to weigh the costs of both types of errors before selecting the alpha level and the desired power (Overall & Dahal, 1965). For these reasons, effect size is important for anyone who designs a research study.

Methods for Improving NHST

The need for reforming traditional NHST is certainly widespread in the field of psychology. Many authors such as Hunter and Schmidt (1997) have outlined plans for reforming NHST. Hunter and Schmidt's (1997) approach would ban NHST and use every study as a single data point for meta-analyses, with no interpretations or discussion made for individual studies. Although this is an intriguing idea, it is unlikely to occur in the near future due to the slow change of NHST in the last 3 decades. While there has been a large resistance to reforming significance testing over the past three decades, there are several ways that significance testing can be improved without radically changing the way that data is analyzed and without the need to totally retrain researchers in statistical methods. These four simple methods for reforming significance testing are: the use of confidence intervals, testing for probable upper bounds when the null hypothesis cannot be rejected, power surveys of the published literature and meta-analytic techniques for combining results. Each of these methods for improving significance testing will be described in turn.

When a significance test is conducted and the results reveal that $p < .001$, this tells us very little. All we know from this is that it is very unlikely that our treatment group came from the same population as our control group. Hence, our treatment is supported. However, as Meehl (1967) showed, almost everything in the social sciences is related to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

everything else to some degree. With large enough sample sizes, we are almost guaranteed to get statistical significance due to what had been called the "crud factor"(Meehl, 1967). A classical example of this is ask undergraduate students, which they would be more impressed with, results that indicated that $p < .01$ with 20 subjects or with 200 subjects and they will undoubtedly answer 200 because it is more representative of the population. This is where effect size is very important, since it is independent of sample size. Ask the same students which they would be more impressed with, a study that indicated that $d = .50$, with 20 students or 200 students. The original answer is now correct. Effect size gives us an idea of how much separation there is between the two groups. It is also necessary when determining clinical significance compared to statistical significance. For example, if we conducted a country-wide trial of the SMART program and found that it increased the average IQ by 2 points this would be significant at some p value for example .000001. However, does an increase of 2 IQ points mean that little Jimmy is going to Harvard instead of Nowhereville Community College? Highly unlikely. So as you can see from this example, significance testing with its dichotomous decision making often gives us incorrect information when power is low and gives us too little information when power is high. The use of confidence intervals (CI) provide an easy method for improving the knowledge gleaned from a significance test. Not only does a 95% confidence interval provide the same information as a traditional NHST, but it also includes a measure of how large the effect is and an error band around that effect. CIs make NHST much more informative than just the dichotomous decision that is usually discovered in NHST, while still allowing the researcher a measure for making

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

statistical decisions. However it should be noted that adequately powered studies are still needed to avoid the same trap that NHST has fallen into.

In addition to CIs, probable upper bounds can be tested for when the null hypothesis cannot be rejected. Rossi (1990) has also outlined how we can test for probable upper bounds when a result is not found to be significant. In this scenario, say we conduct another trial of the SMART program, but this time the researcher is tired of getting inconsistent results so she decides to get a bigger sample size of, say 200 a group. This person again finds non-significant results, but this time tests for probable upper bounds. By using her sample size of 200 and $\underline{d} = .50$, she discovers that power = .99, so she rules that out and proceeds to the next effect size. Always the careful fellow, she stops at $d = .35$, here power is = .85. She believes that she can be fairly confident that the effect of the SMART program is less than $d = .35$ and writes an article suggesting that any researcher interested in examining the effect of the SMART program should design a study with a sample size large enough to detect an effect smaller than .35. Thus, a probable upper bound of the SMART program has been discovered and an acknowledgment of the probable effect size is narrowed in on. This could have several practical implications including selecting sample size for future research or suspending a research program because even if an effect was found it would not be clinically significant. Since a researcher cannot prove the null hypothesis and results that do not reach a critical value are regarded are inconclusive, this methodology will improve the interpretability of null results.

A third method for improving NHST would be to conduct power surveys of published literature in relevant fields. Little was known about the power of published

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

articles in the field of psychology until Cohen's (1962) groundbreaking study on the power of articles published in the Journal of Abnormal and Social Psychology. Cohen discovered that power for medium effects was unacceptably low with a mean of .48. After this study and the publication of Cohen's (1969) book on statistical power, power analyses grew exponentially. Power analyses have been done in dozens of areas in the twenty-five years following the first power study in dozens of fields (Cohen, 1988). Rossi (1990) replicated Cohen's original study using the 1982 editions of the Journal of Personality and Social Psychology and the Journal of Abnormal Psychology. The results of this analysis indicated a small increase in power for medium effect size (M = .59) which was still inadequate. Since the publication of Cohen's (1962) first study, effect size estimates for small, medium, and large effects have been readjusted downward indicating that these power studies may have inflated the power estimates for these journals! Rossi (1984) used Cohen's (1977) revised estimates to examine the power of the articles in the 1982 editions of the Journal of Personality and Social Psychology and the Journal of Abnormal Psychology Using these estimates, the average power for medium size effects was only .55.

Power analyses have been performed in many areas including: abnormal psychology (Sedlmeier & Gigerenzer, 1989), animal behavior (Thomas & Juanes, 1996), applied psychology (Chase & Chase, 1976). communication (Katzer & Sodt, 1973; Chase & Tucker, 1975), clinical trials (Freiman, Chalmers, Smith, & Kuebler, 1978; Davis, Janicak, Wang, & Gibbons, 1992), counselor education (Haase, 1974), education (Brewer, 1972; Jones & Brewer, 1972, Pennick & Brewer, 1972; Wooley & Dawson, 1983; Daly & Hexamer, 1983; Sindelar, Allman, Monda, & Vail, 1988), educational

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

measurement (Brewer & Owen, 1973), medical education, (Wooley, 1983), occupational therapy, (Ottenbacher, 1982), HIV transmission prevention research (Kalichman, Carey & Johnson, 1996), personnel selection (Katzell & Dyer, 1977), market research (Sawyer & Ball, 1981), physical education (Christensen & Christensen, 1977), speech pathology (Kroll & Chase, 1975), social work (Crane, 1976; Judd & Kenny, 1981; Orme & Tolman, 1986) and vocational evaluation research (Kosciulek, 1993). These analyses indicated that power varies widely between different fields and subspecialties within psychology. Knowledge of power for more specific areas of psychology provide the researcher with an index of confidence about the research. While many of these analyses have been performed they are fairly rare within psychology. With the number of sub-disciplines in psychology growing rapidly, it is important to examine levels of power within varied areas.

The fourth area of improvement for NHST is in the increased use of meta-analytic techniques for combining research results. As noted above, lack of power in individual studies has in part lead to psychology's failure as a cumulative science (Rossi, 1997). Studies that had inadequate power were often found to have non-significant effects even when they really existed. Traditional narrative reviews helped to increase the debate over these inconsistencies by using "head counting" methods instead of impartial, statistically controlled methods. Meta-analysis provides a much more powerful way to examine effect sizes in specific areas of research (Rosenthal, 1991). The use of meta-analysis has exploded over the last two decades and is widely accepted by many researchers (Hunter & Schmidt, 1990; Bailar, 1997), although there are still several researchers who question the utility of meta-analysis (Abelson, 1997; LeLorier, Gregoire, Benhaddad, LaPierre, &

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Derderian, 1997). Although meta-analytic techniques still have some methodological flaws due to differences in primary studies, this technique is quite powerful for synthesizing research results. Error bands which are quite large in individual studies are greatly reduced in meta-analysis. This technique allows the researchers the ability to calculate the "true effect" of an area within psychology much better than an individual study. Moderators of effects can also be examined to see which factors are relevant in increasing or decreasing effect size.

NHST is a controversial, yet widely used practice in psychology. Over the last three decades this practice has come under numerous criticisms, but the technique has been fairly impervious to change. Much of the controversy has arisen from the lack of statistical power typically demonstrated in the social sciences. This lack of power leads to inconsistent findings and debate within the field. This paper has suggested four ways in which NHST can be improved within radically changing the way that data is analyzed. These methods are the use of confidence intervals, testing for probable upper bounds when the null hypothesis cannot be rejected, power surveys of the published literature and meta-analytic techniques for combining results. If these methods are assimilated into mainstream psychology, the conclusions derived from psychological research will be much more consistent and help the field to develop into a more cumulative science.

In the following sections, two of the four methods in which NHST can be improved will be examined in the field of health psychology. In the first study, a power analysis was conducted on three journals, Health Psychology, Addictive Behaviors, and Journal of Studies on Alcohol. Results from this study provide researchers with an indication of the average power across the journals for small, medium and large effect

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

sizes. This information is essential for researchers designing future studies and for assessing the reliability of present findings. In the second study, a meta-analysis on individual interventions to reduce college student drinking was performed. Results from this study indicate that overall, individual interventions to reduce college student drinking are effective. Also, cognitive-behavioral interventions are more efficacious in changing behavior than traditional educational approaches, and in general studies need more subjects to provide adequate power for research of this nature. Together these two studies build on the arguments postulated in this section by providing practical applications of the strategies outlined above.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 117-144). Mahwah, NJ, Lawrence Erlbaum Associates.

Aron, A. & Aron, E. N. (1999). Statistics for Psychology (2nd Edition). Upper Sadle River, NJ: Prentice-Hall, Inc.

Bailar, J. C. (1997). The promise and problems of meta-analysis. The New England Journal of Medicine, 337, 559-560.

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

Borenstien, M., Cohen, J. & Rothstein, H. (1997). Power and Precision. Teaneck, NJ: Biostat.

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

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

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

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Christensen, J. E. & Christensen, C. E. (1977). Statistical power analysis of health, physical education, and recreation research. Research Quarterly, 48, 204-208.

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

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

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Academic Press.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 47, 997-1003.

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

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

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

Davis, J. M., Janicak, P. G., Wang, Z. & Gibbons, R. D. (1992). The efficacy of psychotropic drugs: Implications for power analysis. Psychopharmacology Bulletin, 28, 151-155.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Deyo, R. A. & Patrick, D. L. (1995). The significance of treatment effects: The clinical perspective. Medical Care, 33, AS286-AS291.

Fisher, R. A. (1928). Statistical methods for research workers (2nd ed.). London: Oliver & Boyd.

Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). London: Oliver & Boyd.

Fisher, R. A. (1949). The design of experiments. New York: Hafner.

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

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

Gravetter, F. J. & Wallnau, L. B. (1996). Statistics for the behavioral sciences (4th Edition). New York: West Publishing Co.

Haase, R. F. (1976). Power analysis of research in counselor education. Counselor Education and Supervision, 14, 124-132.

Harlow, L. L. (1997). Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 1-20). Mahwah, NJ, Lawrence Erlbaum Associates.

Harlow, L. L., Mulaik, S.A. & Steiger, J. H. (1997), What if there were no significance tests? Mahwah, NJ, Lawrence Erlbaum Associates.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

Hunter, J. E. & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

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.

Kalichman, S. C., Carey, M. P., & Johnson, B. T. (1996). Prevention of sexually transmitted HIV infection: A meta-analytic review of the behavioral outcome literature. Annals of Behavioral Medicine, 18, 6-15.

Katzell, R. A. & Dyer, F. J. (1977). Differential validity revived. Journal of Applied Psychology, 62, 137-145.

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

Kazis, L. E., Anderson, J. J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. Medical Care, 27, S178-S189.

Kosciulek, J. F. (1993). The statistical power of vocational evaluation research. Vocational Evaluation and Work Adjustment Bulletin, 26, 142-145.

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

LeLorier, J., Gregoire, G., Benhaddad, A., LaPierre, & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large, randomized, controlled trials. The New England Journal of Medicine, 337, 536-542.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115.

Minium, E. W., Clarke, R. C. & Coladarci, T. (1999). Elements of Statistical Reasoning (2nd Edition). New York: John Wiley and Sons, Inc.

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

Mulaik, S. A., Raju, N. S. & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-198). Mahwah, NJ, Lawrence Erlbaum Associates.

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

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

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

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

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

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

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

Orme, J. G. & Tolman, R. M. (1986). The statistical power of a decade of social work education research. Social Service Review, 60, 620-632.

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

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

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.

Prentice, D. A. & Miller, D. T. (1992). When small effects are impressive. Psychological Bulletin, 112, 160-164.

Rosenthal, R. (1990). How are we doing in soft psychology? American Psychologist, 45, 775-776.

Rosenthal, R. (1991). Meta-analytic procedures for social research. Newbury Park, CA: Sage.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Rossi, J. S. (1984). Statistical power of psychological research. Unpublished doctoral dissertation, University of Rhode Island.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-198). Mahwah, NJ, Lawrence Erlbaum Associates.

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

Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-198). Mahwah, NJ, Lawrence Erlbaum Associates.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Sindelar, P. T., Allman, C., Monda, L., & Vail, C. O. (1988). The power of hypothesis testing in special education efficacy research. Journal of Special Education, 22, 284-296.

Spence, J. T., Cotton, J. W., Underwood, B. J., & Duncan, C. P. (1990). Elementary Statistics (5th Edition). Englewood Cliffs, NJ: Prentice Hall.

Thomas, L. & Juanes, F. (1996). The importance of statistical power analysis: An example from Animal Behaviour. Animal Behaviour, 52, 856-859.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Wooley, T.W. (1983). A comprehensive power-analytic investigation of research in medical education. Journal of Medical Education, 58, 710-715.

