A Post Hoc Statistical Power Analysis and Survey of the Research Published in the *Journal of Athletic Training*

Mark Andrew Tener

A dissertation presented to the
Graduate Faculty of Middle Tennessee State University
In partial fulfillment of the requirements
for the degree Doctor of Arts

December 2000

UMI Number: 9992329

UMI®

UMI Microform 9992329

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

A Post Hoc Statistical Power Analysis and Survey of the Research Published in the *Journal of Athletic Training*

APPROVED:

Graduate Committee:

____
Major Professor

____
Committee Member

____
Committee Member

____
Head of Department of HPERS

____
Dean of Graduate College

ABSTRACT

A Post Hoc Statistical Power Analysis and Survey of the Research Published in the Journal of Athletic Training

Mark Andrew Tener

Statistical power is the probability of finding a true effect in the population under investigation. When a true effect is found, the researcher rejects the null hypothesis and accepts the research hypothesis. Therefore, statistical power is the probability of rejecting the null hypothesis when it is false.

The purpose of this study was to perform a post hoc power analysis of published athletic training research. A sample of articles was selected from the three most recently completed volumes of the Journal of Athletic Training (vols. 32, 33, and 34) to answer the proposed research questions. The information collected from the articles concerned the use of an a priori power analysis, the report of observed power, the report of observed effect size, the reported level of significance, and the report of nonsignificant results. Specific articles were designated for a post hoc power analysis. Using well-known power tables, the statistical power for each article was calculated assuming the authors attempted to detect small, medium, and large effect sizes.

Of the 77 articles surveyed, two articles contained a report of using an a priori power analysis. Four articles contained a report of observed power. None of the articles contained a report of observed effect size. Forty-seven articles contained a report of nonsignificant results, yet only nine of these articles contained a report of insufficient power as a possible reason. Thirty-six articles were used for the post hoc power analysis. The mean powers were .18, .53, and .75 for the assumed small, medium, and large effect sizes, respectively.

It was concluded that athletic training researchers had a poor chance of finding true effects and rejecting the null hypothesis, unless the researchers attempted to detect large effects. Recommendations included: (1) requiring certain information be reported in research articles; (2) reporting clear, well-understood results statements; (3) researching the possible reasons for the neglect of power among athletic training researchers; and (4) having editors and reviewers of the research help to change article submission policies.

ACKNOWLEDGEMENTS

I want to acknowledge Mr. Larry Cook, former athletic director and head football coach of Circleville High School. He helped me get started, as a high school student, in the field of athletic training by providing the money necessary to attend student athletic training camps.

I want to acknowledge Mr. Brian Siddall, head athletic trainer for Ashland University. He taught me the discipline and professionalism of athletic training.

I want to acknowledge Mr. Charles "Skip" Vossler, former head athletic trainer for Ohio University. He gave me the opportunity to sharpen my skills as an athletic trainer during my pursuit for a master's degree.

Finally, I want to acknowledge my dissertation committee: Dr. Jon MacBeth, professor of Health, Physical Education, Recreation, and Safety; Dr. William Whitehill, associate professor of Athletic Training; and Dr. Jane Williams, professor of Educational Leadership. They were persistent with me and extremely understanding, especially when I wanted to change my dissertation topic. I want to include Dr. David Rowe for his input for my dissertation before he accepted another teaching position.

DEDICATIONS

This dissertation is dedicated to the glory of God through His Son, my Lord and Savior, Jesus Christ. Without God's help and provision, my education at Middle Tennessee State University would not have been accomplished.

This dissertation is also dedicated to my wife, Sandra, parents, and family. The tongues of mankind and angels cannot adequately express my love and gratitude for Sandra's sacrifice of moving to Tennessee, leaving friends and family, and becoming the primary "bread winner" for my education. I thank God for my parents and family who have loved and supported me throughout my lifetime.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Introduction

Researchers have been using statistics for many years to analyze the data from their experiments. Statistics is one method in which data can be objectively and uniformly reported and enables researchers to make accurate conclusions and comparisons. Researchers use the statistical techniques to describe the sample data and to test the effect of the treatment. The effect could be a relationship between two variables, or the effect could be a mean difference between groups of subjects that may have received some type of treatment.

Statistical power is the probability of detecting a true effect that researchers expect to find in the population under investigation. True effects are ones that occur due to the relationship or treatment(s) under investigation, not the differences that may occur due to chance or error. When the true effect is discovered, the researcher is able to reject the null hypothesis(es) under investigation. As a researcher increases the statistical power of a study, he/she increases the chance of finding true effects and rejecting the null hypothesis.

Conversely, if a researcher neglects adequate statistical power, the researcher decreases the chance of finding true effects and fails to reject the null hypothesis.

The process of establishing an adequate level of statistical power (power analysis) within an investigation should be performed in the planning stages of the study before the data are collected (a priori). Despite this general rule, there is evidence that may indicate that researchers are not planning for adequate levels of statistical power. The problem of inadequate statistical power usually comes to light when the researcher does not find a true effect. The researcher then performs the power analysis after the data have been collected (post hoc) and realizes that some aspect of the research design was overlooked which produced low statistical power, e.g., sample size was too small.

The problem of inadequate statistical power in research was first identified in the early 1960s. Cohen's (1962) landmark study of the statistical power of the research published in the Journal of Abnormal and Social Psychology brought the issue to the forefront of research design and evaluation. He explained the status of statistical power analysis at that time.

> The problem of power is occasionally approached indirectly by the concern of sample size to be used in an investigation. Other things equal, power is a monotonic function of sample size, but decisions as to sample size are typically reached by recourse to local tradition, ready availability of data, unaided intuition, usually called "experience," and negotiation (the latter between doctoral candidate and sponsor, or author and editor) and rarely on the basis of a Type II error analysis, which can always be performed prior to the collection of data. These non-rational bases for setting sample size must often result in investigations being undertaken which have little chance of success despite the actual falsity of the null hypothesis, and probably less often in the use of a far larger sample than is necessary. Either of these circumstances is wasteful of research effort (p. 145).

Despite Cohen's concern and warning about inadequate statistical power almost 40 years ago, several researchers still overlook the necessity for adequate statistical power. For example, Davlin, Holcomb, and Guadagnoli (1999) investigated the effect of electromyographic biofeedback training and hip position on the strength ratio between the vastus medialis oblique (VMO) and vastus lateralis (VL) muscles. An imbalance between the VMO and VL is associated with patellofemoral pain syndrome (PFPS), and the authors wanted to determine if the biofeedback training and hip position would positively affect the VMO/VL strength ratio.

The authors used 36 subjects who were randomly assigned to one of three experimental groups. The first group used the biofeedback training with the hip in a neutral position. The second group used the training with the hip in external rotation. The third group used the training with the hip in internal rotation. The treatment protocol spanned five consecutive days and was initiated with a pretest on the first day, proceeded by the biofeedback training on the second, third, and fourth days, and concluded with a posttest on the fifth day.

Davlin, Holcomb, and Guadagnoli (1999) reported a significant difference between the pretest and posttest; therefore, the subjects received a benefit from the biofeedback training. However, the authors also reported nonsignificant results between the three groups (hip positions). The authors explained that the lack of significant differences between the groups was attributable to the small sample size. Through further investigation, the authors explained that they would have needed to double the sample size to demonstrate significant differences among the groups.

In essence the study lacked the statistical power to detect the true difference between the three treatment groups. The problem of small sample size could have been detected with an a priori power analysis, and changes could have been made to the study design to ensure adequate statistical power.

Keppel (1991) stated that it is unclear why this situation exists considering procedures are readily available for researchers to design studies with acceptable statistical power. These procedures would allow researchers to escape the statistical power problems associated with most research.

Other statisticians have expressed their concern for this oversight in more contemporary research. Sedlmeier and Gigerenzer (1989) stated that research was in a "paradoxical state," given that a high priority is placed on significant findings, yet studies are designed and conducted with a small chance of resulting in a significant difference if there was a true effect. In a bold statement, the authors stated, "Researchers paradoxically seem to prefer probable waste of time, money, and energy to the explicit calculation of power." Keppel (1991) stated, "The reality is that most researchers appear to pay little attention to power," and presented the question, "Why should we waste time and resources undertaking a project that has a relatively low probability of detecting treatment effects and producing significant results?" Shavelson (1996) stated that researchers should take a "power trip" and conduct studies designed with as much statistical power as possible. Like the others, Shavelson stated that it is not reasonable to waste time and money on an experiment that has a small probability of detecting a difference when a true difference is present.

Discipline of Athletic Training

Many disciplines of study (e.g., education, psychology, and marketing) have been examined for existing levels of statistical power. This study centers on the discipline of athletic training.

The discipline of athletic training deals with the prevention, evaluation, treatment, and rehabilitation of athletic injuries. The Journal of Athletic Training, published by the National Athletic Trainers' Association

(NATA), is the primary source of research in athletic training. The articles include original research, literature reviews, clinical techniques, case reports, and special communications that encompass the performance domains of the Certified Athletic Trainer (ATC). These domains are: (1) prevention, (2) recognition, evaluation, and assessment, (3) immediate care, (4) treatment, rehabilitation, and reconditioning, (5) organization and administration, and (6) professional development and responsibility (National Athletic Trainers Association Board of Certification, 1999).

In an editorial, Perrin (1997), editor of the journal, explained the roles of a more scholarly journal for the profession of athletic training and made recommendations for the Journal of Athletic Training to become a scholarly journal. He stated that the original research published in a scholarly journal would create a foundation for professional position statements. He explained that a scholarly journal would contribute to the progress of the athletic training profession by establishing a scientific body of knowledge. Further in the editorial, he recommended that improvement was needed in the quality of the writing and research published in the journal, and improvement was needed in the peer review process.