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.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 1: Possible Outcomes of Statistical Test Decisions

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 2: Power analysis of three health psychology journals

Introduction

The issue of statistical power had long been ignored in the field of psychology until Cohen's (1962) ground breaking study on the power of articles published in the Journal of Abnormal and Social Psychology. Cohen discovered that power for medium effects was unacceptably low with a mean of .48. After this study and the publication of Cohen's (1969) book on statistical power, power analyses grew exponentially. Power analyses have been done in dozens of areas in the twenty-five years following the first power study in dozens of fields (Cohen, 1988). Rossi (1990) replicated Cohen's original study using the 1982 editions of the Journal of Personality and Social Psychology and the Journal of Abnormal Psychology. The results of this analysis indicated a small increase in power for medium effect size (M = .59) which was still inadequate. Since the publication of Cohen's (1962) first study, effect size estimates for small, medium, and large effects have been readjusted downward indicating that these power studies may have inflated the power estimates for these journals! Rossi (1984) used Cohen's (1977) revised estimates to examine the power of the articles in the 1982 editions of the Journal of Personality and Social Psychology and the Journal of Abnormal Psychology Using these estimates, the average power for medium size effects was only .55.

Power analyses have been performed in many areas including: abnormal psychology (Sedlmeier & Gigerenzer, 1989), animal behavior (Thomas & Juanes, 1996), applied psychology (Chase & Chase, 1976). communication (Katzer & Sodt, 1973; Chase & Tucker, 1975), clinical trials (Freiman, Chalmers, Smith, & Kuebler, 1978; Davis, Janicak, Wang, & Gibbons, 1992), counselor education (Haase, 1974), education

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(Brewer, 1972; Jones & Brewer, 1972, Pennick & Brewer, 1972; Wooley & Dawson, 1983; Daly & Hexamer, 1983; Sindelar, Allman, Monda, & Vail, 1988), educational measurement (Brewer & Owen, 1973), medical education, (Wooley, 1983), occupational therapy, (Ottenbacher, 1982), HIV transmission (Kalichman, Carey & Johnson, 1996), personnel selection (Katzell & Dyer, 1977), market research (Sawyer & Ball, 1981), physical education (Christensen & Christensen, 1977), speech pathology (Kroll & Chase, 1975), social work (Crane, 1976; Judd & Kenny, 1981; Orme & Tolman, 1986) and vocational evaluation research (Kosciulek, 1993).

Power analyses are essential to research in psychology because they elaborate areas in which real group differences would be unlikely to be detected. If researchers were aware that they had less than a 50% chance to detect their hypothesized effects they should be less likely to perform these studies. Luckily, power is largely under the control of the researcher. Power consists of only three variables: significance criterion, sample size, and effect size (Cohen, 1992). All of these areas are under direct control of the researcher.

The significance criteria or alpha level is set a priori to the study and is traditionally .05 or less. Psychologists tend to regard this number with magical significance. Beware the researcher who violates this proscription. However, as Cowles and Davis (1982) point out the arrival at this convention is quite arbitrary and has little mathematical significance. Unfortunately, since this convention is unlikely to change, this dissertation will instead examine the problem of using even smaller alpha conventions. Researchers are often so worried about type I error that they are willing to reduce the alpha level to .001. As Cohen (1988) points out this can reduce power of .80

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

with alpha at .05, to power of .10 at alpha .001. The ratio of beta to alpha then becomes 900 to 1. It is difficult to think of an area where rejecting the null hypothesis when it is true is 900 times more important than accepting the null hypothesis when it is false! Cohen (1988) recommends an alpha level of .05 and power of .80 which results in a 4 to 1 ratio, a much more realistic assumption. Examination of the significance criterion for power indicates that little flexibility is given to the researcher due to long standing conventions regarding the alpha level.

The next variable of interest in power calculations is the reliability of sample results. The reliability of the sample relates to how well the sample can approximate the relevant population value. Hence, the more reliable the sample, the greater the power, since differences between the populations are more easily detected. Reliability can be controlled in two ways: through experimental design and manipulation of sample size. Experimental design is often an under-examined area of power analysis. This is true because it is not quantifiable a priori. The effect of experimental design is simple, whatever reduces within-group variation will increase power. However, this is often not desirable since highly homogeneous groups are often not generalizable to larger populations. The recommendation for this area is to use the strongest experimental design that is applicable for the research question.

Sample size is the variable that is most under the control of the researcher. The larger the sample size is, the smaller the error and the greater the reliability of the sample. The smaller the error of the sample is, the greater the power is. While some populations are difficult to recruit into studies it is often necessary to increase sample size as much as possible to ensure adequate power to detect effects.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The final variable entered into the power equation is effect size. "Effect size is the degree to which a phenomenon is present in a population" (Cohen, 1988, p. 9). The null hypothesis posits an effect size of 0. Deviations from 0 indicate the effect size in the population. The greater the effect size the greater the power is. Also larger effect sizes indicate the need for smaller sample sizes without loss of power. Effect size is often the stopping point for many would be power calculators. How does one know the effect size before a study is conducted? Simply, one doesn't. However it can often be estimated from theory and other studies in closely related areas. If this is not possible, Cohen (1988) has constructed a series of effect sizes for each of the statistical tests in this book. For each of these tests he has constructed effect sizes that correspond to "small", "medium" and "large" effect sizes. While these are just conventions, they do allow the researcher some guidelines when conducting power analyses. While effect sizes are not generally alterable, the researcher is at liberty to decide which effects to examine. If the researcher knows that the effect is expected to be small, a design can be selected that will be powerful enough to detect these size effects. Conversely, in an area where a large effect is expected, scarce resources can be saved by employing the appropriate number of subjects.

While effect sizes are essential to psychological research, the size of the effect differs in importance depending on the aims of the study. The researcher also has to examine the clinical significance of the finding. A small effect in medicine, that is inexpensive and saves many lives, may be important while a larger effect that is expensive to implement may not be justified (Kazis et al., 1989; Rosenthal, 1990). An example of this may be an expensive, new teaching method in which the effect size is .05

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

or one-twentieth of a standard deviation. It is unlikely that this intervention would be employed on a large scale even though it has been shown to be effective (Prentice and Miller, 1992). On the other hand, in the physicians aspirin study in which over 22,000 physicians took part in a randomized double-blind experiment the effect was only .034. Yet, when examined in terms of the binomial effect size display this equaled a 45% decline in heart attacks. This result was seen as so strong that the trial was terminated prematurely because it was deemed unethical to continue giving the control group a placebo (Rosenthal, 1990). A study using a large population may have a statistically significant effect, but may not be clinically significant. Conversely, a result may be found to be not statistically significant, but the effect size might justify the need for a larger study to analyze the possibility of an important finding (Deyo & Patrick, 1995). Also, as Prentice and Miller (1992) have pointed out, small effects are impressive when minimal interventions are used or when it seems unlikely that a dependent variable will be influenced by an independent variable. It is essential for the researcher to weigh the costs of both types of errors before selecting the alpha level and the desired power (Overall & Dahal, 1965). Huysamen (1996) has argued that effect size and clinical significance are essential to the study of health psychology. He believes that highly powered studies can lead to the publication of statistically significant but unimportant results if very large sample sizes are used.

Health psychology has been a rapidly growing sub-field in psychology since its beginnings in 1973. The implications of research in health psychology can be enormous. Research in this field has lead to large scale clinical trials such as the Minnesota Heart Health Trial, MRFIT and project MATCH (Project MATCH Research Group, 1997).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

While these trials employed large number of participants and adequate measures of statistical power, much of the basic research used to design these treatments have had much smaller samples which indicate potentially inadequate levels of power. The problem of inadequate power could possibly lead to the rejection of numerous interventions that could have saved thousands of lives if employed on a population level. Only by examining the power of these studies can we know if the field is rejecting interventions because they are ineffective or because they were under powered.

Health psychology is also in a unique position among many of the fields of psychology because of its interdisciplinary nature. While a knowledge of effect size and power are increasingly found in the social and behavioral sciences they are still relatively new to the field of medicine (Kazis, Anderson, & Meenan, 1989). Use of effect size measures allows researchers and clinicians the opportunity to compare traditional results (i.e. changes in blood pressure and physical functioning) with psychometric measures that do not have concrete measurement scales associated with them (Kazis et al., 1989).

A review of Psychlit revealed no power analyses of the field of health psychology or the journals Health Psychology, Addictive Behaviors, and Journal of Studies on Alcohol. These three journals were selected for two reasons. First, each of these journals was selected because of the high readership (circulation > 1,200) and high citation rates of these journals (Howard & Howard, 1992). Selection of these journals is deliberate, since these journals are highly read and cited, power problems in these journals would affect most health psychology researchers. Secondly, these journals encompass three levels of specificity in health psychology. The Journal of Studies on Alcohol is the most specific accepting almost exclusively studies dealing with alcohol. Addictive Behaviors

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

includes studies on alcohol along with other studies on drugs, smoking, and other addictive behaviors. Health Psychology accepts studies on addictive behaviors, but also includes non-addictive health related studies. By using journals with different levels of specificity, results can be compared to assess if a power problem is ubiquitous within the field or if it is confined to a specific area.

Consistent with earlier power studies in the field of psychology, it is hypothesized that all three of the selected journals will demonstrate inadequate levels of power to detect all but large effects. Small and medium effects will have power less than .80. Although it is expected that some of the larger studies will have adequate power, the proliferation of small studies will greatly lower the average power. It was expected that these journals are not fundamentally different from the studies done by Cohen (1962), Rossi (1990) and others.

Method

Selection Procedure: Articles

All of the articles published in the 1997 volume of the journals, Health Psychology, Addictive Behaviors, and Journal of Studies on Alcohol were examined and only those articles containing statistical tests were selected. Of these articles, those for which power could not be computed because of the tests used were also discarded.

Selection Procedure: Statistical Tests

The articles were examined and statistical tests will be placed into two categories "major" and "peripheral" statistical tests. Major tests are based directly on the research hypothesis of the study, while peripheral tests are not. Peripheral tests were not included in this study. These tests can include correlation coefficients of a factor or principal

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

components analysis, unhypothesized higher order interactions in the analysis of variance, manipulation checks, inter-rater reliability coefficients, reliabilities of psychometric tests (internal consistency, test-retest), post hoc analysis of variance procedures (simple effects, multiple comparisons), and tests of statistical assumptions. This exclusion criteria is similar to other studies of this type and will result in little loss of information since for these tests power is often not appropriate or important. Examination of major tests was limited to the tests included in Cohen's (1988) power handbook. These tests include all the techniques included in Cohen's (1962) initial power study as well as multiple regression which was included in Rossi's (1984) replication of Cohen's work. Power for the multivariate analysis of variance (MANOVA) will also be included due to the recent development of computer programs for this test (J. S. Rossi, March 15, 1998). Table 1 displays the eligible tests. Not included in this analysis are most nonparametric techniques (e.g. Mann-Whitney U test, rank order correlation tests, ect.), most multivariate methods (e.g. two-way multivariate analysis of variance), and tests where the concept of power is not appropriate (e.g. factor analysis).

Determination of Statistical Power

Cohen's (1988) tables were used to determine power for the following statistical tests: differences between correlation coefficients, sign tests, differences between proportions, and chi-square tests. The tables require knowledge of sample sizes and effect size to compute power values. Two-tailed testing at $\alpha = .05$ was assumed. The

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

values in the tables are given to two decimal places and are accurate to about one digit in the last decimal place when compared to exact values.

Computer programs were written for the t, r, and F tests due to the ubiquitous nature of these tests. These programs were based on Rossi's (1984) BASIC programs designed for the same purpose. The power of the t test will be based on the normal approximation to the noncentral t distribution given by Cohen (1988). This formula was modified slightly to permit unequal N power calculations. Power for the Pearson correlation coefficient was based on the normal score approximation for r provided by the hyperbolic arctangent transformation, plus a correction factor for small sample sizes (Cohen, 1988). The cube root normal approximation of the noncentral F distribution was employed to compute power for the analysis of variance (Laubscher, 1960; Severo & Zelen, 1960).

Range and Type of Effect Size Indices

Power determinations were made using Cohen's (1969) definitions of small, medium, and large effect size. Cohen's (1962) earlier definitions of effect size were not examined since these estimates are not recommended for current use and have not been used in recent surveys (Cohen, 1988). The following sections list the measures of effect size that were used for each of the nine statistical tests in this study.

1. t-test: The effect size index for student's t test is Cohen's d, the standardized difference between group means (Cohen, 1988):

$$d = (M1 - M2) / s, \quad (1)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where M1 is the mean of the first group, M2 is the mean of the second group, and s is the common (pooled) standard deviation. d is related to delta, the noncentrality parameter for the noncentral t distribution, as follows:

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

where n is the sample size for each group.

The definitions for small, medium, and large effect sizes are .20, .50, and .80 respectively.

2. Pearson r: The effect size index for Pearson r is the correlation coefficient itself. The definitions for small, medium, and large effects sizes are .10, .30. and .50 respectively.

3. Differences between correlation coefficients: The effect size index for the difference between correlations is Cohen's (1969) q. This index is based on the Fisher r to z transformation:

$$q = | z(1) - z(2) |, \quad (3)$$

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

$$z = \ln((1+r)/(1-r))/2, \quad (4)$$

or equivalently by

$$z = \mathrm{arctanh}\,(r) \quad (5)$$

Small, medium, and large values of q are .10, .30, and .50, respectively. Rossi (1985) provides tables for computing q. These tables will be employed to expedite these calculations.

4. Sign test: The effect size here is just the departure of a proportion from .50:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\underline{q} = | \underline{p} - .50 |, \quad (6)$$

where p is the observed proportion Small, medium, and large effect sizes are defined as .05, .15, and .25 respectively.

5. Differences between proportions: Cohen (1988) presents a convenient index of effect size for calculating the difference between proportions as

$$\underline{h} = | \phi_1 - \phi_2 |, \quad (7)$$

where ϕ_1 and ϕ_2 are the arcsine transformation for the two proportions. The arcsine transformation was suggested by Eisenhart (1947) to stabilize the variance and normalize the distribution of proportions:

$$\phi = 2 \text{ arcsine } \sqrt{\underline{p}}. \quad (8)$$

Rossi (1985) again provides tables to easily calculate $\underline{h}$.