If the Journal of Athletic Training were to assume the roles proposed by Perrin (1997), then statistical power must be considered within the research published in the journal. Position statements cannot be established nor the profession of athletic training advanced if the information in the foundational body of knowledge is not reliable. Statistical power is used by the researcher to detect true effects, but statistical power also deals with the ability to reproduce the researcher's findings when other researchers attempt to replicate the study. The ability to reproduce the results is a reliability issue. The results of a study with low statistical power cannot be reproduced adequately which makes the results unreliable and places an unsupportive "brick" in the foundation of position statements and professional advancement.

Statistical power is a factor in the recommendations presented by Perrin (1997). A priori power analyses would assist researchers in maximizing their research design to find true effects, which, in turn, would improve the quality of the research published. Also, the editors and reviewers of athletic training research can encourage the practice of statistical power analysis by stating that they expect the authors to explain their choices of sample sizes and to report the statistical power and effect sizes in the results.

Purpose

The primary purpose of this study was to conduct a post hoc power analysis within the athletic training research published in the Journal of Athletic Training. A secondary purpose of this study was to identify and evaluate the use of a priori power analyses and the report of observed power, observed effect size, level of significance, and nonsignificant results.

Significance of the Study

The issue of statistical power must be brought to the attention of the producers, sponsoring agencies, and readers of athletic training research. Individuals who conduct research in athletic training must address statistical power in order to detect the true treatment differences they expect to find. Members of sponsoring agencies, such as the review board of the Journal of Athletic Training and the NATA Research and Education Foundation, should consider the statistical power of submitted studies and proposals to determine if the submitted materials warrant publication or financial support. The readers of athletic training research, primarily athletic training practitioners, should understand statistical power in order to evaluate the reliability of research results before incorporating treatments into their respective practices.

Research Questions

Through consideration of the purpose and need for statistical power analysis in research, statisticians' concerns regarding inadequate statistical power, and previous power analysis studies, the following research questions were formulated. In the research studies published in volumes 32, 33, and 34 of the Journal of Athletic Training:

1. What were the types of research published?
2. How many studies contained a report of the use of a power analysis?
3. How many studies contained a reported calculation of observed power?
4. How many studies contained a reported calculation of observed effect size?
5. What was the reported level of significance (alpha) for each of the research studies?

6. How many studies contained a report of nonsignificant results?
7. Of the studies that contained nonsignificant results, how many studies contained a report of insufficient statistical power as a possible reason for the nonsignificant results?
8. How many studies were designated for post hoc power analyses?
9. What were the frequencies and means of the calculated statistical power values for the designated articles?

Limitations of the Study

The following limitations were established due to the scope and design of this study.

1. The sample of athletic training research was limited to the three most recently completed volumes of the *Journal of Athletic Training* (volume 32, 1997; volume 33, 1998; volume 34, 1999) and did not reflect the studies published in previous volumes.
2. The studies designated for post hoc power analyses were limited to those studies whose authors used the statistics for which Cohen's (1988) tables were designed and did not reflect those studies whose authors used within-subjects (repeated measures) designs.

3. The post hoc power calculations were limited to the use of the power tables presented in Cohen's (1988) Statistical Power Analysis for the Behavioral Sciences and did not reflect the use of other tables and computer programs that can be used to calculate statistical power.

Definitions of Terms

For consistency and clarity of meaning throughout this study, the following definitions were used for the listed terms.

Alpha (α) level (level of significance) - the probability that defines how rare or unlikely the sample data must be before the researcher can reject the null hypothesis (Huck, Cormier, & Bounds, 1974); the probability of rejecting a true null hypothesis (Type I error) (Keppel, 1991)

A priori - estimated from available facts without close examination (Webster's School Dictionary, 1980)

Beta (β) - the probability of failing to reject a false null hypothesis (Type II error) (Keppel, 1991)

Description - one of the uses of statistics that describes the sample taken from the population; statistics may include measures of central tendency (mean, median, mode) and variability (standard deviation, variance, range, standard error of the mean) (Thomas & Nelson, 1996)

Effect size - the magnitude or strength of the relationships or treatment differences under investigation (Cohen, 1988)

Extraneous variable (a.k.a., nuisance or confounding variable) - the factor that may explain the effect other than the independent variable under investigation (Shavelson, 1996)

Inference - the ability of the researcher to generalize (or infer) the results from the sample to the general population (Shavelson, 1996)

Observed effect size - the size of the effect that was calculated by the researcher(s) once the data were collected and analyzed (Cohen, 1988)

Observed power - the statistical power of a research study that was calculated by the researcher(s) after the data were collected and analyzed (Cohen, 1988)

Population - the large group of people, designated by the researcher, that is studied and from which the sample is taken (Shavelson, 1996)

Post hoc power analysis - the calculation of statistical power after the data have been collected and analyzed (Cohen, 1988)

Power - the probability of finding true treatment effects which allows the researcher to reject a false null hypothesis (Cohen, 1988)

Power analysis - the a priori process of determining the statistical power of a study for an established level of significance, effect size, and sample size, or the a priori process of determining the required sample size for an established effect size, level of significance, and statistical power (Cohen, 1988)

Sample - the subset of subjects taken from the population (Shavelson, 1996)

Sample size - the number of subjects included in the sample (Shavelson, 1996)

Statistics - an objective method of interpreting a set of observations or scores (Thomas & Nelson, 1996)

Type I error - the rejection of the null hypothesis when in fact the null hypothesis is true (Keppel, 1991)

Type II error - the failure to reject the null hypothesis when in fact the null hypothesis is false (Keppel, 1991)

Chapter 2

Review of Literature

The following subjects are reviewed within this chapter: (1) the uses of statistics, (2) the hypothesis testing procedure, (3) errors in hypothesis testing, (4) definition of statistical power, (5) the relationship of statistical power, α, sample size, and effect size, (6) the methods of increasing statistical power, (7) previous research of post hoc power analysis, (8) implications of the neglect of statistical power in research, (9) reasons for the neglect of statistical power in research, and (10) statistical power analysis in athletic training.

The Uses of Statistics

Thomas and Nelson (1996) explained that statistics is a way of interpreting collected data, and researchers use statistics in three ways. One use of statistics is to describe the characteristics of the sample and data in which the researcher is studying. The most common way of describing these characteristics is with measures of central tendency, i.e., the mean, median, and mode, and variability, i.e., range and standard deviation (p. 92).

Thomas and Nelson (1996) further explained that another use for statistics is to establish and test relationships between sets of data. The relationship is referred to as a correlation, and the most common type of correlation is the Pearson _r_ (simple correlation) (p. 92). The quantitative value of the Pearson _r_ correlation, called a correlation coefficient, is a measure of the degree of relationship between two variables. The two variables can be positively correlated or negatively correlated. Therefore, the values for a correlation coefficient can range from -1.00 to +1.00, where -1.00 is a perfect negative relationship, +1.00 is a perfect positive relationship, and 0.00 is no relationship (p. 116).

Thomas and Nelson (1996) further explained that the third use of statistics is to measure the differences among groups of data. This process involves calculating a value from the data, dependent upon the test used, and comparing the calculated value to a value from a table to determine if the groups are significantly different (p. 92). Commonly used tests that measure the difference among groups include the _t_ test (difference between two groups), the analysis of variance (difference between two or more levels of the independent variable), and the factorial analysis of variance (difference between the levels of two or more independent variables) (p. 140).

Thomas and Nelson (1996) also explained that the latter two uses of statistics are calculated from a comparison of components of the following formula: total variance = true variance + error variance. Researchers use different parts of the formula to measure relationships between variables and test for differences among groups. To calculate a relationship, the researcher uses the proportion of total variance accounted for by the true variance (true variance/total variance) and takes the square root of that value. To calculate a difference among groups, the researcher uses the ratio of true variance to error variance (true variance/error variance) (p. 152).

Often there is a misunderstanding of the uses of statistical techniques. Description and inference, i.e., the ability to generalize the results to the larger population, are sometimes considered the uses of statistics. Shavelson (1996) defined inferential statistics as, "a set of methods used to draw inferences about a large group of people from data available on only a representative subset of the group." Shavelson (1996) defined descriptive statistics as, "a set of concepts and methods used in organizing, summarizing, tabulating, and describing collections of data."

Thomas and Nelson (1996) argued that description and inference are confused as statistical techniques, and the confusion, "is the result of saying that correlations describe relationships and that cause-and-effect is inferred by techniques for testing differences between groups." Actually, the sample of subjects is described by the results of any statistical technique. Inference is based on the sample, procedures, and context and has nothing to do with the statistical technique used. If the sample has been properly selected and represents a larger group, then the results can be inferred to the larger group (p. 94).

The Hypothesis Testing Procedure

Before any data can be collected, ideas, or hypotheses, about the outcome of the experiment must be established. Keppel (1991) and Shavelson (1996) explained that the hypothesis to be tested is the null hypothesis (denoted H_0). The authors explained that with the null hypothesis the researcher predicts there is no effect in the population(s) under investigation. When using statistics to measure the difference between groups, the researcher tests if there is a difference between the means of the groups. When using statistics to test for relationships between variables, the researcher demonstrates whether there is a relationship between the variables under investigation.

Keppel (1991) and Shavelson (1996) explained that an alternative hypothesis (denoted H_1) is also considered in the study. The authors further explained that the alternative hypothesis specifies that an effect does exist in the population(s) under investigation. A researcher may state that the group means are not equal (nondirectional), or the mean of one group will be greater than the other(s)(directional). When studying relationships, a researcher may state that a relationship exists between two variables (nondirectional), or the relationship is positive or negative (directional).

Shavelson (1996) explained that researchers must consider previous studies when choosing one of the forms of the alternative hypothesis. If little previous research is found and/or the topic is relatively new, the researcher may design the study to be more exploratory and use the nondirectional alternative hypothesis. If the researcher is fairly certain that the observed difference will favor one group or the other, the researcher selects the appropriate directional alternative hypothesis (p. 220).

In all cases, the null hypothesis is set up in an attempt to knock it down (reject it). Cohen (1990) stated, "It is called a 'null' hypothesis because the strategy is to nullify it." If this is accomplished, the alternative hypothesis is supported indirectly, and the alternative hypothesis is what led to the study in the first place.

The level of significance is another item that needs to be set before the data are collected. Huck, Cormier, and Bounds (1974) explained that the level of significance, denoted by the Greek letter α (alpha), is a probability (p) that defines how rare or unlikely the sample data must be before the researcher can reject the null hypothesis.

Shavelson (1996) explained that the establishment of a significance level leads to the following decision: The null hypothesis is rejected if the probability of obtaining a sample mean at or beyond a certain value is less than or equal to the specified α. If the probability of the sample mean is greater than the specified α, then the null hypothesis is not rejected.

In summary, the researcher formulates a null hypothesis to be tested. The researcher also establishes a level of significance as a criterion to determine if a true relationship or difference exists. The researcher applies the specified statistical technique to the data and receives an observed value for that technique. The researcher then compares the observed value to a table designed to give critical values for the specific technique within the established significance level. If the observed value exceeds the critical value, the researcher rejects the null hypothesis and accepts the respective alternative hypothesis.

Modern statistical software for computers calculates the actual probability for the specific statistical test value, and referral to a table is not needed. If the probability of the statistical test is less than or equal to the preset level of significance, i.e., $p \le \alpha$, then the null hypothesis can be rejected.

Errors in Hypothesis Testing

Kraemer and Thiemann (1987) explained that errors can be made in the hypothesis testing process even though the researcher may have followed the correct procedures. Using an analogy of a court trial, even though the prosecutor follows the procedure for prosecuting a defendant the jury can make mistakes in deciding whether to return a guilty verdict or uphold the innocence of the defendant.

Sometimes, the jury may find a defendant guilty when, in reality, he/she is truly innocent. Other times, the jury may find a defendant innocent when, in reality, he/she is truly guilty.

Kraemer and Thiemann (1987) further explained that the problem is similar to rejecting or failing to reject the null hypothesis in deciding if a relationship or difference is significant. Sometimes the researcher rejects the null hypothesis when, in reality, the null hypothesis is true and should not be rejected. This error is known as a Type I error. Other times the researcher fails to reject the null hypothesis when, in reality, the null hypothesis is false and should be rejected. This error is known as a Type II error.

Keppel (1991) explained the two types of errors another way. In reality, the null hypothesis is either true or false, and the researcher either rejects or does not reject the null hypothesis. The combination of these decisions results in four possible situations, of which two are the correct decisions, specifically, rejecting the null hypothesis when it is false and not rejecting the null hypothesis when it is true. Table 1, on page 23, summarizes the possible situations and correct decisions.

Table 1

Errors in Null Hypothesis Testing

	Reality	
Decision	H_0 is true	H_0 is false
Reject H_0	Incorrect (Type I error)	Correct
Retain H_0	Correct	Incorrect (Type II error)

There is an inherent problem when researchers try to control for Type I and Type II errors. Huck, Cormier, and Bounds (1974), Keppel (1991), Shavelson (1996), and Thomas and Nelson (1996) explained that an inverse relationship exists when controlling for the two types of errors. Therefore, when the researcher decreases the likelihood of committing a Type I error, the probability of making a Type II error increases, and vice versa.

The researcher has another consideration within the study design as a result of this inherent problem between the two types of error. Shavelson (1996) explained that researchers must decide which of the two types of error is more serious. Type I errors may lead the researcher into a blind alley by relying on an effect that does not exist in the population. Franks and Huck (1986) stated, "In exploratory studies, Type II errors are very costly because they may eliminate factors that need to be included in future studies and overall theory."

Keppel (1991) and Shavelson (1996) explained that researchers should use common sense when considering the risks of committing a Type I or Type II error, and, after careful consideration, the researcher must design the study to control for the specific error deemed more serious. Miller and Knapp (1971) discussed the relative implications of Type I and Type II errors.

> The prevalent preoccupation with the avoidance of Type I error bespeaks a commendable concern for the integrity of knowledge by subjecting H_1 to rigorous tests against H_0. Yet the cause of science may be prejudiced far more gravely in the long run by the erroneous, and perhaps permanent, abandonment of a true H_1. By its very nature the incorrect rejection of H_0 invites ultimate exposure. Type II error, however, is more likely to escape detection (p. 21).

Keppel (1991) and Shavelson (1996) explained that in any study researchers need to use a balanced research design that adequately controls for both types of error. Although the balanced approach in research design is needed and admirable, the balanced approach is easier said than done.

Huck, Cormier, and Bounds (1974), Keppel (1991), Shavelson (1996), Thomas and Nelson (1996) explained that researchers control the probability of making Type I errors by using the alpha (α) level, which is the same alpha used for the level of significance. The α level not only serves to determine significance, but the α also represents the probability of rejecting a true null hypothesis. As α is decreased, e.g., from .10 to .05 to .01, the researcher decreases the probability of committing a Type I error and increases the probability of making a correct decision when the null hypothesis is true.

Huck, Cormier, and Bounds (1974), Keppel (1991), Shavelson (1996), and Thomas and Nelson (1996) also explained that researchers establish a beta (β) level to control for a Type II errors. Restricting the β will decrease the probability of committing a Type II error and increase the probability of making a correct decision when the null hypothesis is false.

Definition of Statistical Power

Cohen (1970) defined statistical power as the probability of rejecting a false null hypothesis, i.e., $1 - \beta$. Cohen (1988) expanded the definition and stated, "The power of a statistical test is the probability that it will yield statistically significant results."

The theoretical definition of statistical power goes beyond the mathematical definition of $1 - \beta$. Keppel (1991) explained, "Power reflects the degree to which we can detect the treatment differences we expect and the chances that others will be able to duplicate our findings when they attempt to repeat our experiments." Therefore, statistical power includes the ability to detect true effects (relationships and differences) that allow the researcher to reject a false null hypothesis. Also, the concept of statistical power includes the ability to reproduce the results of the original study.

For example, if a researcher found that his/her study has a statistical power value of .5, the researcher had a 50% chance of rejecting a false null hypothesis. If the study were repeated 100 times, the researcher would reject a false null hypothesis in 50 studies. The lack of statistical power in this example has a profound effect on the researcher's ability to support his/her hypothesis and the reliability of the researcher's results. Thomas, Lochbaum, Landers, and He (1997) stated, "A real difference is a reliable one, because the null hypothesis is rejected consistently in replications of the research."

Relationship of Statistical Power, α, Sample Size, and Effect Size

Sample size is the number of subjects used for the study. Keppel (1991) explained that researchers have to decide on the proper number of subjects based on the cost and availability of subjects, research design constraints, and the statistical power of the design.

Sawyer and Ball (1981) and Kosciulek & Szymanski (1993) stated that effect size is the degree of group differences or the strength of relationships among variables. Cohen (1988) explained that effect size may also be considered as the degree to which the phenomenon of interest is observable in the population or the degree to which the null hypothesis is false. West (1985) explained that effect size, "refers to the extent to which an alternative hypothesis is true in the population." Katzer and Sodt (1973) stated that the effect size, "tells the researcher how important his results are; it helps answer the 'so what' question." Cohen (1992) and Shavelson (1996) stated that effect size is an index of the discrepancy between the null and alternative hypotheses.

This "index" can be expressed in a quantitative value. Keppel (1991) and Thomas and Nelson (1996) explained that a common index for effect size, also considered treatment magnitude, is the omega squared (ω^2). The ω^2 value varies between 0, when effects are not present, and 1.0 when effects are present. The measure is an indication of the percentage of total variance explained, or accounted for, by the treatment(s) or condition(s) under investigation.

Other common expressions of effect size are formulas presented by Cohen (1988). He developed the effect size formulas as a way to standardize the size of the treatment effects. The formulas represent the influence of the treatment or grouping variable on the dependent variable reported in standard deviation units.

The effect size is important to the interpretation of the results. Ahrens (1971) stated, "The researcher should not be limited to tests of significance. Rather, he should measure the magnitude of the relationships." Cohen (1990) agreed as he stated, "I have learned and taught that the primary product of a research inquiry is one or more measures of effect size, not *p* values." Chase and Tucker (1975) explained that after conducting the study the power-related data should be included in an effort to analyze the results of the study. They stated further,

> The experimenter may or may not choose to compute the observed effect size, but he should provide the necessary information so that his readers may do so. This information enables scholars to determine the amount of variance in the dependent variable which was accounted for by the independent variable, and as such provides valuable data concerning the experimental effort (p. 40).

Katzer and Sodt (1973) stated that it seems obvious that a researcher would want to know these pieces of information. Brown (1989) explained that none of the information requires any additional effort from the researcher, because the information is included in the output of most statistical software packages.

Statistical power, effect size, α, and sample size should be considered when designing a study. Thomas et al. (1997) explained that if three of these factors are known, or can be estimated, the fourth factor can be calculated. Researchers should use statistical power, α, effect size, and sample size, and the relationships between them, to properly design a study. These design parameters should be taken into consideration before the study (a priori).