6. Chi-square tests: The chi-square test for $\underline{k}$ proportions (goodness-of-fit) test was difficult for Cohen to devise. In 1977, Cohen totally revised his early formulation and devised an index he called $\underline{w}$, where

$$\underline{w} = \text{sqr}(\underline{e}) = \text{sqr}(\underline{l}). \quad (9)$$

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

$$\text{lambda} = \underline{w}^2 * \underline{N}, \quad (10)$$

where $\underline{N}$ is the total sample size.

Definitions of small, medium, and large effect size are 0.10, 0.30, and 0.50 respectively.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

7. F tests in the analysis of variance and covariance: The effect size index is f, the standard deviation of the k standardized population means:

$$f = s(m) / s, \tag{11}$$

where s is the common (pooled) standard deviation of the k groups, and s(m) is the standard deviation of the k groups, and s(m) is the standard deviation of the means. Explicitly,

$$s(m) = sqr(sum((m(i) - m)^2)/k), \tag{12}$$

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

$$f = d / 2. \tag{13}$$

The index f is also closely related to phi, the noncentrality parameter of the noncentral F distribution introduced by Tang (1938):

$$\phi = f * sqr(n), \tag{14}$$

Furthermore, f is related to lambda, the noncentrality parameter used by Patnaik (1949) and others (Laubscher, 1960), as follows:

$$\lambda = f^2 * n * k. \tag{15}$$

Small, medium, and large effect sizes were defined by Cohen (1988) as .10, .25, and .40, respectively.

8. F tests in multiple regression/correlation analysis: Cohen (1977) suggested f^2 as a measure of effect size:

$$f^2 = R^2 / (1 - R^2), \tag{16}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

$$\text{lambda} = \underline{f}^2 * \underline{v}, \tag{17}$$

where

$$\underline{v} = \underline{N} - \underline{k} - \underline{1}. \tag{18}$$

Here $\underline{v}$ is the error degrees of freedom, $\underline{N}$ is the total sample size, and $\underline{k}$ is the number of groups.

Small, medium, and large effect sizes, in terms of $\underline{f}^2$, are defined as .02, .15, and .35, respectively. In terms of $\underline{R}^2$ these values are equal to .02, .13, and .26, respectively.

9. F test for the one-way multivariate analysis of variance: The equations described above for multiple regression and correlational analysis prove quite valuable when computing power for the one-way MANOVA because of their generality. However, multiple regression and correlation are only a realization of the univariate linear model, because it can only deal with one dependent variable at a time. Set correlation is the realization of the multivariate general linear model and a generalized version of multiple regression (Cohen, 1988). Therefore, MANOVA as well as other multivariate techniques are special cases of set correlation.

Cohen (1988) again uses f^2 as the measure of effect size. However, this time it is defined as:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$f^2 = L^{-1/S} - 1, \quad (19)$$

where

$$L = |E|/|E + H|, \quad (20)$$

Here, L = Wilks' Lamba, E is an error matrix, and H is an hypothesis matrix and

$$S = \sqrt{(k^2_Y k^2_X - 4)/(k^2_Y + k^2_X - 5)}, \quad (21)$$

where k_Y and k_x are the denominator degrees of freedom.

Cohen (1988) does not provide tables for computing power for the generalized one-way MANOVA test. Fortunately, Rossi (1991) has recently developed BASIC software to complete this type of analysis. This program was used for computing the power of one-way MANOVA in this study.

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

Once the power has been computed for each test the results will be combined to assess the average power for each journal. When doing this the unit of analysis will be the article, since the number of tests in each article has been found to vary widely (Cohen, 1962; Rossi, 1984). The power of each study will be determined by averaging across statistical tests. The studies will then be averaged together producing estimates for each journal for small, medium, and large effect sizes.

Results

A total of 216 articles were examined, 65 in Health Psychology (HP), 83 in Addictive Behaviors (AB) and 68 in the Journal of Studies on Alcohol (JSA). Twenty-nine articles were excluded, 9 in HP, 13 in AB and 7 in JSA. Twelve of the articles

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

contained no statistical tests at all and 3 were meta-analyses and were not included for theoretical reason described earlier. The remaining fourteen articles contained tests for which power could not be computed, mainly data reduction techniques, odds ratios, and structural equation modeling.

Power was computed for statistics reported in the remaining 187 articles: 56 in HP, 70 in AB and 61 in JSA. Power was calculated for 8,266 eligible tests in these 187 articles: 2,429 in HP, 2,449 in AB, and 3,388 in JSA. The frequency of each test was coded and is reported in Table 2. The sample was dominated by "traditional" statistical tests: Pearson r, analysis of variance and covariance, and the t test, which accounted for 84% of the eligible tests. Pearson r was the most common statistical test used ranging from 44% in AB to 58% in JSA. The distributions in these three journals are highly similar to each other and to Rossi (1984) indicating a stability of preference of tests across journals and across time.

The number of eligible statistical tests per article varied greatly from 1 to 334, the study was used as the unit of analysis (following Cohen, 1962; Rossi, 1984). The power of each study was determined by averaging across each statistical test.

Three separate power estimates were made for each statistical test. All estimates were made using Cohen's (1988) definitions of small, medium, and large effect sizes. Power estimates for the three journals are reported in Table 3. The percentage of studies with power less than .50 and .80 are reported in Table 4. Power results are broken down further and displayed in Table 5. Striking similarities were found between the three journals. Power for small effects exhibits the largest variations in mean differences. However, a one-way analysis of variance revealed no significant differences between the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

three journals for small effects, $F(2,186) = 1.3$, $p > .05$. Similarly, no significant differences existed between the journals for medium and large effects either ($p > .05$). Power for these ANOVAs was .20 for small effects, .89 for medium effects and .99 for large effects. Since none of the results were significant, probable upper bounds were examined. This test revealed that when $f = .23$, power was .80. Therefore, the difference in power between the journals was likely to be less than $f = .23$.

It was hypothesized that federally funded studies would have greater power than those studies which did not receive extramural funding. Each study was coded as funded if it acknowledged an outside funding source. Some articles from researchers outside the United States did not report funding even though it appeared that they must have received some type of outside funding (i.e. an intervention study with $n > 10,000$), so this method was not exact. Yet, it is believed that most articles were correctly coded, and any incorrectly coded articles would only minimize the difference between the groups. Most of the studies (n = 113, 60%) reported outside funding, though a substantial portion did not report funding (n = 74, 40%). Independent samples t-tests were used to compare power for small, medium and large effects for funded and unfunded studies. Results indicated that no differences existed for large effects. However, funded studies (M = .80) had significantly higher power to detect medium effects than did unfunded studies (M = .71), $t(185) = 2.3$, $p < .05$, $d = .36$. This result also held true for small effects, where funded studies (M = .41) had significantly higher power than unfunded studies (M = .28), $t(185) = 2.9$, $p < .01$, $d = .45$. These results are displayed in Table 6.

Power for these three journals were then compared to power studies done in other areas of psychology. Although dozens of power studies have been completed, most of

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

these are done in areas outside of psychology such as education and communication. Power for the three health psychology journals are compared to other power studies in psychology and presented in Table 7. Results indicate that the field of health psychology generally employs studies using higher power than other areas. See Rossi (1990) for a more in depth treatment of this issue.

Discussion

The results for this study are very encouraging. As displayed in Tables 4 and 5 the average overall power for small effects is inadequate (.36), however power to detect medium effects is good (.77) and power to detect large effects is excellent (.92). Cohen (1988) has recommended power of .80 to detect effects. While only 15% of the studies had adequate power for small effects, 60% of the studies had adequate power for medium effects, and 86% had adequate power for large effects.

Of particular note is the striking similarity among the three journals. The power estimates for HP and AB are almost exactly the same. The results for JSA are slightly higher due to 12 (20%) large studies ($n > 2000$) inflating its results, though not significantly. However, it appears that the overall estimates are a good estimate of power for the field of health psychology because of their degree of similarity. Although, results might differ slightly from year to year it is not expected that this would be more that 5% overall from 1997.

The comparison to other areas where power has been computed is also striking. These three journals have quite a bit more power than most other areas of psychology examined so far, although these analyses has been scant. Comparing these results to the 25 power studies analyzed by Rossi (1990) indicates that the field of health psychology is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

one of the most highly powered areas similar to marketing, sociology, and applied psychology.

Power was computed in this study for alpha = .05, two-tailed (following Cohen, 1962; Rossi, 1984). Most of the articles in this study used this common convention. Nine articles (4.8%) used some type of alpha correction for at least some of their tests. Although alpha corrections are often cited as an area where power is greatly decreased, it is doubtful that it had any differential effect on the results on this analysis, since even when it was performed it accounted for less than 50% of the eligible tests in that article making the impact on the power study negligible. However the effect of alpha corrections on the individual test is quite significant. One article actually used $\alpha < .10$. The authors did not mention power as a reason for this increase in alpha but instead reasoned that this was an exploratory study so they were interested in finding relationships. Interestingly, power for this study using $\alpha < .05$ was .49 for small effects, .99 for medium effects and .99 for large effects. Also 21 tests were performed and 20 of these were ANOVAs with no alpha corrections. This technique is dubious and highly inflates the Type I error rate. It is not recommended as a technique to increase power.

Even though power in three journals is improving, the reaction of researchers in the field should still be one of cautious optimism. Since significant results are more likely to be published, while non-significant results are more likely to be placed in "file drawers" never to be seen, it is likely that the average study in health psychology has a much lower power than those reported in these journals.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Methodologists have long decried the lack of acknowledgment of statistical power in the social sciences. In this study, nine (4.8%) articles at least mentioned statistical power, while seven (3.7%) actually computed power. It is encouraging that power is being mentioned more in articles. One reason for this is the requirement that power calculations being included for government funded grant applications. This should increase the number of people who are knowledgeable about how to calculate power and also increase the power of studies. Funded studies in this analysis had significantly greater power than non-funded studies. This could be due to the requirement of power calculations in the grant application or more likely the ability to recruit more subjects with an increase in funds. However, even though it appears that power is being mentioned and calculated more often, a review of studies that mentioned power indicated that statistical power is still largely misunderstood and abused. One of the articles reported calculating power to detect effects and having adequate subjects for medium effects using the t-test. However this author calculated power using the medium effect of $\underline{d} = .30$ instead of .20 and calculated power for .75 instead of .80. In another study, the authors calculated power for $\underline{p} < .05$ and then conducted all of their analysis with $\underline{p} < .01$. In a final case of misunderstanding of statistical power, the authors computed power correctly and then reported that $\beta = .75$ instead of $1-\beta = .75$! These examples clearly show that while there is a realization that power is an important component of significance testing it is still largely misunderstood or misused. This lack of knowledge about power underscores the need to teach power at the undergraduate level in introductory statistics courses. My experience indicates that with a little extra work

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

students are able to comprehend power and effect size and are able to calculate power for tests using Cohen's (1988) tables.

Although power in the field of health psychology has been shown to be adequate to detect medium and large effects, for small effects it is still below .50 in 71% of the studies and below .80 in 85% of the studies. It is now logical to ask: How many subjects are needed to adequately detect small effects? Table 8 displays the number of subjects needed to have power of .50 and .80 for the most common tests in psychology: ANOVA/ANCOVA with three groups, Pearson r, t test with equal sample sizes, and chi-square test with 1 degree of freedom. Examination of this table reveals that it takes approximately 280 subjects to detect small effects with power = .50 for the Pearson r and over 900 to detect power for small effects at .80 for a three group one-way ANOVA. Making .80 a standard to detect small effects would make research on small populations for unfunded researchers prohibitive. What can be done to remedy this situation? As Rossi (1990) points out the publishing bias against null results leads to an inflation in Type I errors because articles that achieved significant results are capitalizing on chance. Schmidt and Hunter (1997) suggest that a way to remedy the problem of low power is to publish or make available all of the data from primary studies to meta-analysts. Thus, each study is considered a data point from which more valid conclusions can be drawn without the aid (hindrance?) of significance testing. Obviously, this is a difficult problem. Rossi (1990) recommends using significance tests to develop a probable upper bounds on an effect. In this case, if an effect is non-significant a researcher can determine by calculating power that a certain effect is very unlikely in this case. Once this is established, researchers would know that larger sample sizes would be necessary to detect

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

a possible smaller effect of the phenomenon. Harlow (1997) notes that one area that proponents and adversaries of significance testing agree on is the need to test strong theories. Hopefully if strong theories are tested the average effect size will be larger. Eventually, psychologists have to ask how important small effects are? Are they central to discovering the truths of human behavior or are they just part of Cohen's (1994) so called "crud factor"? This question remains to be answered.

Even though power analyses have been seen as extremely valuable for assessing power in research fields, they still face one common limitation, the true effect size is not known. To circumvent this problem, researchers have typically used Cohen's definitions of small, medium and large effect sizes to give three estimates of power for the area. Rossi (1997) has indicated that there is a more efficient way to assess power for particular areas. He notes that the use of meta-analysis provides the researcher with an index of effect size that can be then imputed back into the original studies, so an average effect size for the area can be established. By using this method, power for specific areas can easily be computed. However, when knowledge about an entire area of study such as health psychology, where many areas of inquiry exist and naturally many effect sizes are present, power surveys still appear to be the best way of assessing power.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

References

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

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

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

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

Christensen, J. E. & Christensen, C. E. (1977). Statistical power analysis of health, physical education, and recreation research. Research Quarterly, 48, 204-208.

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

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

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Academic Press.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

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

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

Davis, J. M., Janicak, P. G., Wang, Z. & Gibbons, R. D. (1992). The efficacy of psychotropic drugs: Implications for power analysis. Psychopharmacology Bulletin, 28, 151-155.

Deyo, R. A. & Patrick, D. L. (1995). The significance of treatment effects: The clinical perspective. Medical Care, 33, AS286-AS291.