With statistical power, α, effect size, and sample size in mind, Cohen (1988) described four possible types of power analysis in which one parameter is determined as a function of the other three. A researcher can determine: (1) power as a function of α, effect size, and sample size, (2) sample size as a function effect size, α, and power, (3) effect size as a function of α, sample size, and power, and (4) α as a function of sample size, power, and effect size.

Cohen (1988) explained that analyses 1 and 2 are the most commonly used types. Analysis 1 is used in post hoc power analysis where the researcher used the published α, effect size, and sample size to determine the statistical power of the published study. Analysis 2 is used to predetermine the needed sample size for a study with a predetermined α, effect size, and power level. Cohen (1988) further explained that the use of analysis 3 is less common, but may be quite useful in special circumstances, e.g., comparing research results in literature surveys. Cohen (1988) also explained that the use of analysis 4 is very uncommon, because researchers may not want to use an α level larger than the significance criterion.

Methods of Increasing Statistical Power

When designing a study, researchers must be familiar with the methods that help increase the statistical power of a study. The other components (sample size, effect size, and α) can be manipulated by the researcher to produce a sound research study with adequate statistical power to support the research (alternative) hypothesis.

Cohen (1962, 1988, 1990) stated that sample size is the factor normally used to increase statistical power. If the effect size and α are held constant, a researcher can increase the statistical power of the study by increasing the sample size. However, many deterrents to using larger sample sizes exist, e.g., time, cost and availability of subjects. Therefore, the researcher must use the other components (α and effect size) to boost the statistical power.

The α level can be adjusted according to the needs of the study. Franks and Huck (1986) explained that with the effect size and sample size held constant, researchers can increase statistical power by increasing the α level. Researchers can relax the α level and use less stringent levels, e.g., $\alpha = .10$ or .25, for studies that are exploratory, have limited samples sizes, or for which a Type II error is deemed more costly than a Type I error. Keppel (1991) agreed and stated that this method is a reasonable option if researchers realize that they cannot attain adequate statistical power in the study without increasing the significance level. O'Brien and Israel (1987) stated, "Only after addressing all design options...should the Type I error be considered for manipulation."

Chase and Tucker (1975), West (1985), and Mazen, Graf, Kellogg, and Hemmasi (1987) stated that of the three components of statistical power effect size is perhaps the most difficult to estimate. Ideally, the researcher needs to fully review previous well-conceived studies to ascertain some idea of expected treatment effect. Cohen (1988) stated that when it is not feasible to calculate such an index from previous research, researchers can utilize one of Cohen's (1988) three conventional levels representing small, medium, and large effect sizes present in a population.

Thomas and Nelson (1996) explained that even though it is difficult to estimate an effect size, researchers are able to control the effect size in several ways. Methods of controlling the effect size can be determined by examining the ratio used to calculate differences among groups, i.e., true variance/error variance ratio, and the ratio used to determine relationships, i.e., true variance/total variance. Therefore, by maximizing the true effect (effect size) in the numerator and controlling the error variance in the denominator, the researcher can increase the statistical power within the study.

Thomas and Nelson (1996) explained that there are several methods of increasing effect size that the researcher can use, especially if the researcher is studying the difference among groups. Applying stronger, more concentrated treatments increases the differences among the groups. The researcher can plan a longer treatment time, e.g., 12 weeks instead of eight weeks, so the treatment can show a better effect. The researcher can select subjects in which the effect is more obvious, i.e., comparing novice and expert subjects.

Keppel (1991), Shavelson (1996), and Thomas and Nelson (1996) explained that reducing the error variance in the ratio increases the statistical power. First, the researcher must choose a reliable method of measuring the dependent variable. The instrument used to measure the dependent variable must be consistent, accurate, reproducible, and dependable. Second, the researcher can eliminate the amount of measurement error to increase the statistical power of a study. The researcher must maintain proper calibration of the equipment, incorporate well-trained assistants and technicians, and use special testing rooms that maintain consistent environmental factors, i.e. noise, temperature, and illumination. Third, the researcher can choose a statistical technique that reduces the error variance, e.g., the analysis of covariance (ANCOVA).

Finally, Cohen (1990), Keppel (1991), and O'Brien and Israel (1987) explained that researchers increase statistical power by decreasing the number of treatment conditions considered within the study. Researchers can increase the sample size of the remaining treatment conditions, therefore, increasing the statistical power of the study. Researchers should not consider this maneuver if all of the treatment conditions are essential for the study. However, some treatment conditions can be eliminated if the conditions provide information that is secondary to the main purpose of the study and/or the conditions are too time consuming and costly to measure. The researcher should complete a statistical power analysis, a priori, to decide whether some treatment conditions can be eliminated without jeopardizing the integrity of the study.

Previous Research of Post Hoc Power Analysis

Jacob Cohen initiated the topic of research involving the post hoc power analysis of completed research in 1962. The purpose of his study was three-fold: (1) to bring the issue of statistical power to the attention of researchers and consumers of research, (2) to develop tables and uniform standards which facilitate the use of statistical power analyses for common statistical tests, and (3) to survey the psychological research literature and determine the statistical power of the literature.

Cohen (1962) stated that the main difficulty for psychological researchers in performing a statistical power analysis was the determination of the proper effect size that the researchers wanted to detect, or the effect size that existed in the population. With this in mind, he developed ranges of small, medium, and large effect sizes for frequently used statistical tests. He hoped that the development of uniform standards would encourage the use of statistical power analysis in the design of psychological research, and that the standards would be applicable throughout the diverse content areas of psychological research.

After he determined and gave the rationale for the ranges of effect sizes, Cohen (1962) used the standards to investigate the status of statistical power in published psychological research. He surveyed the articles published in volume 61 (1960) of the Journal of Abnormal and Social Psychology. The mean statistical power of the articles was .18 for small effect sizes, .48 for medium effect sizes, and .83 for large effect sizes. This meant that psychological researchers had an 18% chance of rejecting a false null hypothesis assuming the authors attempted to detect a small effect size, a 48% chance assuming the authors attempted to detect a medium effect size, and an 83% chance assuming the authors attempted to detect a large effect size. Cohen (1962) concluded that the author(s) who published research studies in this particular volume had a poor chance of rejecting a false null hypothesis, unless the author(s) was expecting a large effect size.

With these results, Cohen developed standards for statistical power analysis and compiled the standards into a handbook called Statistical Power Analysis for the Behavioral Sciences, which was published in 1969. The text was later revised in 1977 and in 1988. This text and Cohen's concept of statistical power analysis were the basis for subsequent research in other disciplines.

Brewer (1972) conducted a study of the statistical power present in the American Educational Research Journal. He surveyed the issues from November 1969 through May 1971. Brewer (1972) found that the mean statistical power for the journal articles was .14, .58, and .78 for small, medium, and large effect sizes, respectively.

As a comparison, Brewer (1972) also calculated the statistical power for two other journals. He surveyed articles from the Journal of Research in Science Teaching and The Research Quarterly covering a comparable time span as the American Educational Research Journal. Brewer (1972) reported that the mean statistical power for the

Journal of Research in Science Teaching was .22, .71, and .87 for the assumed small, medium, and large effect sizes, respectively. Brewer (1972) reported that the mean statistical power for The Research Quarterly was .14, .52, and .80 for the assumed small, medium, and large effect sizes, respectively.

Penick and Brewer (1972) surveyed the articles published in the 1969 and 1970 issues of the Journal of Research in Science Teaching. Penick and Brewer (1972) found that the mean statistical power of the surveyed articles was .22, .71, and .87 for the assumed small, medium, and large effect sizes, respectively.

Katzer and Sodt (1973) surveyed articles published in the 1971 and 1972 issues of the Journal of Communication. Katzer and Sodt (1973) found that the mean statistical power of the articles surveyed was .23, .56, and .79 for the assumed small, medium, and large effect sizes, respectively.

Haase (1974) surveyed the articles published in the 1968 to 1971 issues of Counselor Education and Supervision. Haase (1974) found that the mean statistical power of the surveyed articles was .10, .37, and .74 for the assumed small, medium, and large effect sizes, respectively.

Chase and Tucker (1975) surveyed the articles published in the 1971 to 1972 issues of the Journal of Communication. Chase and Tucker (1975) found that the mean statistical power of the surveyed articles was .18, .52, and .79 for the assumed small, medium, and large effect sizes, respectively.

Kroll and Chase (1975) surveyed the articles published in the 1973 to 1974 issues of the Journal of Communication Disorders and the Journal of Speech and Hearing Research. Kroll and Chase (1975) found that the combined mean statistical power of the surveyed articles was .16, .44, and .73 for the assumed small, medium, and large effect sizes, respectively.

Chase and Baran (1976) surveyed the articles published in the 1974 issues of Journalism Quarterly and the Journal of Broadcasting. Chase and Baran (1976) found that the mean statistical power of the surveyed articles was .34, .76, and .91 for the assumed small, medium, and large effect sizes, respectively.

Chase and Chase (1976) surveyed the articles published in the 1974 issues of the Journal of Applied Psychology. Chase and Chase (1976) found that the mean statistical power of the surveyed articles was .25, .67, and .86 for the assumed small, medium, and large effect sizes, respectively.

Schmelkin (1976) surveyed the articles published in the 1972 to 1974 issues of Exceptional Children. Schmelkin (1976) found that the mean statistical power of the surveyed articles was .11, .49, and .82 for the assumed small, medium, and large effect sizes, respectively.

Christensen and Christensen (1977) surveyed the articles published in the 1975 issues of The Research Quarterly. Christensen and Christensen (1977) found that the mean statistical power of the surveyed articles was .18, .39, and .62 for the assumed small, medium, and large effect sizes, respectively.

Sawyer and Ball (1981) surveyed the articles published in the 1979 issues of the Journal of Marketing Research. Sawyer and Ball (1981) found that the mean statistical power of the surveyed articles was .41, .89, and .98 for the assumed small, medium, and large effect sizes, respectively.

Woolley and Dawson (1983) performed a follow-up study to investigate the changes in statistical power that occurred since Penick and Brewer's (1972) study. Using the 1977-1980 issues of the Journal of Research in Science

Teaching, Woolley and Dawson (1983) found that the mean statistical power values for small, medium, and large effect sizes were .29, .63, and .85, respectively, compared to Penick and Brewer's (1972) results of .22, .71, and .87 calculated from the 1969 and 1970 volumes of the same journal.

West (1985) surveyed the articles published in the 1971 to 1982 issues of Adult Education. West (1985) found that the mean statistical power of the surveyed articles was .22, .63, and .85 for the assumed small, medium, and large effect sizes, respectively.

Mazen, Graf, Kellogg, and Hemmasi (1987) surveyed the articles published in the 1984 issues of the Academy of Management Journal and the Journal of Management and in the 1984 Proceedings of the Midwest Division of the Academy of Management. Mazen, Graf, Kellogg, and Hemmasi (1987) found that the mean statistical power of the surveyed articles was .31, .77, and .91 for the assumed small, medium, and large effect sizes, respectively.

Brown (1989) surveyed the articles published in two issues of the following criminology/criminal justice journals: the Journal of Criminal Law and Criminology, Criminology, the Journal of Research in Crime and

Delinquency, Crime and Delinquency, the Journal of Criminal Justice, the Journal of Police science and Administration, Criminal Justice Review, and Criminal Justice and Behavior. Brown (1989) found that the mean statistical power for the surveyed articles was .41, .84, and .96 for the assumed small, medium, and large effect sizes, respectively.

Sedlmeier and Gigerenzer (1989) performed a follow-up study to Cohen's (1962) original study. The Journal of Abnormal and Social Psychology was divided into the Journal of Abnormal Psychology and the Journal of Personality and Social Psychology. Sedlmeier and Gigerenzer (1989) used the 1984 issues of the Journal of Abnormal Psychology to assess the status of statistical power in psychological research 24 years after Cohen's (1962) study. Sedlmeier and Gigerenzer (1989) found that the mean statistical power values for small, medium, and large effect sizes were .21, .50, and .84, respectively. They compared their results to Cohen's (1962) results of .18, .48, and .83 for small, medium, and large effect sizes, respectively. Sedlmeier and Gigerenzer (1989) concluded that authors who published articles in the 1984 volume did not properly address the issue of statistical power, and there was no increase in the statistical power of studies from 1960 to 1984.

McKean (1991) surveyed doctoral dissertation research within the field of educational psychology. McKean (1991) found that the mean statistical power of the surveyed dissertations was .17, .54, and .79 for the assumed small, medium, and large effect sizes, respectively.

Daniel (1993) surveyed the articles published in the 1987 to 1991 issues of the Journal of Research in Music Education. Daniel (1993) found that the mean statistical power of the surveyed articles was .13, .64, and .97 for the assumed small, medium, and large effect sizes, respectively.

Kosciulek and Szymanski (1993) surveyed the articles published in the 1990 to 1991 issues of Rehabilitation Bulletin, the 1990 issues of Rehabilitation Psychology, the 1990 issues of the Journal of Applied Rehabilitation Counseling, the 1990 issues of the Journal of Rehabilitation, and the 1990 issues of Rehabilitation Education. Kosciulek and Szymanski (1993) found that the combined mean statistical power of the surveyed articles was .15, .63, and .90 for the assumed small, medium, and large effect sizes, respectively.

Mone, Mueller, and Mauland (1996) surveyed the articles published in the 1992 to 1994 issues of the following journals: the Academy of Management Journal, Administrative Science Quarterly, the Journal of Applied Psychology, the Journal of Management, Organizational Behavior and Human Decision Processes, Personnel Psychology, and the Strategic Management Journal. Mone, Mueller, and Mauland (1996) found that the combined mean statistical power of the surveyed articles was .27, .74, and .92 for the assumed small, medium, and large effect sizes, respectively.

Clark-Carter (1997) surveyed the articles published in the 1993 and 1994 issues of the British Journal of Psychology. Clark-Carter (1997) found that the mean statistical power of the surveyed articles was .20, .60, and .82 for the assumed small, medium, and large effect sizes, respectively.

Coblick (1998) surveyed the articles published in the 1995 issues of Nursing Research and the Western Journal of Nursing Research. Coblick (1998) found that the mean statistical power of the surveyed articles in Nursing Research was .25, .80, and .94 for the assumed small, medium, and large effect sizes, respectively. Coblick (1998) also found that the mean statistical power for the surveyed articles in the Western Journal of Nursing

Research was .27, .77, and .92 for the assumed small, medium, and large effect sizes, respectively.

Implications of the Neglect of Statistical Power in Research

Chase and Tucker (1975) and Kroll and Chase (1975) explained that the first, and most important, implication of neglecting the issue of statistical power is the commission of a Type II error. Associated with this implication is the likelihood of presenting misleading results in a research study.

Chase and Baran (1976) and Wooley and Dawson (1983) explained that the lack of adequate statistical power in the study may prevent the study from being published. This may be troublesome for the researcher employed by a university whose faculty review board requires publication for tenure and promotion.

A consequence of disregarding a proper power analysis is overpowered research studies. Chase and Tucker (1976) and Mone, Mueller, and Mauland (1996) explained that studies with sample sizes that are too large lead researchers into finding trivial or irrelevant results. Keppel (1991) explained that the researcher should use the minimally required sample size corresponding with acceptable statistical power to keep the possibility of loss of life to a minimum if animal subjects are utilized.

Reasons for the Neglect of Statistical Power in Research

Several reasons have been proposed to explain the neglect of statistical power in research. Kraemer and Thiemann (1987) proposed an educational problem. They stated the mathmatical complexity of statistical power is difficult to teach to researchers, and researchers are regularly taught to deal with significance level but rarely with statistical power. Kraemer and Thiemann (1987) stated that the consideration of statistical power is an important part of designing a research study, and this process requires formal training and experience in a particular field of research. These skills in power analysis research design are not easily taught in coursework (p. 17).

Sedlmeier and Gigerenzer (1989) stated the problem of the neglect of statistical power in research was created by the differences in theories for testing the null hypothesis. R. A. Fisher's theory was based on the strict rejection of the null hypothesis, usually at the .05 level, with no consideration of alternative hypotheses. Jerzy Neyman and Egon Pearson's theory of null hypothesis testing included the statement of alternative hypotheses and the consideration of statistical power. Sedlmeier and Gigerenzer (1989) stated that the differences in theory were a source of bitter controversies between the Fisher camp and the Neyman-Pearson camp.

Sedlmeier and Gigerenzer (1989) explained that there is a historical reason why researchers neglected statistical power in the first place. Fisher's statistical theory was developed about 1935, and the Neyman-Pearson statistical theory was developed and presented during World War II. Therefore, researchers have been familiar with Fisher's theory for a longer period of time than the Neyman-Pearson theory.

Sedlmeier and Gigerenzer (1989) also explained that subsequent blending of the two theories over the years into a "hybrid" theory of null hypothesis testing has continued the neglect of statistical power in research studies. The authors stated that the blending of the two theories into a single theory has produced confusion about the meaning of basic concepts. The continued neglect of statistical power is a direct result of this problem. Sedlmeier and Gigerenzer (1989) concluded, "It is important to understand the unresolved issue of statistical power in psychological studies against the background of this unresolved debate in statistics rather than as an isolated issue."

Statistical power may be neglected in research due to researchers' misunderstanding of the factors that determine sample size, i.e., statistical power, α, and effect size. In a telephone survey of 28 authors, Sawyer and Ball (1981) reported responses to an open-ended question about what factors determine sample size. Budget limitations, availability of subjects, or convenience were cited most often as the factors that determine sample size. Five of the 28 authors surveyed cited a statistical rule of thumb or an expected effect size as a criterion for determining sample size.

Another possible reason for the neglect of statistical power in research may lie in the editors and reviewers of the research. Mone, Mueller, and Mauland (1996) surveyed authors of management research and concluded that the respondents believed that journal editors and reviewers generally do not call for greater usage of power analysis. Mone, Mueller, and Mauland (1996) commented that the editors and reviewers did not intend to convey the idea that power analysis was not important, but their silence concerning the use of power analysis seemed to have great influence on the respondents' beliefs and behaviors.

Another possible reason for neglecting power analysis is that researchers simply do not want to take the time to conduct the power analysis. Mone, Mueller, and Mauland (1996) came to this conclusion after reporting that, of the 169 respondents, 108 (63.9%) of the authors surveyed reported never using power analysis. Also, of the 108 respondents that did not use power analysis, 67.4% disagreed strongly or moderately with the statement, "I am not familiar with power analysis." Mone, Mueller, and Mauland (1996) stated that they believe that researchers are aware of power analysis, but choose not to use it.

Statistical Power Analysis in Athletic Training

Researchers in the field of athletic training have been calling for more research. Osternig (1988) stated that the research component of athletic training has progressed much more slowly than the practice and education components of athletic training. He further stated, "Ongoing research is essential if the field is to be credible within paramedical disciplines and must play a far greater role if athletic training is to continue to be recognized as a true allied health profession."

Osternig (1988) cited the current practices of athletic trainers as examples to explain the need for research. He stated, "Much of what is taught as athletic training does not have a sound base of principles and practice which have been subjected to and validated by scientific scrutiny." Osternig (1988) explained that varied responses and opinions would be submitted if athletic trainers were asked the efficacy of a particular modality for a specific physical condition. He further explained that most of the responses would be based on anecdotes and testimonials instead of established research.