Eisenhart, C. (1947). Inverse sine transformation of proportions. In C. Eisenhart, M. W. Hastay, & W. A. Wallis, (Eds.), Selected techniques of statistical analysis for scientific and industrial research production and management engineering (pp. 395-416). New York: McGraw-Hill.

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

Haase, R. F. (1976). Power analysis of research in counselor education. Counselor Education and Supervision, 14, 124-132.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Harlow, L. L. (1997). Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 1-20). Mahwah, NJ, Lawrence Erlbaum Associates.

Howard, M. O. & Howard, D. A. (1992). Citation analysis of 541 articles published in drug and alcohol journals: 1984-1988. Journal of Studies on Alcohol, 53, 427-434.

Hunter, J. E. & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Huysamen, G. K. (1996). Some methodological issues in health psychology research. South African Journal of Psychology, 26, 10-15.

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.

Kalichman, S. C., Carey, M. P., & Johnson, B. T. (1996). Prevention of sexually transmitted HIV infection: A meta-analytic review of the behavioral outcome literature. Annals of Behavioral Medicine, 18, 6-15.

Katzell, R. A. & Dyer, F. J. (1977). Differential validity revived. Journal of Applied Psychology, 62, 137-145.

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

Kazis, L. E., Anderson, J. J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. Medical Care, 27, S178-S189.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Kosciulek, J. F. (1993). The statistical power of vocational evaluation research. Vocational Evaluation and Work Adjustment Bulletin, 26, 142-145.

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

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

Orme, J. G. & Tolman, R. M. (1986). The statistical power of a decade of social work education research. Social Service Review, 60, 620-632.

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

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

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

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.

Prentice, D. A. & Miller, D. T. (1992). When small effects are impressive. Psychological Bulletin, 112, 160-164.

Project MATCH Research Group (1997). Matching alcoholism treatments to client heterogeneity. Project MATCH post-treatment drinking outcomes. Journal of Studies on Alcohol, 58(1), 7-29.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

Rosenthal, R. (1990). How are we doing in soft psychology? American Psychologist, 45, 775-776.

Rossi, J. S. (1984). Statistical power of psychological research. Unpublished doctoral dissertation, University of Rhode Island.

Rossi, J. S. (1985). Tables of effect size for z score tests of difference between proportions and between correlation coefficients. Educational and Psychological Measurement, 45, 737-743.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Rossi, J. S. (1991). Power for k-groups MANOVA. Unpublished computer program. University of Rhode Island.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-198). Mahwah, NJ, Lawrence Erlbaum Associates.

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

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Sindelar, P. T., Allman, C., Monda, L., & Vail, C. O. (1988). The power of hypothesis testing in special education efficacy research. Journal of Special Education, 22, 284-296.

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

Thomas, L. & Juanes, F. (1996). The importance of statistical power analysis: An example from Animal Behaviour. Animal Behaviour, 52, 856-859.

Wooley, T.W. (1983). A comprehensive power-analytic investigation of research in medical education. Journal of Medical Education, 58, 710-715.

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.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 1: Statistical Tests Included in Power Analysis

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 2:

Frequency distribution of tests used in power survey

Statistical Test	Health Psychology		Addictive Behaviors		Studies on Alcohol		Overall	
	Frequency	Proportion	Frequency	Proportion	Frequency	Proportion	Frequency	Proportion
Pearson r	1401	57.7%	1077	44.0%	1976	58.3%	4454	53.9%
ANOVA /ANCOVA	481	19.8%	430	17.5%	449	13.3%	1360	16.5%
t test	256	10.5%	475	19.4%	420	12.4%	1151	13.9%
Chi-square	182	7.5%	281	11.5%	292	8.6%	755	9.1%
Multiple Regression	84	3.5%	150	6.1%	56	1.7%	290	3.5%
MANOVA / DFA	19	0.8%	27	1.1%	20	0.6%	66	0.8%
Difference between r's	6	0.2%	0	0.0%	16	0.5%	22	0.3%
Difference between p's	0	0.0%	9	0.4%	159	4.7%	168	2.0%
Total	2429		2449		3388		8266	

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 3

Average power for the three journals

Effect Size	Mean	SD	Median	95% C.I.
Health Psychology	(n = 56)			
Small	.34	.31	.18	.26-.42
Medium	.74	.25	.83	.68-.81
Large	.92	.12	.99	.89-.95
Addictive Behaviors	(n = 70)			
Small	.34	.28	.19	.27-.41
Medium	.75	.27	.83	.69-.81
Large	.90	.17	.99	.86-.94
Studies on Alcohol	(n = 61)			
Small	.41	.32	.29	.33-.49
Medium	.81	.25	.92	.74-.87
Large	.92	.18	.99	.88-.97
Overall	(n=187)			
Small	.36	.30	.22	.32-.40
Medium	.77	.26	.86	.73-.81
Large	.92	.16	.99	.90-.94

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4

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

Effect Size	< .50	< .80
Health Psychology	(n = 56)	
Small	79%	86%
Medium	23%	46%
Large	2%	16%
Addictive Behaviors	(n = 70)	
Small	71%	89%
Medium	19%	40%
Large	6%	17%
Studies on Alcohol	(n = 61)	
Small	62%	80%
Medium	10%	34%
Large	7%	10%
Overall	(n=187)	
Small	71%	85%
Medium	17%	40%
Large	5%	14%

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 5

Power of 187 Studies Published in Health Psychology, Addictive Behaviors, and the journal of Studies on Alcohol in 1997.

	Small effects		Medium effects		Large effects	
Power	Frequency	Cumulative %	Frequency	Cumulative %	Frequency	Cumulative %
.99	14	100	58	100	106	100
.95-.99	4	93	18	68	24	43
.90-.94	2	90	11	58	21	30
.80-.89	8	89	25	52	8	19
.70-.79	6	85	14	39	9	14
.60-.69	9	82	13	32	6	10
.50-.59	12	77	15	25	3	6
.40-.49	13	71	3	17	4	5
.30-.39	12	64	15	15	3	3
.20-.29	18	57	5	7	1	1
.10-.19	59	48	7	4	1	1
.05-.09	30	16	1	1	---	0
N	187		187		187	
M	.36		.76		.91	
Mdn	.21		.85		.99	
SD	.30		.26		.16	
Q_1	.13		.59		.93	
Q_3	.55		.99		.99	

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 6:

Differences in the Power of Studies by Funding Status

	Funded N = 113 M (SD)	Nonfunded N = 74 M (SD)	t(185)	p	d
Small	.41 (.31)	.28 (.27)	2.9	.004	.45
Medium	.80 (.25)	.71 (.27)	2.3	.02	.36
Large	.93 (.15)	.89 (.16)	1.4	.17	.25

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 7:

Power Surveys in Psychology

Source	Journals included	Years covered	Sample size		Statistical power estimate		
			Articles	Tests	Small	Medium	Large
Maddock (1999)	***Journal of Studies on Alcohol***	**1997**	**61**	**3388**	**.41**	**.81**	**.92**
Mone et al. (1996)	*Journal of Applied Psychology*	1992-94	30	100	.35	.82	.95
Maddock (1999)	***Health Psychology***	**1997**	**56**	**2429**	**.34**	**.74**	**.92**
Maddock (1999)	***Addictive Behaviors***	**1997**	**70**	**2449**	**.34**	**.75**	**.90**
Mone, et al. (1996)	*Personnel Psychology*	1992-94	30	105	.30	.83	.97
Chase & Chase (1976)	*Journal of Applied Psychology*	1974	121	3,373	.25	.67	.86
Sedimeier & Gigerenzer (1989)	*Journal of Abnormal Psychology*	1984	54	NR	.21	.50	.84
Cohen (1962)	*Journal of Abnormal and Social Psychology*	1960	70	2,088	.18	.48	.83
Mone et al. (1996)	*Organizational Behavior and Human Decision Process*	1992-94	30	113	.17	.60	.87
Rossi (1990)	*Journal of Abnormal Psychology* *Journal of Consulting and Clinical Psychology* *Journal of Personality and Social Psychology*	1982	221	6,155	.17	.57	.83

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 8:

Number of Subjects Needed to Detect Small Effects

Test	Conditions[1]	N for power = .50[2]	N for power = .80
Pearson r		280	620
Chi-square	df = 1	380	780
t test	equal sample sizes	272	620
ANOVA	3 groups	492	945

[1]$p < .05$ is assumed

[2]Total n needed

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 3: Meta-Analysis Of Interventions To Reduce Alcohol Consumption Among College Students

Introduction

Integrating Findings across Research Studies

Lack of power in individual studies has lead to psychology's failure as a cumulative science (Rossi, 1997). Studies that had inadequate power were often found to be non-significant even when the effect really existed. Traditional narrative reviews help to increase the debate over these inconsistencies by using "head counting" methods instead of an impartial statistically controlled methods. Meta-analysis provides a much more powerful way to examine effect sizes in specific areas of research (Rosenthal, 1991). The use of meta-analysis has exploded over the last two decades and is widely accepted by many researchers (Hunter & Schmidt, 1990; Bailar, 1997), although there are still several researchers who question the utility of meta-analysis (Abelson, 1997; LeLorier, Gregoire, Benhaddad, LaPierre, & Derderian, 1997).

Effect of Interventions for Alcoholics

Several reviews and meta-analyses have been conducted on the outcome of treatments with alcoholics. Qualitative reviews have come to largely different conclusions. Schuckit (1992) notes that treatments are quite effective with 60 to 70% of alcoholics remaining abstinent one year after treatment. Valliant (1988) does not concur with this conclusion. He states that treatment does not decrease long term morbidity or mortality of alcoholics. Lindstrom (1992) observed only weak and short-term effects of treatment. Goodwin (1988) concluded that there is insufficient evidence of efficacy or

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

cost effectiveness for any existing treatment. These differential conclusions indicate a problem with traditional narrative reviews of the literature.

Even in the area of meta-analysis, there are different conclusions drawn depending on the variable of interest. Agosti (1994) found that only 20% of studies reported significantly different levels of abstinence between the treatment and control groups. However, in a later report, a strong effect (E. S. = 1.17) for reduction in quantity consumed in the control group was noted (Agosti, 1995).

These differential results point out the problems in the field of alcohol with traditional narrative review. In addition, it also points out the need for meta-analytic studies to examine several outcome variables.

Alcohol Interventions with College Students

Alcohol abuse is a major problem on college campuses. Alcohol abuse is considered the number one problem on many campuses by college presidents (Wechsler, Davenport, Dowdall, Moeykens, & Castillo, 1994). The seriousness of this problem along with the availability of funds from the Department of Education's Fund for the Improvement of Post-Secondary Education and the Safe and Drug Free Schools Act has led to an estimated 95% of the Nation's post-secondary institutions instituting substance abuse policies and alcohol or other drug (AOD) prevention programs (Commission on Substance Abuse at Colleges and Universities, 1994). Despite the overwhelming presence of these programs very few studies have employed sufficient methodological rigor to evaluate the effectiveness of these programs (Wood, 1998).

Moskowitz (1989) reviewed the effects of programs and policies for reducing the incidence of alcohol problems. He found strong support for raising the minimum legal

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

drinking age, increasing alcohol taxes and enforcement of drunk driving laws. However, he concluded that little evidence existed for the efficacy of primary prevention programs. Wood (1998) has recently compiled a narrative review of interventions to reduce college drinking. He identified eighteen studies that met minimal methodological requirements. This review indicated that cognitive behavioral / self-regulation approaches were superior to traditional educational approaches. However, research has shown that narrative reviewers are not good at estimating significance from a group of studies and are often too conservative (Cooper & Rosenthal, 1980). A review of the literature reveals that no meta-analyses have been conducted on college interventions for drinking. Employing meta-analytic procedures on this set of studies will produce the average effect size for both of these alcohol reduction approaches and produce an estimate of the average power in these studies. Once an average effect size is found for each type of intervention, researchers will be able to easily compute power for future research efforts.

Need for Meta-analysis

The need for a meta-analysis in this area is crucial. A meta-analysis in this area will provide the answers to a number of important questions in this area: 1) Do interventions on college drinking have any effect? 2) What is the typical effect size for college alcohol interventions? 3) What is the power for this area? 4) How many subjects should be used in future studies? 5) What type of interventions have the strongest results? 6) What effect do poorly controlled studies have on the results? (Bangert-Drowns, 1986). The answers to these questions will be invaluable to anyone conducting interventions in this area. Since an estimated 95% of the Nation's post-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

secondary institutions have instituted substance abuse policies and alcohol or other drug (AOD) prevention programs the results of this study could be important.

Justification of Major Hypothesis

Narrative reviews of the alcohol intervention literature have produced conflicting results. Some studies have shown positive results, while others have shown negative results. As Rossi (1990) points out this can be due to lack of power and small effect sizes. Since the results have been inconsistent, it is hypothesized that when taken as a whole a small effects exist for alcohol intervention studies. Wood (1998) discovered more support in his narrative review for cognitive social learning over traditional educational approaches. Moskowitz (1989) found no support for the traditional educational approaches. This relationship is expected to hold up, although the magnitude of difference is questionable. Poorly controlled studies are abundant in the college alcohol intervention literature (Wood, 1998). The lack of control groups in these studies will influence the results in unpredictable ways. Confounds such as history effects and subject maturation will not be controlled for which will increase the variability of the effect sizes for these studies. Consistent with earlier hypotheses, power will be inadequate to detect small and medium effects. This finding will demonstrate Rossi's (1990) hypothesis that controversy over whether findings exist is largely due to lack of statistical power in research designs.