Osternig (1988) seemed optimistic about the future of research in athletic training. He stated, "Although such research has been slow to evolve, there can be no question that research activity by athletic trainers is increasing rapidly." However, Osternig (1988) gave a warning if this increase in athletic training research does not continue. He stated, "Failure to meet this need for systematic inquiry into our educational and clinical practice and thereby more adequately meet society's needs, can only lead to the eventual demise of the profession."

Knight (1990) agreed with Osternig's (1988) call for more research. Knight (1990) stated,

> "One of the most important tasks of the NATA is to delineate, develop, and expand the athletic training body of knowledge; or stated another way, to increase our understanding of the skills and techniques used by the everyday athletic trainer in the trenches to care for athletes and their injuries." (p. 8)

Further in the article, Knight (1990) stated, "We must continue to upgrade the quality and quantity of the research articles that we publish. Research is essential to our profession - we'll die without it."

Perrin (1997) addressed the role of a more scholarly journal for the profession of athletic training. He stated, "The original research published in a scholarly journal also forms the basis for professional position statements. A professional position statement establishes standards for clinical practice that can also be used for educational and legal purposes." Perrin (1997) further stated, "For the educator/researcher, a scholarly journal serves as an outlet for publication of findings related to athletic training, basic science, and clinical research." Perrin (1997) also stated, "These publications help to move the profession forward by establishing a scientific body of knowledge."

The concept of adequate statistical power has not been addressed in the athletic training literature, yet the statements made by Osternig (1988), Knight (1990), and Perrin (1997) are dependent upon the existence of adequate statistical power in athletic training research. Statistical power involves the proper research design to detect true effects in the population under investigation. Statistical power also involves the ability to reproduce the results if the study is replicated. With this in mind, the profession of athletic training cannot progress if the body of knowledge that supports the profession is poorly designed and unreliable. Position statements cannot be used for educational and legal purposes if the research upon which the position statement is based is poorly designed and unreliable.

Statistical power and its components (sample size, effect size, and α) have been addressed in the physical education literature. Teraslinna (1967) criticized the criteria used for accepting research articles in The Research Quarterly. She stated,

> How many research workers in our field ever speculate about the set α level before the experiment? How many care to know what would be practically significant?

> Who would ever take the pains to find out the contamination of the error term by uncontrolled or unknown variance, which would certainly affect the β level, hence influencing our decision in setting the proper α level for the test? (p. 155)

Uncontrolled and unknown variance, α, and β influence statistical power. It seemed from Teraslinna's (1967) comments that researchers who published articles in The Research Quarterly were not considering these items, and, consequently, the researchers were not properly considering the statistical power of their respective studies.

Ahrens (1971) expressed the need for physical education researchers to incorporate the effect size of the study results instead of relying on the significance test to determine differences or relationships. Ahrens (1971) stated,

> The researcher should not be limited to tests of significance. Rather, he should measure the magnitudes of the relationships. Depending upon the problem and researcher's aim, he may measure the magnitudes of the relationships in one of several ways: the difference of two means, the proportion of

the total variance "explained," the coefficient of regression and of correlation, or other measures of association.

Baumgartner (1974) criticized the sampling techniques used by the authors of the articles published in The Research Quarterly. He stated,

> Physiological, kinesiological, and psychological studies with faulty and outdated methodological techniques are rejected for publication in The Research Quarterly and I feel that this should also be true with faulty and outdated statistical techniques. We are still accepting the same faulty sampling practice accepted 10 years ago even though correct procedures are now well explained in many statistics books on the market (p. 215).

Faulty sampling is especially evident in terms of sample sizes of the published research, i.e., the sample are frequently too small. A consequence of small sample sizes is low statistical power.

Christensen and Christensen (1977) performed a post hoc power analysis on volume 46 of The Research Quarterly. They found that the surveyed articles had a mean statistical power of .17 assuming the authors attempted to detect a small effect size, .38 assuming the authors attempted to detect a medium effect size, and .62 assuming the authors attempted to detect a large effect size. Christensen and Christensen (1977) concluded, "This finding seems counterproductive in light of the usual research goal of finding relationships between variables."

Taking these statements and criticisms into account, this study of the post hoc power analysis of the athletic training research was conducted for the following reasons.

1. Statistical testing plays an important role in athletic training research. When properly applied, the statistical techniques contribute to the development of athletic training. Proper application is needed in order to avoid wasted effort, incorrect conclusions, and flawed theory. To assure proper application, periodic assessment of the state of statistical usage in the field of athletic training is beneficial.
2. Statistical power is probably inadequate in athletic training, because it is evident in a variety of disciplines, especially the related discipline of physical education.
3. The discipline of athletic training is comparatively new to other disciplines, e.g., physical education, or other

allied health professions, e.g., nursing. Athletic training researchers could avoid the "pitfall" of inadequate statistical power if the problem is identified and resolved in the early stages of research development.

4. The problem of inadequate statistical power in athletic training research has not been addressed in the athletic training literature.

Summary

The use of statistics, the hypothesis testing procedure, and the possible errors that can be committed need to be understood in order to understand the definition and purpose of statistical power. The relationship between statistical power, effect size, sample, and the α level is used to increase the statistical power of a research study. Increased statistical power in research studies is needed because authors of previous studies reported low statistical power, especially when small and medium effect sizes were assumed. Finally, the implications and reasons for the neglect of statistical power analysis must be overcome to produce adequately designed research studies that detect true treatment effects for the researchers and produce reliable results for the practitioner.

Chapter 3

Methods

The contents of this chapter describe the methods used to conduct this study. The subjects of discussion are (1) the identification of the population and sample, (2) the instruments and techniques for measurement, (3) the procedure for data collection, and (4) the data analysis.

The Identification of the Population and Sample

The articles published in volumes 1-34 of the *Journal of Athletic Training* constituted the population of athletic training research. The journal is peer-reviewed and published quarterly by the National Athletic Trainers' Association (NATA). The journal contains a variety of articles that include original research, literature reviews, case reports, clinical techniques, and special communications. The articles relate to one or more of the domains of athletic training: (1) prevention, (2) recognition, evaluation, and assessment, (3) immediate care, (4) treatment, rehabilitation, and reconditioning, (5) organization and administration, and (6) professional development and responsibility (National Athletic Trainers' Association Board of Certification, 1999).

The three most recently completed volumes (volume 32, 1997; volume 33, 1998; and volume 34, 1999) of the Journal of Athletic Training constituted the sample that was surveyed to answer the following research questions:

1. What were the types of research published?
2. How many studies contained a report of the use of a power analysis?
3. How many studies contained a reported calculation of observed power?
4. How many studies contained a reported calculation of observed effect size?
5. What was the reported level of significance (alpha) for each of the research studies?
6. How many studies contained a report of nonsignificant results?
7. Of the studies that contained nonsignificant results, how many studies contained a report of insufficient power as a possible reason for the nonsignificant results?
8. How many studies were designated for post hoc power analyses?
9. What were the frequencies and means of the calculated power values for the designated articles?

Three volumes were chosen due to the quarterly publishing of the journal. Three volumes of the Journal of Athletic Training equate to a journal that is published monthly (12 issues) and include an ample pool of articles to evaluate for power analysis.

The Instruments and Techniques for Measurement

Cohen's (1988) power tables were used to calculate the statistical power of the designated articles. The tables were developed to help researchers design their studies and conduct post hoc power analyses on completed studies. Cohen (1988) designed the power tables for the following types of statistical tests: (1) the *t* test for means, (2) Pearson correlation coefficient *r*, (3) differences between correlation coefficients, (4) tests of proportion and the sign test, (5) differences between proportions, (6) chi-square (χ^2) tests, (7) analysis of variance (ANOVA) and covariance (ANCOVA), (8) multiple regression and correlation analysis, and (9) multivariate analysis.

Cohen (1988) proposed small, medium, and large effect size values as conventions or operational definitions for each of the statistical techniques. Cohen (1992) explained that a medium effect size would be visible to the naked eye of a careful observer. He further explained that he set the small effect size to be noticeably smaller than the medium effect size but not so small that the effect would be trivial. He set the large effect size the same distance above the medium effect size as the small effect size was below it. Cohen (1992) stated that these values were at first proposed subjectively, but through many adjustments over the years the values have come into general use by researchers.

The values proposed by Cohen (1988) for the small, medium, and large effect sizes were used for this study. The values for the small, medium, and large effect sizes are (1) .20, .50, and .80 for the t test for means; (2) .10, .30, and .50 for Pearson correlation coefficient r; (3) .10, .30, and .50 for the difference between correlation coefficients; (4) .05, .15, and .25 for the tests of proportion and the sign test; (5) .20, .50, and .80 for the difference between proportions; (6) .10, .30, and .50 for the chi-square (χ^2) tests; (7) .10, .25, and .40 for the analysis of variance (ANOVA) and covariance (ANCOVA); (8) .02, .15, and .35 for the multiple regression and correlation analysis; and (9) .02, .15, and .35 for the multivariate analysis; respectively.

The Procedure for Data Collection

The articles published in volumes 1-34 of the Journal of Athletic Training constituted the population of athletic training research. The three most recently completed volumes (volume 32, 1997; volume 33, 1998; volume 34, 1999) constituted the sample that was surveyed to answer the research questions. Volume 32 had 45 articles published, volume 33 had 51 articles published, and volume 34 had 45 articles published for a total of 141 articles in the sample. Each of the 141 articles was identified as either original research, a literature review, a case report, a clinical technique, or a special communication in order to answer the first research question.

The articles identified as original research were further categorized into types of research in order to answer the first research question. Each of the 89 original research articles was identified as either experimental, quasi-experimental, or survey/questionnaire research.

The original research articles were utilized to answer research questions 2-7. Only the articles whose authors used statistical techniques to test significant differences and relationships were incorporated into this study. The authors of 12 of the survey/questionnaire studies did not use any significance testing. These 12 articles were excluded, and 77 articles were used for research questions 2-7. Each of the remaining 77 articles was surveyed for the following information: (1) a report of an a priori power analysis, (2) a report of observed power, (3) a report of observed effect size, (4) the reported α level, (5) a report of nonsignificant results, and (6) a report that low statistical power was a possible reason for the nonsignificant results.

The studies designated for post hoc power analyses were limited to those studies whose authors used statistics for which Cohen's (1988) power tables were designed. Of the 77 articles surveyed, 36 were designated for post hoc power analysis in order to answer research questions 8 and 9.

Cohen (1962) established conditions within his original study that have been used by previous researchers and were included in this study. One condition of the study implemented the use of the .05 level of significance (α) for all the eligible studies, whether or not a different α was indicated. The second condition implemented the use of the non-directional version of the null hypothesis for all the eligible studies, whether or not a different version was indicated. The third condition stipulated the calculation of statistical power for the specific test that supported the primary purpose presented by the author(s).

The statistical power of the 36 studies was calculated with power tables provided in Cohen's (1988) handbook. The tables are structured according to the specific statistical test (t test, r, ANOVA, etc.), the α level, and the tailing (one-tailed or two-tailed) of the t test and r. The statistical technique that supported the primary purpose was identified for each article. The specific table corresponding to the statistical technique was then utilized. Using the sample size reported in the article and the proposed effect size values from Cohen (1988) the statistical power for each of the 36 articles was calculated for the assumed small, medium, and large effect sizes.

Data Analysis

The data were analyzed according to the individual research question. For research question 1, the categories and types of research were analyzed as frequencies and percentages of the total sample. The results of research question 2-8 were analyzed as frequencies and percentages of the total number of articles surveyed. The results of research question 9 were analyzed as frequencies within ranges of statistical power, cumulative percentages, mean statistical power and standard deviation for each of the small, medium, and large effect sizes. The frequencies, percentage distributions, mean statistical power and standard deviation were computed using SPSS-PC software (version 10.0, Chicago, IL).

Chapter 4

Results

The results are reported according to the individual research question. The data were analyzed according to the method explained in the Data Analysis section in Chapter 3.

Research Question No. 1

The first research question asked, "What were the types of research published?" There were 45 articles published in volume 32, 51 articles in volume 33, and 45 articles in volume 34 of the Journal of Athletic Training totaling 141 articles in the sample. Of these articles, 89 (63%) were original research, 22 (16%) were case reports, 14 (10%) were literature reviews, nine (6%) were special communications, and seven (5%) were clinical techniques. These results are summarized and presented in Table 2 on the following page.

The articles categorized as original research were surveyed to answer research questions 2-8. Of the 89 original research articles, 41 (46%) were considered experimental research, 24 (27%) were considered quasi-experimental, and 24 (27%) were considered surveys/ questionnaires. These resulted are summarized and presented in Table 3 on the following page.

Table 2

Types of Research within the Sample

Type	Number	Percentage
Original Research	89	63%
Case Reports	22	16%
Literature Reviews	14	10%
Special Communications	9	6%
Clinical Techniques	7	5%
Total	141	100%

Table 3

Original Research Articles Divided into Types of Design

Type of Design	Number	Percentage
Experimental	41	46%
Quasi-experimental	24	27%
Survey/Questionnaire	24	27%
Total	89	100%

Twelve of the 24 articles that were categorized as surveys/ questionnaires contained statistics used to describe the respondents and did not contain statistics used to measure significant relationships or differences among groups. These articles were excluded from the survey. Therefore, the total number of articles used to answer research questions 2-8 was 77.

Research Question No. 2

The second research question asked, "How many studies contained a report of the use of a power analysis?" Of the 77 articles surveyed, two (3%) articles contained a report of the use of a power analysis before the data were collected.

Research Question No. 3

The third research question asked, "How many studies contained a calculation of observed power?" Of the 77 articles surveyed, four (5%) articles contained a calculation of observed power.

Research Question No. 4

The fourth research question asked, "How many studies contained a calculation of observed effect size?" Of the 77 articles surveyed, none (0%) of the articles contained a calculation of observed effect size.

Research Question No. 5

The fifth research question asked, "What was the reported level of significance (α) for each of the research studies?" Of the 77 articles surveyed, 58 (75%) articles contained a report of $\alpha = .05$. The remaining 19 (25%) articles contained no report of the α level.

Research Question No. 6

The sixth research question asked, "How many studies contained a report of nonsignificant results?" Of the 77 article surveyed, 47 (61%) articles contained a report of nonsignificant results.

Research Question No. 7

The seventh research question asked, "Of the studies that contained nonsignificant results, how many studies contained a report of insufficient power as a possible reason for the nonsignificant results?" Of the 47 articles that contained a report of nonsignificant results, nine (19%) articles contained a report of insufficient power as a possible reason for the nonsignificant results.

Research Question No. 8

The eighth research question asked, "How many studies were designated for post hoc power analysis?" Of the 77 articles surveyed, 36 (47%) articles were designated for post hoc power analysis.

Research Question No. 9

The ninth research question asked, "What were the frequencies and means of the calculated power values for the designated articles?" The sample size and α were used to calculate the power of each statistical test assuming the author(s) attempted to detect a small, medium, and large effect size. Assuming the authors of the designated articles attempted to detect a small effect size, the mean power was .18 (SD = .16). The frequencies, mean, and standard deviation for the small effect size are summarized and presented in Table 4 on the following page. Assuming the authors of the designated articles attempted to detect a medium effect size, the mean power was .53 (SD = .33). The frequencies, mean, and standard deviation for the medium effect size are summarized and presented in Table 5 on page 71. Assuming the authors of the designated articles attempted to detect a large effect size, the mean power was .75 (SD = .27). The frequencies, mean, and standard deviation for the large effect size are summarized and presented in Table 6 on page 72.

Table 4

Frequency and Cumulative Percentage of the Power to Detect Small Effect Sizes (n = 36)

Power	Frequency	Cumulative Percentage
.99 +		
.90-.99		
.80-.89		
.70-.79		
.60-.69	1	100
.50-.59	2	97
.40-.49	2	92
.30-.39	3	86
.20-.29	2	78
.10-.19	8	72
.01-.09	18	50
Mean	.18	
SD	.16	

Table 5

Frequency and Cumulative Percentage of the Power to Detect Medium Effect Sizes (n = 36)

Power	Frequency	Cumulative Percentage
.99 +	3	100
.90-.99	5	92
.80-.89	1	77
.70-.79	4	75
.60-.69	2	64
.50-.59	1	58
.40-.49	4	56
.30-.39	2	44
.20-.29	7	39
.10-.19	6	19
.01-.09	1	3
Mean	.53	
SD	.33	

Table 6

Frequency and Cumulative Percentage of the Power to Detect Large Effect Sizes (n = 36)

Power	Frequency	Cumulative Percentage
.99 +	8	100
.90-.99	9	77
.80-.89	3	53
.70-.79	1	44
.60-.69	3	42
.50-.59	4	33
.40-.49	1	22
.30-.39	5	19
.20-.29	1	6
.10-.19	1	3
.01-.09		
Mean	.75	
SD	.27	

Chapter 5

Discussion, Conclusions, and Recommendations

The producers, sponsors, and readers of athletic training research must address the issue of statistical power. Therefore, the primary purpose of this study was to conduct a post hoc power analysis to evaluate the current status of statistical power within the athletic training research published in the Journal of Athletic Training. The secondary purpose of this study was to identify and evaluate the use of a priori analyses and the report of observed power, observed effect size, level of significance, and nonsignificant results. The following discussion, conclusions, and recommendations were produced from the consideration of the data collection and results.

Discussion

The data collection process was difficult due to the absence of any consistent method of presenting statistical results. This difficulty was experienced by others (Katzer & Sodt, 1973; Chase & Tucker, 1975; Chase & Baran, 1976; Woolley & Dawson, 1983; West, 1985; Brown, 1989).

On several occasions pertinent data were missing. The authors of one article (Sharpe, Knapik, & Jones, 1997) reported a significant χ^2, but did not give the calculated value for the statistic. The authors of another article (Smith, Szczerba, Arnold, Martin, & Perrin, 1997) reported no probability value for the *t* test. The authors of another article (Myrer, Measom, Durrant, & Fellingham, 1997) reported significant correlations but did not report the values of the correlations. The authors of another article (Boyle, Sitler, Rogers, Duffy, & Kimura, 1997) restructured the groups they were comparing and did not give a report of the new sample size for the next comparison of group differences.

Another problem was identified from the method used to report the probabilities of the statistical tests. A typical *t* test in the *Journal of Athletic Training* was reported as, "($t(28) = 9.63$, $p < .001$)" This type of report made it difficult to approximate the statistical power of the test or determine the α used unless $\alpha = p$. If the α level is larger than .001, e.g., .05, then it should be stated. The statement "$p < .001$" is misleading considering $p < .05$ means the same thing as far as rejection at $\alpha = .05$ is concerned. The reporting of the probability at some smaller value, e.g., $p < .001$, may cause the reader to believe that a large effect size has resulted from the statistical analysis. Neither α nor p provide any evidence about the effect size. The researcher must give a value of observed effect size in order for the reader to interpret the strength of the relationship or the magnitude of the difference among groups.