Methods

Study Inclusion Criteria

Studies were obtained from a variety of resources. First, electronic databases were examined including Psychlit, MedLine, and NIAAA's ETOH database. Secondly,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

examination of the references for several reviews of the literature including Moskowitz (1989) and Wood (1998) were examined for missed studies. Studies conducted between 1974 and 1999 were included in the analyses. This twenty-five year period was selected because it has been electronically cataloged by the various search engines. Studies were then coded based on a minimal methodological standard. The studies were coded into two groups based on methodological standards. The first group must have random assignment to control or intervention groups, or baseline differences must have been statistically controlled in lieu of random assignment. The second group included studies that do not meet these standards. Methodological rigor of studies should always be examined when conducting a meta-analysis and great care should be taken when combining randomized and nonrandomized experiments (Heinsman & Shadish, 1996). For this reason the second group was not included in the meta-analysis.

Wood (1998) has identified two major types of approaches for alcohol interventions among college student. The first he called traditional alcohol education approaches. These approaches include information and values clarification. Only one of eleven of these studies met minimal methodological requirements to be included in his study (Wood, in press). The second approach encompasses cognitive-behavioral/self-regulation interventions. Sixteen studies met criteria here.

Studies will be coded by the methodological criteria outlined above as well as type of approach used. Appendix A contains the studies included in this meta-analysis.

Meta-Analytic Procedures

Meta-analysis is a quantitative synthesis of results from many separate studies (Glass, 1976). Meta-analysis is not a technique in itself but rather a set of techniques.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Several schools of meta-analytic thought exist to provide a quantitative synthesis of research results. The three main schools of meta-analysis have been identified as the Hedges and Olkin (1985) techniques, the Rosenthal and Rubin (1978, 1988) techniques, and the Hunter, Schmidt and Jackson (1982) techniques. A recent study by Johnson, Mullen & Salas (1995) compared these three approaches using systematically differing databases. They discovered that the Hedges and Olkin techniques and the Rosenthal and Rubin techniques provided convergent results that conformed to a priori predictions made by the researchers. The Hunter, Schmidt, and Jackson techniques varied widely from the other two and did not conform to the a priori predictions. The researchers (Johnson et al., 1995) suggested strong caution when using Hunter and his colleagues techniques. For this study, the techniques outlined in Cooper & Hedges (1994) were employed. These techniques provide an update of earlier techniques and also incorporate some of the adjustments made by Hedges and Olkin (1985) and Rosenthal (1991).

The general analytic questions that are typically answered using these techniques are central tendency, variability, and prediction. Central tendency is measured by combinations of effect size and combinations of significance levels. Variability is measured by homogeneity tests of effect size. Prediction is measured by comparing the study outcomes as a function of their discrete or continuous characteristics (Johnson, Mullen & Salas, 1995).

Choice of Effect Size

Reviews of the literature in this area have revealed that traditional alcohol education approaches have demonstrated differences in knowledge but no differences in drinking levels (Wood, 1998). Studies designed to change expectancies have

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

demonstrated differences in expectancies, while studies designed to change normative perceptions have also found differences. Since all of these studies have used different mediating variables, which may or may not be equated with drinking outcomes, effect sizes for differences in mediating variables will not be examined. In this study, only drinking related variables or negative consequences from drinking variables will be assessed in the meta-analysis. Due to differences in the number of effect sizes per study, each study will only be allowed to contribute one effect size to the overall analysis. For the overall analysis, the effect size from the longest follow-up assessment will be chosen (following Baillie, Mattick, Hall & Webster, 1994).

Analyses to determine the central tendency of this area will be conducted in two main categories. The first will be to discover if a significant effect exists. The second will be to assess how strong the effect is for this area of research. The first category will be analyzed by techniques that can be thought of as significance tests on the results of separate studies. The unit of analysis here is the study itself, and the data analyzed may be the probability level attained by the separate significance tests, or the significance test values, or some transformation of these. The second category will be assessed using techniques which average the effect size results of the separate studies.

Significance level

Procedures for assessing the overall level of statistical support for an effect are most useful when the results of some studies have been statistically significant while the results of other studies have not. This is the case in college alcohol intervention research which has been noted in traditional narrative reviews (Moskowitz, 1989; Wood, 1998).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Rosenthal (1978, 1980, 1982, 1991) and others (Hunter, Schmidt, & Jackson, 1982; Hunter & Schmidt, 1990) have described various techniques for summarizing significance levels. Perhaps the simplest and most generally useful technique is the "method of adding z scores" (Rosenthal, 1991):

$$z = \text{sum}(z(I)) / \text{sqr}(N), \tag{22}$$

where z is the normal score associated with the overall level of significance, z(I) is the normal score associated with the probability level attained by the significance test for the i^{th} experiment, and N is the number of experiments included in the analysis. Although simpler than other combining procedures, the method of adding z scores nevertheless gives results in good agreement with the results of other techniques, such as the well-known Fisher (1932; Winer, 1971) chi-square method for combining probabilities.

The method of adding z scores requires a knowledge of the exact value of the test statistic and its associated degrees of freedom so that a z score corresponding to the significance level (p) attained by the test statistic can be determined. For non-significant findings this level is often not reported. When this is the case a z score of zero will be assigned to the variable when the study suggests no effect or a non-significant positive effect. If the non-significant score is in the negative direction, a negative score will be recorded.

Scientific journals have demonstrated a preference towards publishing significant findings. This creates a "file drawer" problem where many non-significant findings never find their way into the literature (Rosenthal, 1979). Interestingly, while many researchers decry journals unwillingness to publish non-significant findings, these findings could be more often than not type II errors instead of a confirmation of the null hypothesis since

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the average power to find medium effects is often less than .50! Rosenthal (1991) provides a method for assessing how many non-significant studies it would take to make a positive finding non-significant. By rearranging equation 19 for significance and solving for N, a fail safe number of non-significant studies is produced. This estimate will indicate if the file drawer concern is a problem.

Average Effect Size

If the accumulated z scores produce a significant result, average effect sizes will be computed. Since effects sizes are rarely reported they will generally have to be estimated for each experiment. The most widely used measure of effect size is standardized mean difference statistic (Cohen, 1988), which is computed from group means, standard deviation and sample size:

$$\underline{d} = (M_t - M_c)/s_p, \tag{23}$$

where M_T = mean of the treatment group, M_c = the mean of the control group, and s_p = pooled standard deviation. s_p is calculated as:

$$S_p = sqr[[(n_e - 1)s^2_e + (n_c - 1)s^2_c]/(n_e+n_c - 2)], \tag{24}$$

Where n_e is the sample size for the treatment group, s^2_e is the variance of the treatment group, n_c is the sample size for the control group, s^2_c is the variance of the control group.

Since complete information is not always reported in journal articles, several other algebraically equivalent ways of computing d have been devised (Ray & Shadish, 1996). These will be used if information is not available to compute equation 23. These methods are test specific and are as follows: Between-groups t test on posttest scores with sample sizes for each group.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\underline{d} = t^*sqr((1/n_e) + (1/n_c)), \quad (25)$$

where t is the value of the t statistic.

Two-group between-groups one-way ANOVA is defined as

$$\underline{d} = sqr(F(1/n_e + 1/n_c)), \quad (26)$$

where F is the value of the F statistic.

Two-factor between groups ANOVA on posttest scores is computed by using equation 23 and substituting the following equation for equation 24:

$$s_p = sqr((SS_b + SS_{ab} + SS_w)/(df_b + df_{ab} + df_w)), \quad (27)$$

Where SS_b is the sum of squares for the second factor, SS_{ab} is the sum of squares for the interaction, SS_w is the sum of squares for the residual, and the degrees of freedom have parallel notation. This method can also be extended to larger between-groups factorial designs by extending the same logic.

Several other methods for estimating $\underline{d}$ have also been devised. Ray and Shadish (1996) have examined several other methods for estimating $\underline{d}$ that were not algebraically equivalent. These methods varied widely in their equivalence to $\underline{d}$ and could lead to biased effect size measures and inaccurate significance tests. Therefore, these techniques will not be used in the proposed study. Some researchers also use only the standard deviation of the control group instead of the pooled standard deviation (Glass, McGaw & Hill, 1981). While there is some debate over which measure of effect size is a better estimate, the pooled standard deviation tends to provide a better estimate in the long run (Rosenthal, 1991).

Another measure of effect size that is preferred by many researchers is $\underline{r}$ (Rosenthal, 1991). This effect size can be computed easily for the chi-square test,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$r = sqr(\chi^2(1)/N), \quad (28)$$

the t test,

$$r = sqr(t^2/(t^2+df)), \quad (29)$$

where $df = n_1 + n_2 - 2$, and for the F test

$$r = sqr(F(1, -)/F(1,-) + df_{error}), \quad (30)$$

where F(1,-) indicates any F with df = 1 in the numerator.

If none of these tests of significance have been used or reported r can be computed from a p value and N. This is done by converting p to its standard normal deviate equivalent using a table of Z values (Rosenthal, 1991).

$$r = Z/sqr(N), \quad (31)$$

The prior equations all produce product moment correlation coefficients (Pearson's r) and can be interpreted in the same way regardless of whether the data was originally continuous or dichotomous or ranked (Rosenthal, 1991).

The effect size indicated r can also be obtained from d as follows

$$r = d/(sqr(d^2) + (1/pq), \quad (32)$$

where p is the proportion of the total population in the first of the two groups, and q is the proportion in the second. When the proportions are equal this equation is simplified as

$$r = d/(sqr(d^2) + 4). \quad (33)$$

We can also convert r to d easily

$$d = 2r/sqr(1-r^2). \quad (34)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

For this proposed study, d will serve as the final index of group differences. R will be computed when d cannot be, due to lack of summary information and then it will be converted to d using equation 34.

Unfortunately, as the population value of r gets further away from zero the distribution of r's sample from the population becomes more and more skewed. Fisher (1928) provides a transformation of r (z_r) that is almost normally distributed. The relation ship between r and z_r is summarized as:

$$z_r = \frac{1}{2} \log_e((1+r)/(1-r)). \quad (35)$$

The estimation of effect size is slightly too large by r-population/[2(N-1)]. However, for practical purposes this bias is only substantial when r is very large and N is very small, so a correction is usually not necessary (Snedecor & Cochran, 1989).

Effect sizes (z_r) for two studies can be combined using the following equation:

$$z_r = (z_{r1} + z_{r2})/2. \quad (36)$$

Tables can then be used to convert z_r to r. Additionally, if weighted means are desired to correct for differences in sample size, the following equation can be computed:

$$\text{weighted mean } z_r = (w_1 z_{r1} + w_2 z_{r2})/w_1 + w_2. \quad (37)$$

These equations can then be easily converted to accommodate more studies following the same logic outlined above.

It is expected that some studies will use percentage of participants reaching a preset criteria as an outcome measure. When the outcome measure is dichotomous a fourfold table will be used. In a fourfold table, two dichotomous variables are crossed with one another to form four possible categories (Haddock, Rindskopf & Shadish, 1998). In this example, treatment condition would be the first variable and treatment

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

success would be the second variable. This would create four categories: participants in treatment who met criteria, participants in treatment would did not meet criteria, participants in the control group who met criteria, and participants in the control group who did not meet criteria. The most appropriate measure of effect size for the fourfold table is the odds ratio (Agresti, 1990; Fliess, 1981; Sandercock, 1989). The odds of improving given that one is treated can be expressed as the following ratio of conditional probabilities,

$$\Omega_E = P(I|E)/P(I'|E). \quad (38)$$

where P(I|E) is the probability of improving given treatment and P(I'|E) = 1 – P(I|E) is the probability of not improving given the same treatment. Similarly, $\Omega_{E'}$ can be defined for participants not receiving treatment. The overall measure of effect size is then computed as:

$$\omega_i = \Omega_E/\Omega_{E'}, \quad (39)$$

where $\underline{i}$ refers to the $\underline{i}^{th}$ study (Haddock, Rindskopf & Shadish, 1998). Since ω is not easily transferable to $\underline{r}$ or $\underline{d}$, studies that use dichotomous outcomes will be analyzed separately. While $\underline{d}$ or the product-movement correlation coefficient is often used in psychology to analyze dichotomous data, they will not be used in this study since they tend to underestimate the size of the effect (Fleiss, 1981, p. 60).

Once an overall effect size is calculated, it will be necessary to assess the homogeneity of effect sizes (Hedges, 1981). This is done by calculating a chi-square test such that:

$$\chi^2 = \Sigma(w(d-d_i)^2) \quad (40)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where d is the weighted mean d of the studies to be aggregated, d_i is the effect size for each study, and w is the reciprocal of the estimated variance of each d. The result is a χ^2 distribution with K-1 df where K = the number of studies (Rosenthal & Rubin, 1982). Weighted d can be found using the following formula:

$$d = \Sigma wd / \Sigma w \quad (41)$$

where d is the unweighted effect size and w is the reciprocal of the estimated variance of d in each of the studies to be aggregated in the meta-analysis. This can be calculated using the following equation:

$$w = 2N/(8+d^2) \quad (42)$$

Where N= the total sample size in the study for both the experimental and control groups. If the resulting chi-square is significant the effect sizes in the meta-analysis are heterogeneous. When this occurs moderators of effect sizes will be examined.

Power Analysis

Using the average effect size estimate as the alternative, power was calculated for each experiment based on the degrees of freedom for each significance test used in the original analysis. Power calculations will be based on the cube root normal approximation of the noncentral F distribution (Laubscher, 1960; Severo & Zelen, 1960), as previously described in Chapter II. Power will also be computed for each study using Cohen's (1988) definitions of small, medium, and large effect size for all eligible tests (for example f = .1, .25, and .4, respectively). All power computations will be performed with $\alpha = .05$. Four power estimates were obtained for each experiment. These estimates

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

were then averaged across experiments to obtain summary power estimates. Here the experiment, not the study, was the unit of analysis.

Results

An exhaustive search of electronic databases and review articles revealed 18 articles which met the minimum methodological criteria outlined above. Three additional unpublished studies were obtained by contacting researchers working in the area. The coding sheets for this study are displayed in Table 1. Fifteen studies were examined and not included in the meta-analysis because they did not employ random assignment or use statistical control for baseline differences.

Description of eligible studies

The eighteen studies included 30 separate treatments for which an effect size could be computed. Sixteen of the treatment were classified as cognitive-behavioral and fourteen of the studies were coded as traditional educational. The mean number of participants in the treatment groups was 31, the mean for the control group was 30. However, this was greatly altered by one study that had 299 participants (Marlatt, et al, 1998). After removing this study the average number of participants in the experimental groups was 27, and the mean of the control groups was 25. One of the main reasons for the small sample sizes demonstrated in this study was the high level of attrition. Attrition among these studies ranged from 0% to 41%. The mean attrition was 19%.Subject selection criteria was quite variable with some studies selecting connivance sample of students in a class and other selecting only heavy drinkers. Table 2 displays descriptive information of each of the studies included in the meta-analysis.

Significance Level

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In determining whether the set of studies was significant it was decided that the most rigorous test of the intervention should be examined. For all studies the results of the latest follow-up on drinking rates was examined. Mean or total consumption was used for 19 of the studies, while 2 studies did not report this information so a measure of negative consequences was used. Examining the significance levels reported in the individual studies revealed that 11 (52%) reported significant reductions in drinking or drinking problems, while 10 (48%) did not. Equation 22 was used to determine the overall significance for the interventions. The z score combination of all 21 studies was large ($z = 6.50$, $p < .001$). Since the studies varied greatly in sample size, the analysis was repeated weighting the results for sample size. The resulting z was slightly smaller, but still significant ($z = 5.13$, $p < .001$).

Since only 21 studies were included in the meta-analysis, the result may have been vulnerable to the "file drawer" problem discussed above. Although only 3 studies were culled from unpublished sources, it appeared that journals were not totally unwilling to publishing null findings, since almost half of the studies reported null results. In any case, it is still interesting to examine how many unpublished studies would be necessary to decrease the z found earlier so $p > .05$. This can be done by rearranging equation 22, solving for n, and subtracting a constant equal to the number of studies in the original analysis. The resulting "fail-safe" number for these interventions was 307. The likelihood that there are 18 times as many unpublished studies using sufficient methodological criteria as published studies in this meta-analysis seems highly remote.

Magnitude of the effect

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Effect sizes were computed for 19 of the 21 articles. Two of the articles did not report adequate information to calculate an effect size. Of the 19 articles that effects sizes could be calculated for, many had multiple interventions leading to effect sizes being created for 30 different treatments. Seventeen of the nineteen articles used an index of total alcohol consumption as their primary outcome measure. Effect sizes were computed on this measure for all studies where it was available. The other two articles used a measure of negative consequences as their outcome measure. For these two studies effect sizes were calculated using negative consequences. All 30 interventions reported outcomes at a post test shortly following the end of the intervention. In addition to this, 14 of the treatments reported results from a longer follow-up time. These results were coded separately to assess long term effects of treatment. The 30 treatments at post-test revealed an average effect in the small to medium range, $\underline{d}$ = .36 (95% CI, .23 - .49). The range of the effects was quite diverse, from -.45 to 1.01. The effect sizes for all of the interventions are displayed in Table 3. Using equation 40, the effect sizes were then examined for homogeneity. A significant chi-square revealed that the effect sizes were heterogeneous indicating the need to examine potential moderators.

Effect of moderators

The testing of potential moderators proved to be a difficult task. Only one study broke out effects by gender. Other potential moderators were also excluded because of limited information provided in the published reports. However, the main moderator of interest, type of intervention was reported in all studies. The studies were tested to examine if cognitive-behavioral interventions produced differential effects than educational treatments. These intervention types were chosen based on the qualitative

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

distinction made by Wood (1998). Traditional educational approaches consisted of both information/values clarification and experiential programs. This approach is based on early social psychological approaches to attitude change (e.g., Hovland, Janis & Kelley, 1953). Cognitive behavioral approaches were even more varied. These approaches are based in social learning theory (Bandura, 1977; 1982) and attitude change and self regulation theory (Higgins, 1996; Petty & Cacioppo, 1984). These interventions consisted of self monitoring, normative feedback, lifestyle change, skills training, expectancy challenge, or motivational interviewing. Sixteen interventions used cognitive-behavioral approaches, while 14 used traditional educational approaches. A t-test indicated that cognitive-behavioral approaches ($\underline{d}$ = .53, 95% CI = .34-.72) were superior to traditional educational approaches ($\underline{d}$ = .17, 95% CI = .02-.31; $\underline{t}(28) = 3.2$, $\underline{p} <$ .01). Figure 1 displays this relationship graphically. Homogeneity tests were then performed on both groups of effect sizes. The chi-square test was not significant for the cognitive-behavioral approaches, χ^2 (15) = 24.59, $\underline{p} > .05$, or the traditional educational approaches, χ^2 (13) = 14.54, $\underline{p} > .05$, indicating homogeneity of effect sizes within both type of treatment groups.

Long Term Effects

Finally, the long term effects of treatment were examined by taking the 14 interventions that computed long-term follow-up and comparing them to the short-term follow-up results reported in the same study. Both cognitive-behavioral and traditional educational approaches were included in this analysis because this analysis was within study. There were not a sufficient number of studies to examine long-term effects within

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

treatment type. A dependent t-test indicated no significant differences between short-term (d = .26) and long-term (d = .39) follow-up, t(13) = 1.74, p = .105.

Power Analysis

A power analysis was then conducted on the eligible studies, using d = .53 for tests examining cognitive-behavioral interventions and d = .17 for educational interventions. Only outcome measures involving alcohol consumption variables were examined since this reflects the effect sizes used. Fourteen studies used at least one cognitive-behavioral intervention. Power to detect an effect size of d = .53 was .42 for these studies. Seven studies used at least one educational intervention. Power to detect an effect size of d =.17 was .13. Power to detect small, medium and large effect sizes for these studies is reported in Table 4.

Discussion

Results indicate that interventions to reduce college student drinking are effective. However, the effectiveness of these interventions varies widely by type of intervention used. Traditional educational approaches produced small effect sizes (d = .17), while cognitive-behavioral (CB) interventions produced medium sized effects (d = .53). This rather large difference indicates the superiority of CB interventions. In light of this finding, CB interventions are strongly recommended over traditional educational approaches. However, since relatively few interventions using CB techniques have been performed, not enough information exists to recommend one type of CB treatment over another. More research in this area is needed comparing different types of CB interventions. Components research is needed in the area to isolate the "active" influences of these interventions. This meta-analysis also focused specifically on

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

individual level interventions to reduce college student drinking. Since, environmental level interventions were not included in the analyses no evaluation of there effectiveness can be assessed.

One variable that was not discussed in any of the articles included in the meta-analysis was stage of change. This variable has been used to tailor interventions to individuals at different levels of readiness (Prochaska & DiClemente, 1983). Interventions tailored to stage of change have been effective in many other areas of behavior change including smoking cessation and ultraviolet light reduction (Prochaska et al., 1993; Rossi et al., 1998). As I have argued elsewhere, (Maddock, Laforge & Rossi, 1998), treatments designed to enhance motivation such as motivational interviewing have been shown to be more effective than cognitive behavioral treatments for individuals in precontemplation (Project MATCH Research Group, 1998). A logical course of action following this would be to tailor interventions to an individual's stage of change. Thus, a person in precontemplation would receive motivational enhancement or normative feedback, while an individual in preparation would receive an action based intervention such a skills training. This type of continuum of treatment might prove more efficacious in reducing college student drinking than the current one size fits all approach.

Results from the power analysis indicate that as a whole these studies have been severely under-powered. Now that an average effect size for CB intervention is known, power analyses for future studies should be greatly enhanced. For example, to perform a simple independent t-test, the study would need 64 participants per group to achieve power = .80 assuming $\underline{d} = .51$. This sample size is greater than all of the studies

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

employing cognitive-behavioral interventions except one (Marlatt et al., 1998), indicating a severe lack of power to detect effects for this type of treatment.

Limitations of the meta-analysis

All meta-analyses face some limitations. This study is no exception. The studies included in this analysis varied on length of follow-up, outcome measures, and statistical control procedures. While these measures were kept as consistent as possible, differences did occur. Reporting procedures in the articles also produced somewhat biased results. Only three of the articles reported effect sizes. Standard deviations of treatment outcomes were often not reported. Estimations of these deviations were gleaned from summary statistics, yet this often produces some unknown amount of bias. Methodological quality also varied greatly by study. Although only those studies that used random assignment or statistical control of baseline differences were included several other differences existed such as high levels of attrition or mixed units of assignment and analysis. Although all participants were college students, study populations also differed. In some studies only high-risk drinkers were selected, and even this definition varied from study to study. In other studies, convenience samples from college classes were used and compared to other classes. Measurement instruments also differed by study. No one measure of alcohol use or problems was employed by even a majority of the studies. Although almost all of the studies reported some measure of drinking outcome this was often not the same. Despite all of these limitations, the use of meta-analysis is still seen as an improvement for synthesizing findings across studies. Hopefully, increased use of this technique will lead to better standardization in

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

instruments and procedures across studies, which will increase the validity of this technique.

Implications for future research

Heavy drinking on college campuses remains a central problem of many higher education administrators even though many efforts have been used to alter problematic usage. Better methodological standards must be adhered to. Only 21 studies met minimal methodological standards to be included in the meta-analysis. Among these studies several employed unsound methodological practices such as randomizing individuals by group and analyzing data on individuals. This ignores the interclass correlation and distorts effect sizes. In short, well-designed methodological studies are still sorely needed in this area. On a positive note, CB interventions appear to be effective in reducing college student drinking. These interventions, if combined with environmental level interventions could prove to be highly effective in reducing college student drinking for entire campuses. Further research is still needed to isolate the active ingredients of these interventions to narrow the broad range of cognitive-behavioral interventions.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 117-144). Mahwah, NJ, Lawrence Erlbaum Associates.

Agosti, V. (1994). The efficacy of controlled trials of alcohol misuse treatments in maintaining abstinence. The International Journal of the Addictions, 29, 759-769.

Agosti, V. (1995). The efficacy of treatments in reducing alcohol consumption: A meta-analysis. The International Journal of the Addictions, 30, 1067-1077.

*Agostinelli, G., Brown, J. M. & Miller, W. R. (1995). Effects of normative feedback on consumption among heavy drinking college students. Journal of Drug Education, 25, 31-40.

Agresti, A. (1990). Categorical data analysis. New York: Wiley.

*Baer, J. S., Marlatt, G. A., Kivlahan, D. R., Fromme, K., Larimer, M. E., & Williams, E. (1992). An experimental test of three methods of alcohol risk reduction with young adults. Journal of Consulting and Clinical Psychology, 60, 974-979.

Bailar, J. C. (1997). The promise and problems of meta-analysis. The New England Journal of Medicine, 337, 559-560.

Ballie, A.J., Mattick, R.P., Hall, W. & Webster, P.(1994). Meta-analytic review of the efficacy of smoking cessation interventions. Drug and Alcohol Review, 13, 157-170.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. Psychological Bulletin, 99, 388-399.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Academic Press.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Commission on Substance Abuse at Colleges and Universities (1994, June). Rethinking rites of passage: Substance abuse on America's campuses. New York: Center on Addiction and Substance Abuse at Columbia University.

Cooper, H. M. & Hedges, L. V. (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Cooper, H. M. & Rosenthal, R. (1980). Statistical verses traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442-449. New York: McGraw-Hill.

*Cronin, C. (1996). Harm reduction for alcohol-use related problems among college students. Substance Use and Misuse, 31, 2029-2037.

*Darkes, J., & Goldman, M. S. (1993). Expectancy challenge and drinking reduction: Experimental evidence for a mediational process. Journal of Consulting and Clinical Psychology, 61, 344-353.

*Darkes, J. & Goldman, M. S. (1998). Expectancy challenge and drinking reduction: Process and structure in the alcohol expectancy network. Experimental and Clinical Psychopharmacology, 6, 64-76.

*Dennison, D. & Prevet, T. (1980). Improving alcohol-related disruptive behaviors through health instruction. The Journal of School Health, April, 206-208.

*Engs, R. C. (1977). Let's look before we leap: The cognitive and behavioral evaluation of a university alcohol education program. Journal of Alcohol and Drug Education, 22, 39-48.

Fisher, R. A. (1928). Statistical methods for research workers (2nd ed.). London: Oliver & Boyd.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). London: Oliver & Boyd.

Fleiss, J. L. (1981). Statistical methods for rates and proportions. (2nd ed.) New York: Wiley.

*Fromme, K., Kivlahan, D. R., & Marlatt, G. A. (1986). Alcohol expectancies, risk identification, and secondary prevention with problem drinkers. Advances in Behavior Research and Therapy, 8, 237-251.

*Garvin, R. B., Alcorn, J. D., & Faulkner, K. K. (1990). Behavioral strategies for alcohol abuse prevention with high risk college males. Journal of Alcohol and Drug Education, 36, 23-34.

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

Glass, G. V., McGaw, B. & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

*Gonzalez, G. M. (1980). The effect of a model alcohol education module on college students' attitudes, knowledge, and behavior related to alcohol use. Journal of Alcohol and Drug Education, 25, 1-12.

Goodwin, D. (1988). Alcoholism: Who gets better and who does not. In R. M. Rose & J. Barrett (Eds.), Alcoholism: Origins and Outcomes. New York: Raven Press, pp. 281-292.

Haddock, C. K., Rindskopf, D., Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. Psychological Methods, 3, 339-353.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Hedges, L. V. & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Heinsman, D. T. & Shadish, W. R. (1996). Assignment measures in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? Psychological Methods, 1, 154-169.

*Henderson, M. J. & Goldman, M. S. (1987, November). Effect of a social manipulation on alcohol expectancies and subsequent drinking. Paper presented at the Annual Meeting of the Association for the Advancement of Behavior Therapy, Boston, MA.

Hovland, C. I., Janis, I. L., & Kelly, H. H. (1953). Communication and persuasion: Psychological studies of opinion change. New Haven, CT: Yale University Press.

Hunter, J. E. & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

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

Johnson, B. T., Mullen, B. & Salas, E. (1995). Comparison of three major meta-analytic approaches. Journal of Applied Psychology, 80, 94-106.

*Jones, L. M., Silva, L. Y., & Richman, C. L. (1995). Increased awareness and self-challenge of alcohol expectancies. Substance Abuse, 16, 77-85.

*Kivlahan, D. R., Marlatt, G. A., Fromme, K., Coppel, D. B. & Williams, E. (1990). Secondary prevention with college drinkers: Evaluation of an alcohol skills training program. Journal of Consulting and Clinical Psychology, 58, 805-810.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

LeLorier, J., Gregoire, G., Benhaddad, A., LaPierre, & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large, randomized, controlled trials. The New England Journal of Medicine, 337, 536-542.

Lindstrom, L. (1992). Managing alcoholism: Matching clients to treatments. New York: Oxford University Press, p . 30.

Maddock, J.E., Laforge, R.G., & Rossi, J.S. (1998). The challenge of the precontemplator: Rethinking motivational readiness in Project MATCH. The Addictions Newsletter, 5(3), 21-22.

*Maddock, J. E.; Wood, M. D.; Davidoff, O. J.; Colby S. M.; & Monti P. M..(1999) Alcohol expectancy challenge and alcohol use: Examination of a controlled trial. Paper submitted to the Annual Meeting of the Research Society on Alcoholism, Santa Barbara, CA.

*Marlatt, G. A., Baer, J. S., Kivlahan, D. R., Dimeff, L. A., Larimer, M. E., Quigley, L. A., Somers, J. M., & Williams, E. (1998). Screening and brief intervention for high-risk college student drinkers: Results from a 2-year follow-up assessment. Journal of Consulting and Clinical Psychology, 66, 604-615.

*Marlatt, G. A., Pagano, R. R., Rose, R. M. & Marques, J. K. (1984). Effects of meditation and relaxation training upon alcohol use in male social drinkers. In D. H. Shapiro & R. N. Walsh (Eds.), Meditation: Classic and contemporary perspectives. New York: Aldine Publishing.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

*Massey, R. F. & Goldman, M. S. (1988, August). Manipulating expectancies as a means of altering alcohol consumption. Paper presented at the 96th Annual Convention of the American Psychological Association, Atlanta, GA.

*Meacci, W. G. (1990). An evaluation of the effects of college alcohol education on the prevention of negative consequences. Journal of Alcohol and Drug Education, 35, 66-72.

Moskowitz, J. M. (1989). The primary prevention of alcohol problems: A critical review of the research literature. Journal of Studies on Alcohol, 50, 54-88.

*Murphy, T. J., Pagano, R. R. & Marlatt, G. A. (1986). Lifestyle modification with heavy alcohol drinkers: Effects of aerobic exercise and meditation. Addictive Behaviors, 11, 175-186.

Petty, R. E. & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. Journal of Personality and Social Psychology, 46, 69-81.

Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: Towards an integrative model of change. Journal of Consulting and Clinical Psychology, 51, 390-395.

Prochaska, J., C. DiClemente, C.C., Velicer, W.F., Rossi, J.S. (1993). Standardized, individualized, interactive, and personalized self-help programs for smoking cessation. Health Psychology 12, 399-405.

Project MATCH Research Group (1997). Matching alcoholism treatments to client heterogeneity. Project MATCH post-treatment drinking outcomes. Journal of Studies on Alcohol, 58(1), 7-29.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Ray, J. W. & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? Journal of Consulting and Clinical Psychology, 64, 1316-1325.

*Robinson, J. (1981). A comparison of three alcohol instruction programs on the knowledge, attitudes and drinking behaviors of college students. Journal of Drug Education, 11, 157-166.

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

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

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

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

Rosenthal, R. (1991). Meta-analytic procedures for social research. Newbury Park, CA: Sage.

Rosenthal, R. & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. Behavioral and Brain Sciences, 3, 377-415.

Rosenthal, R. & Rubin, D. (1988). Comment: Assumptions and procedures in the file drawer problem. Statistical Science, 3, 120-125.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-198). Mahwah, NJ, Lawrence Erlbaum Associates.

Rossi, J. S., Redding, C. A., Maddock, J. E., Cottrill, S. D., & Weinstock, M. A. (1998). Effectiveness of stage-matched interventions for skin cancer prevention in high-risk beach bathers. Proceedings of the 19th Annual Meeting of the Society of Behavioral Medicine, Annals of Behavioral Medicine.

*Rozelle, G. R. (1980). Experiential and cognitive small group approaches to alcohol education for college students. Journal of Alcohol and Drug Education, 26, 40-54.

Sandercock, P. (1989). The odds ratio: A useful tool in neurosciences. Journal of Neurology, Neurosurgery and Psychiatry, 52, 817-820.

Schukit, M. (1992). Treatment of alcoholism in office and outpatient settings. In J. H. Mendelson and N. K. Mello (Eds.), Medical Diagnosis and Treatment of Alcoholism.

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

Snedecor, G. W. & Cochran, W. G. (1989). Statistical Methods (8th ed.) Ames: Iowa State University Press.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Valliant, G. E. (1988). What can long-term follow-up teach us about relapse and prevention of relapse in addiction? British Journal of Addiction, 257, 1147-1156.

Wechsler, H., Davenport, A., Dowdall, G., Moeykens, B., & Castillo, S. (1994). Health and behavioral consequences of binge drinking in college. Journal of the American Medical Association, 272, 1672-1677.

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

Wood, M. D. (in press). Preventative interventions to reduce alcohol and other drug abuse among college students: Implications from alcohol abuse prevention interventions. Position paper solicited by the network of Colleges and Universities Committed to the elimination of alcohol and drug abuse, in press.

* Studies included in the meta-analysis

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table1

Coding Sheet Used in the Meta-Analysis

ID # _____

Meta-analysis coding sheet

1. Author(s): _____
2. Source: 1---Journal 2---Book
   3---Thesis/Dissertation 4---Conference presentation
   5---All other sources (define) _____
3. Year: __ __
4. Time in minutes to code report __ __
5. Sampling: 1—Random 2---Judgment or convenience
   3---Reported in another source 4---Not reported
6. Total sample size: __ __ __ __
7. % female __ __%
8. Age data reported on sample: Yes No
9. Mean age: __ __ . __
10. Standard deviation of age: __ __ . __
11. % over 21 __ __%
12. Are effect sizes broken down by gender? Yes No
13. % attrition _____
14. Type of intervention:
    1—Educational
    2---Skills Training
    3---Motivational Enhancement
    4---Expectancy Manipulation
    5---Policy
    6---Other _____
15. Type of sample
    1 – General college
    2—Alcohol offenders
    3—High risk Drinkers (Define)
    4—Incoming students
    5—Students over 21
    6 –Other (define) _____

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 1 Continued

Outcome measures: 1 = drinks/episode 2 = episodes per week/month
3 = peak drinks 4 = problem indices 5= other
Subgroup 1= All 2= males Only 3= females only
Type of Intervention: _____ (use numbers from 14 above)
Covariate 1 = yes 2 = no

	Control group	Treatment group
Mean	_____	_____
S.D.	_____	_____
N	_____	_____
Test Statistic	_____	_____
t test	_____	
d.f.	_____	

analysis of variance

	Sum of square	d.f.	F
Treatment	_____	___	___
Total	_____	___	___

d = _____

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 2

Studies Used in the Meta-Analysis

Study	N treatment	N control	Type of treatment	Subject selection criteria	% attrition	d
Cognitive-Behavioral						
Agostinelli et al. (1995)	12	13	Normative feedback	80 drinks/month	12%	.99
Darkes & Goldman (1993)	15	18	Expectancy Challenge	Males 6-40 drinks week	32%	.59
Darkes & Goldman (1998)	36	18	Expectancy Challenge	Males 6-42 drinks week	19%	.73
Jones et al. (1995)	30	30	Exp. Challenge (A)	Drinkers in a class	10%	.18
Jones et al. (1995)	30	30	Exp. Challenge (B)	Drinkers in a class	10%	.24
Kivlahan et al. (1990)	14	10	Skills Training	At least 1 alc. problem	20%	1.01
Garvin et al. (1990)	20	20	Self monitoring	Fraternity pledges	NR	-.02
Marlatt et al. (1984)	10	14	Meditation	1.5 drinks a day	7%	.79
Marlatt et al. (1984)	8	14	Progressive relaxation	1.5 drinks a day	7%	.78
Marlatt et al. (1984)	9	14	Bibliotherapy	1.5 drinks a day	7%	.89
Murphy et al. (1986)	9	13	Exercise	45 drinks per month	24%	.84
Murphy et al. (1986)	9	13	Meditation	45 drinks per month	27%	.00
Massey & Goldman (1988)	NR	NR	Expectancy Challenge	Females	NR	.73
Henderson & Goldman (1987)	NR	NR	Expectancy Challenge	Females	NR	.44
Maddock et al. (1999)	21	23	Expectancy Challenge	4/5 drinks per episode	19%	.14
Marlatt et al. (1998)	143	156	Motivational Int.	High-risk in H.S.	14%	.14

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Traditional Educational						
Dennison & Prevet (1980)	27	26	Experiential	Class enrollment	0%	-.23
Engs (1977)	50	33	Informational	Residence hall members	17%	.24
Garvin et al. (1990)	20	20	Informational	Fraternity pledges	NR	.05
Gonzalez (1980)	50	44	Informational	Class Volunteers	25%	.40
Kivlahan et al. (1990)	14	10	Informational	At least 1 alc. problem	20%	.33
Robinson (1981)	20	20	Implicit instruction	Students in a class	25%	-.35
Robinson (1981)	23	20	Explicit instruction	Students in a class	12%	.34
Robinson (1981)	23	20	Value clarification	Students in a class	19%	.18
Rozelle (1980)	46	46	Educational	Students in a class	26%	.12
Rozelle (1980)	52	46	Experiential	Students in a class	26%	.23
Meacci (1990)	72	63	Informational	Students in classes	41%	.15
Massey & Goldman (1988)	NR	NR	Informational	Females	NR	.28
Henderson & Goldman (1987)	NR	NR	Informational	Females	NR	.26
Cronin et al. (1996)	41	41	Informational	Students in classes	36%	.43

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 3

Stem and Leaf Display of 30 Effect Sizes for Post-Test

Stem	Leaf
1.0	1
0.9	9
0.8	4,9
0.7	3,3,8,9
0.6	
0.5	9,
0.4	0,3,4
0.3	3,5
0.2	3,4,4,4,8
0.1	2,4,4,5,8
0.0	0,5
-0.0	2
-0.1	8
-0.2	3
-.03	
-.04	5

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4

Average Power for Alcohol Intervention Studies

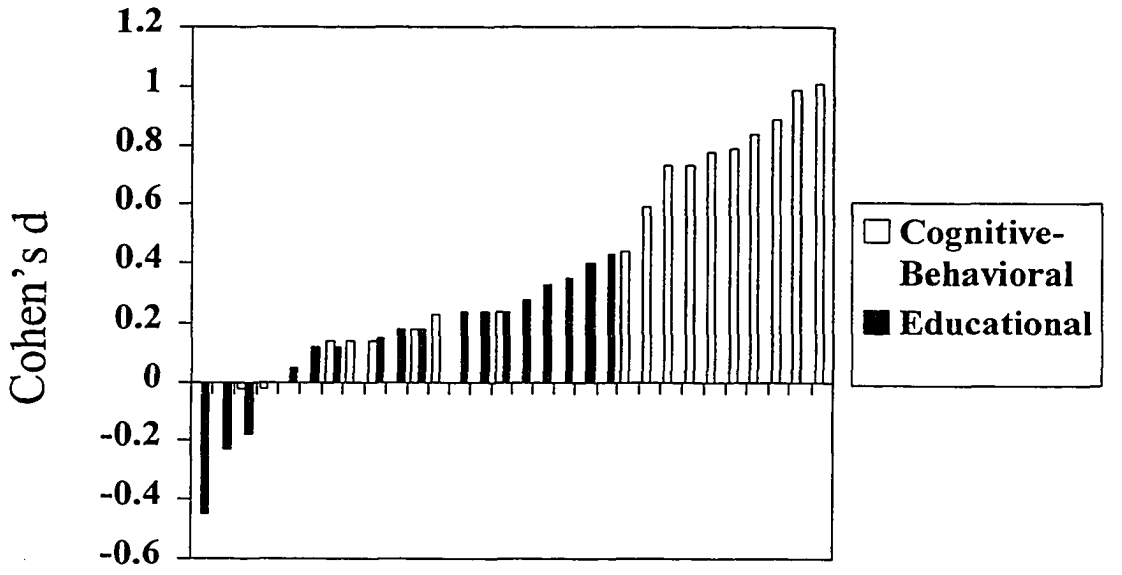
Type of Intervention	n	Effect Size		
		Small	Medium	Large
Cognitive-Behavioral	14	.11	.41	.69
Traditional Educational	7	.18	.69	.94
Total	18	.12	.49	.78

Note: n is the number of experiments in each category. Some experiments contained both cognitive-behavioral and traditional educational interventions.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 1

Comparison of Cognitive-Behavioral and Traditional Educational Effect Sizes



Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Bibliography

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 117-144). Mahwah, NJ, Lawrence Erlbaum Associates.

Agosti, V. (1994). The efficacy of controlled trials of alcohol misuse treatments in maintaining abstinence. The International Journal of the Addictions, 29, 759-769.

Agosti, V. (1995). The efficacy of treatments in reducing alcohol consumption: A meta-analysis. The International Journal of the Addictions, 30, 1067-1077.

Agostinelli, G., Brown, J. M. & Miller, W. R. (1995). Effects of normative feedback on consumption among heavy drinking college students. Journal of Drug Education, 25, 31-40.

Agresti, A. (1990). Categorical data analysis. New York: Wiley.

Aron, A. & Aron, E. N. (1999). Statistics for Psychology (2nd Edition). Upper Sadle River, NJ: Prentice-Hall, Inc.

Baer, J. S., Marlatt, G. A., Kivlahan, D. R., Fromme, K., Larimer, M. E., & Williams, E. (1992). An experimental test of three methods of alcohol risk reduction with young adults. Journal of Consulting and Clinical Psychology, 60, 974-979.

Ballie, A. J., Mattick, R. P., Hall, W. & Webster, P. 91994). Meta-analytic review of the efficacy of smoking cessation interventions. Drug and Alcohol Review, 13, 157-170.

Bailar, J. C. (1997). The promise and problems of meta-analysis. The New England Journal of Medicine, 337, 559-560.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. Psychological Bulletin, 99, 388-399.

Borenstien, M., Cohen, J. & Rothstein, H. (1997). Power and Precision. Teaneck, NJ: Biostat.

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

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

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

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

Christensen, J. E. & Christensen, C. E. (1977). Statistical power analysis of health, physical education, and recreation research. Research Quarterly, 48, 204-208.

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

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Academic Press.

Cohen, J. (1989). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Commission on Substance Abuse at Colleges and Universities (1994, June). Rethinking rites of passage: Substance abuse on America's campuses. New York: Center on Addiction and Substance Abuse at Columbia University.

Cooper, H. M. & Hedges, L. V. (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Cooper, H. M. & Rosenthal, R. (1980). Statistical verses traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442-449.

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

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

Cronin, C. (1996). Harm reduction for alcohol-use related problems among college students. Substance Use and Misuse, 31, 2029-2037.

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Darkes, J., & Goldman, M. S. (1993). Expectancy challenge and drinking reduction: Experimental evidence for a mediational process. Journal of Consulting and Clinical Psychology, 61, 344-353.

Darkes, J. & Goldman, M. S. (1998). Expectancy challenge and drinking reduction: Process and structure in the alcohol expectancy network. Experimental and Clinical Psychopharmacology, 6, 64-76.

Davis, J. M., Janicak, P. G., Wang, Z. & Gibbons, R. D. (1992). The efficacy of psychotropic drugs: Implications for power analysis. Psychopharmacology Bulletin, 28, 151-155.

Dennison, D. & Prevet, T. (1980). Improving alcohol-related disruptive behaviors through health instruction. The Journal of School Health, April, 206-208.

Deyo, R. A. & Patrick, D. L. (1995). The significance of treatment effects: The clinical perspective. Medical Care, 33, AS286-AS291.

Eisenhart, C. (1947). Inverse sine transformation of proportions. In C. Eisenhart, M. W. Hastay, & W. A. Wallis, (Eds.), Selected techniques of statistical analysis for scientific and industrial research production and management engineering (pp. 395-416). New York: McGraw-Hill.

Engs, R. C. (1977). Let's look before we leap: The cognitive and behavioral evaluation of a university alcohol education program. Journal of Alcohol and Drug Education, 22, 39-48.

Fisher, R. A. (1928). Statistical methods for research workers (2nd ed.). London: Oliver & Boyd.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). London: Oliver & Boyd.

Fisher, R. A. (1949). The design of experiments. New York: Hafner.

Fleiss, J. L. (1981). Statistical methods for rates and proportions. (2nd ed.) New York: Wiley.

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

Fromme, K., Kivlahan, D. R., & Marlatt, G. A. (1986). Alcohol expectancies, risk identification, and secondary prevention with problem drinkers. Advances in Behavior Research and Therapy, 8, 237-251.

Garvin, R. B., Alcorn, J. D., & Faulkner, K. K. (1990). Behavioral strategies for alcohol abuse prevention with high risk college males. Journal of Alcohol and Drug Education, 36, 23-34.

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

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

Glass, G. V., McGaw, B. & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Gonzalez, G. M. (1980). The effect of a model alcohol education module on college students' attitudes, knowledge, and behavior related to alcohol use. Journal of Alcohol and Drug Education, 25, 1-12.

Goodwin, D. (1988). Alcoholism: Who gets better and who does not. In R. M. Rose & J. Barrett (Eds.), Alcoholism: Origins and Outcomes. New York: Raven Press, pp. 281-292.

Gravetter, F. J. & Wallnau, L. B. (1996). Statistics for the behavioral sciences (4th Edition). New York: West Publishing Co.

Haase, R. F. (1976). Power analysis of research in counselor education. Counselor Education and Supervision, 14, 124-132.

Haddock, C. K., Rindskopf, D., Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. Psychological Methods, 3, 339-353.

Harlow, L. L. (1997). Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 1-20). Mahwah, NJ, Lawrence Erlbaum Associates.

Hedges, L. V. & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Heinsman, D. T. & Shadish, W. R. (1996). Assignment measures in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? Psychological Methods, 1, 154-169.

Henderson, M. J. & Goldman, M. S. (1987, November). Effect of a social manipulation on alcohol expectancies and subsequent drinking. Paper presented at the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Annual Meeting of the Association for the Advancement of Behavior Therapy, Boston, MA.

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

Hovland, C. I., Janis, I. L., & Kelly, H. H. (1953). Communication and persuasion: Psychological studies of opinion change. New Haven, CT: Yale University Press.

Howard, M. O. & Howard, D. A. (1992). Citation analysis of 541 articles published in drug and alcohol journals: 1984-1988. Journal of Studies on Alcohol, 53, 427-434.

Hunter, J. E. & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

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

Huysamen, G. K. (1996). Some methodological issues in health psychology research. South African Journal of Psychology, 26, 10-15.

Johnson, B. T., Mullen, B. & Salas, E. (1995). Comparison of three major meta-analytic approaches. Journal of Applied Psychology, 80, 94-106.

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

Jones, L. M., Silva, L. Y., & Richman, C. L. (1995). Increased awareness and self-challenge of alcohol expectancies. Substance Abuse, 16, 77-85.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

Kalichman, S. C., Carey, M. P., & Johnson, B. T. (1996). Prevention of sexually transmitted HIV infection: A meta-analytic review of the behavioral outcome literature. Annals of Behavioral Medicine, 18, 6-15.

Katzell, R. A. & Dyer, F. J. (1977). Differential validity revived. Journal of Applied Psychology, 62, 137-145.

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

Kazis, L. E., Anderson, J. J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. Medical Care, 27, S178-S189.

Kivlahan, D. R., Marlatt, G. A., Fromme, K., Coppel, D. B. & Williams, E. (1990). Secondary prevention with college drinkers: Evaluation of an alcohol skills training program. Journal of Consulting and Clinical Psychology, 58, 805-810.

Kosciulek, J. F. (1993). The statistical power of vocational evaluation research. Vocational Evaluation and Work Adjustment Bulletin, 26, 142-145.

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

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

LeLorier, J., Gregoire, G., Benhaddad, A., LaPierre, & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large, randomized, controlled trials. The New England Journal of Medicine, 337, 536-542.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Lindstrom, L. (1992). Managing alcoholism: Matching clients to treatments. New York: Oxford University Press, p . 30.

Maddock, J.E., Laforge, R.G., & Rossi, J.S. (1998). The challenge of the precontemplator: Rethinking motivational readiness in Project MATCH. The Addictions Newsletter, 5(3), 21-22.

Maddock, J. E.; Wood, M. D.; Davidoff, O. J.; Colby S. M.; & Monti P. M..(1999) Alcohol expectancy challenge and alcohol use: Examination of a controlled trial. Paper submitted to the Annual Meeting of the Research Society on Alcoholism, Santa Barbara, CA.

Marlatt, G. A., Baer, J. S., Kivlahan, D. R., Dimeff, L. A., Larimer, M. E., Quigley, L. A., Somers, J. M., & Williams, E. (1998). Screening and brief intervention for high-risk college student drinkers: Results from a 2-year follow-up assessment. Journal of Consulting and Clinical Psychology, 66, 604-615.

Marlatt, G. A., Pagano, R. R., Rose, R. M. & Marques, J. K. (1984). Effects of meditation and relaxation training upon alcohol use in male social drinkers. In D. H. Shapiro & R. N. Walsh (Eds.), Meditation: Classic and contemporary perspectives. New York: Aldine Publishing.

Massey, R. F. & Goldman, M. S. (1988, August). Manipulating expectancies as a means of altering alcohol consumption. Paper presented at the 96th Annual Convention of the American Psychological Association, Atlanta, GA.

Meacci, W. G. (1990). An evaluation of the effects of college alcohol education on the prevention of negative consequences. Journal of Alcohol and Drug Education, 35, 66-72.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115.

Minium, E. W., Clarke, R. C. & Coladarci, T. (1999). Elements of Statistical Reasoning (2nd Edition). New York: John Wiley and sons, Inc.

Moskowitz, J. M. (1989). The primary prevention of alcohol problems: A critical review of the research literature. Journal of Studies on Alcohol, 50, 54-88.

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

Mulaik, S. A., Raju, N. S. & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-198). Mahwah, NJ, Lawrence Erlbaum Associates.

Murphy, T. J., Pagano, R. R. & Marlatt, G. A. (1986). Lifestyle modification with heavy alcohol drinkers: Effects of aerobic exercise and meditation. Addictive Behaviors, 11, 175-186.

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

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

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

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

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

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

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

Orme, J. G. & Tolman, R. M. (1986). The statistical power of a decade of social work education research. Social Service Review, 60, 620-632.

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

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

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

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.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Petty, R. E. & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. Journal of Personality and Social Psychology, 46, 69-81.

Prentice, D. A. & Miller, D. T. (1992). When small effects are impressive. Psychological Bulletin, 112, 160-164.

Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: Towards an integrative model of change. Journal of Consulting and Clinical Psychology, 51, 390-395.

Prochaska, J., C. DiClemente, C.C., Velicer, W.F., Rossi, J.S. (1993). Standardized, individualized, interactive, and personalized self-help programs for smoking cessation. Health Psychology 12, 399-405.

Project MATCH Research Group (1997). Matching alcoholism treatments to client heterogeneity. Project MATCH post-treatment drinking outcomes. Journal of Studies on Alcohol, 58(1), 7-29.

Ray, J. W. & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? Journal of Consulting and Clinical Psychology, 64, 1316-1325.

Robinson, J. (1981). A comparison of three alcohol instruction programs on the knowledge, attitudes and drinking behaviors of college students. Journal of Drug Education, 11, 157-166.

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

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

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

Rosenthal, R. (1990). How are we doing in soft psychology? American Psychologist, 45, 775-776.

Rosenthal, R. (1991). Meta-analytic procedures for social research. Newbury Park, CA: Sage.

Rosenthal, R. & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. Behavioral and Brain Sciences, 3, 377-415.

Rosenthal, R. & Rubin, D. (1988). Comment: Assumptions and procedures in the file drawer problem. Statistical Science, 3, 120-125.

Rossi, J. S. (1984). Statistical power of psychological research. Unpublished doctoral dissertation, University of Rhode Island.

Rossi, J. S. (1985). Tables of effect size for z score tests of difference between proportions and between correlation coefficients. Educational and Psychological Measurement, 45, 737-743.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Rossi, J. S. (1991). Power for k-groups MANOVA. Unpublished computer program. University of Rhode Island.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-198). Mahwah, NJ, Lawrence Erlbaum Associates.

Rossi, J. S., Redding, C. A., Maddock, J. E., Cottrill, S. D., & Weinstock, M. A. (1998). Effectiveness of stage-matched interventions for skin cancer prevention in high-risk beach bathers. Proceedings of the 19th Annual Meeting of the Society of Behavioral Medicine, Annals of Behavioral Medicine.

Rozelle, G. R. (1980). Experiential and cognitive small group approaches to alcohol education for college students. Journal of Alcohol and Drug Education, 26, 40-54.

Sandercock, P. (1989). The odds ratio: A useful tool in neurosciences. Journal of Neurology, Neurosurgery and Psychiatry, 52, 817-820.

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

Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-198). Mahwah, NJ, Lawrence Erlbaum Associates.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Schukit, M. (1992). Treatment of alcoholism in office and outpatient settings. In J. H. Mendelson and N. K. Mello (Eds.), Medical Diagnosis and Treatment of Alcoholism.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

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

Sindelar, P. T., Allman, C., Monda, L., & Vail, C. O. (1988). The power of hypothesis testing in special education efficacy research. Journal of Special Education, 22, 284-296.

Snedecor, G. W. & Cochran, W. G. (1989). Statistical Methods (8th ed.) Ames: Iowa State University Press.

Spence, J. T., Cotton, J. W., Underwood, B. J., & Duncan, C. P. (1990). Elementary Statistics (5th Edition). Englewood Cliffs, NJ: Prentice Hall.

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

Thomas, L. & Juanes, F. (1996). The importance of statistical power analysis: An example from Animal Behaviour. Animal Behaviour, 52, 856-859.

Valliant, G. E. (1988). What can long-term follow-up teach us about relapse and prevention of relapse in addiction? British Journal of Addiction, 257, 1147-1156.

Wechsler, H., Davenport, A., Dowdall, G., Moeykens, B., & Castillo, S. (1994). Health and behavioral consequences of binge drinking in college. Journal of the American Medical Association, 272, 1672-1677.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

Wood, M. D. (1998). Preventative interventions to reduce alcohol and other drug abuse among college students: Implications from alcohol abuse prevention interventions. Position paper solicited by the network of Colleges and Universities Committed to the elimination of alcohol and drug abuse (in press).

Wooley, T.W. (1983). A comprehensive power-analytic investigation of research in medical education. Journal of Medical Education, 58, 710-715.

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.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.