The second research question asked, "How many studies contained a report of the use of a power analysis?" The surveyed athletic training research lacked the use of a priori statistical power analyses. Two articles (Bernier, Perrin, & Rijke, 1997; Kaminski, Perrin, & Gansneder, 1999) contained a report of using an a priori power analysis. Consequently, these two reports are difficult to interpret, because the authors did not define the parameters from which the statistical power was calculated. A thorough a priori statistical power analysis can assist the athletic training researcher in a valid rejection of the null hypothesis and allow the researcher to be comfortable with the results.

The third research question asked, "How many studies contained a calculation of observed power?" The authors of the surveyed articles omitted the report of the observed power. Four articles (Bonacci & Higbie, 1997; Stay, Ricard, Draper, Schulthies, & Durrant, 1998; Kaminski, Perrin, & Gansneder, 1999; Davlin, Holcomb, & Guadagnoli, 1999) contained a report of observed power. The observed power is a valuable tool in the interpretation of the results.

The fourth research question asked, "How many studies contained a calculation of the observed effect size?" None of the 77 articles surveyed contained a report of observed effect size. The effect size is important to the interpretation of the results. The effect size value allows the researcher to observe and interpret the magnitude of the significant difference or relationship.

The fifth research question asked, "What was the reported level of significance (α) for each of the research studies?" Fifty-eight of the 77 articles contained a reported α level, and all of the reported α levels were .05. This may lead a reader to think that the .05 α level is an established research policy for the Journal of Athletic Training, when in reality the policy does not exist. The value of the level of significance lies on a continuum of values that must be considered by the researcher before conducting the study. The α level can be adjusted according to the needs of the study, e.g., studies that are exploratory, have limited size, and/or for which the relative cost of Type II errors is judged to be greater than that of a Type I error.

Another problem was identified after studying the 19 articles that contained no report of an α level. This method of not establishing an α level can be an appropriate technique. The researcher reports the probability of the test equal to a specified value, and the reader is allowed to judge if the effect is significant. This was not the case with the 19 articles that contained no report of an α level. The authors of these articles continued to use the term "significant" as if the α level had been established. This use of the of the term "significant" could be confusing to readers, because the readers do not know if the researcher meant that the results were statistically different or that the results were important or critical.

The sixth research question asked, "How many studies contained a report of nonsignificant results?" Forty-seven of the 77 (61%) articles surveyed contained a report of nonsignificant results. This may dispel Cohen's (1962) belief that reviewers and editors are biased toward publishing only "successful" studies, i.e., studies with significant results. However, when a researcher reports nonsignificant results, no conclusions can be made from the study. The researcher is no further ahead in explaining the treatment or phenomenon under investigation than when the study was initiated.

The seventh research question asked, "Of the studies that contained nonsignificant results, how many studies contained a report of insufficient power as a possible reason for the nonsignificant results?" Out of the 47 articles that contained a report of nonsignificant results, nine (19%) articles contained an explanation that insufficient statistical power was a possible cause for the nonsignificant results. Whenever a researcher reports nonsignificant results, inadequate statistical power should always be considered as a possible explanation. A researcher cannot accept the null hypothesis and conclude that no difference or no relationship exists, but that the experiment is not sufficiently sensitive (powerful) enough to detect the effect if it did exist.

The eighth research question asked, "How many studies were designated for post hoc power analysis?" Thirty-six articles were designated for a post hoc power analysis. The authors of the remaining 41 articles used research designs and statistics that were not included in Cohen's (1988) power tables.

The ninth research question asked, "What were the frequencies and means of the calculated power values for the designated articles?" The statistical power for each of the 36 articles was calculated assuming the authors attempted to detect a small effect size, and the mean statistical power was a dismal 18%. The statistical power for each of the 36 articles was calculated assuming the authors attempted to detect a medium effect size, and the mean statistical power was 53%, which is slightly better than a coin flip. The statistical power for each of the 36 articles was calculated assuming the authors attempted to detect a large effect size, and the mean statistical power was 75%, which did not compare well with the other disciplines discussed in Chapter 2.

Conclusions

Considering the results of this study, the following conclusions were made.

1. Researchers who published articles in volumes 32, 33, and 34 of the Journal of Athletic Training did not give sufficient attention to the concepts of statistical power and effect size.

2. Researchers who published articles in volumes 32, 33, and 34 of the Journal of Athletic Training had a poor chance of finding a true effect and rejecting the null hypothesis unless the effect the authors sought was large.
3. Researchers who published articles in volumes 32, 33, and 34 of the Journal of Athletic Training lacked uniformity in the reporting of results, and often times lacked pertinent information resulting from the statistical technique utilized to test the researchers' hypotheses.

Recommendations

Based on the conclusions of this study, the following recommendations were made.

1. Statistical power should be addressed in the planning and analysis stages of the research study. Researchers who submit articles for the Journal of Athletic Training should perform a priori power analyses and include the following information: the α level, desired statistical power, the minimum effect size expected, and the sample size needed. After the data are collected and statistically analyzed, the observed power and effect size of the study should be reported to help interpret the results.

2. Researchers should investigate the reasons for the neglect of statistical power in the research published in the <u>Journal of Athletic Training</u>. If future researchers report a misunderstanding of statistical power among athletic training researchers, then special instruction of statistical power analysis should be included in existing workshops that instruct individuals who wish to contribute to the <u>Journal of Athletic Training</u>.
3. Researchers who submit articles for the <u>Journal of Athletic Training</u> should have clear, uniform results statements, which include the sample size actually used (and any reasons for subject mortality), the value of the observed statistic, the p value for that statistic, and the observed effect size for all significant findings.
4. Editors and reviewers for the <u>Journal of Athletic Training</u> should assist in changing submission policies by expecting authors to explain their use of sample sizes, statistical power, effect sizes, and α levels, and expecting the authors to report the observed power and effect size in the results.

References

Ahrens, S. J. (1971). Statistical tests of significance: Truth, paradox, or folly. The Research Quarterly, 42(4), 436-440.

Armstrong, R. L. (1987). The use of statistical significance: A commentary related to Franks and Huck. Research Quarterly for Exercise and Sport, 58(1), 81-82.

Baumgartner, T. A. (1974). Remarks concerning sampling used in the Research Quarterly. The Research Quarterly, 45(2), 215-216.

Bernier, J. N., Perrin, D. H., & Rijke, A. (1997). Effect of unilateral functional instability of the ankle on postural sway and inversion and eversion strength. Journal of Athletic Training, 32(3), 226-232.

Bonacci, J. A. & Higbie, E. J. (1997). Effects of microcurrent treatment on perceived pain and muscle strength following eccentric exercise. Journal of Athletic Training, 32(2), 119-123.

Boyle, M., Sitler, M., Rogers, K., Duffy, A. J., & Kimura, I. (1997). Knowledge and attitudes of certified athletic trainers in Pennsylvania toward HIV/AIDS and treating HIV-positive athletes. Journal of Athletic Training, 32(1), 40-44.

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

Brown, S. E. (1989). Statistical power and criminal justice research. Journal of Criminal Justice, 17, 115-122.

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

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

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

Christensen, J. E. & Christensen, C. E. (1977). Statistical power analysis of health, physical education, and recreation research. The Research Quarterly, 48(1), 204-208.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the British Journal of Psychology. British Journal of Psychology, 88, 71-83.

Coblick, G. E. (1998). Statistical power analysis of nursing research (Doctoral dissertation, Auburn University, 1998). Dissertation Abstracts International, 59(01), 288B.

Cohen, J. (1962). The statistical power of abnormal-social psychological research. Journal of Abnormal and Social Psychology, 65(3), 145-153.

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

Cohen, J. (1973). Statistical power analysis and research results. American Educational Research Journal, 10(3), 225-230.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155-159.

Daniel, T. D. (1993). A statistical power analysis of the quantitative techniques used in the Journal of Research in Music Education: 1987-1991 (Doctoral dissertation, Auburn University, 1993). Dissertation Abstracts International, 54(05), 1717A.

Davlin, C. D., Holcomb, W. R., & Guadagnoli, M. A. (1999). The effect of hip position and electromyographic biofeedback training on the vastus medialis oblique:vastus lateralis ratio. Journal of Athletic Training, 34(4), 342-346.

Franks, B. D. & Huck, S. W. (1986). Why does everyone use the .05 significance level. Research Quarterly for Exercise and Sport, 57(3), 245-249.

Franks, D. F. & Huck, S. W. (1987). Response to Armstrong, Brewer, and O'Brien and Israel. Research Quarterly for Exercise and Sport, 58(1), 87-89.

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

Huck, S. W., Cormier, W. H., & Bounds, W. G. (1974). Reading statistics and research. New York: HarperCollins.

Journal of Athletic Training. (1997), 32.

Journal of Athletic Training. (1998), 33.

Journal of Athletic Training. (1999), 34.

Kaminski, T. W., Perrin, D. H., & Gansneder, B. M. (1999). Eversion strength analysis of uninjured and functionally unstable ankles. Journal of Athletic Training, 34(3), 239-245.

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

Keppel, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Knight, K. L. (1990). Expanding our body of knowledge. Athletic Training, 25(1), 8.

Kosciulek, J. F. & Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. Rehabilitation Counseling Bulletin, 36(4), 212-219.

Kraemer, H. C. & Thiemann, S. (1987). How many subjects: Statistical power analysis in research. Newbury Park, CA: Sage Publications.

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

Mazen, A. M., Graf, L. A., Kellogg, C. E., & Hemmasi, M. (1987). Statistical power in contemporary management research. Academy of Management Journal, 30(2), 369-380.

McKean, K. E. (1991). Statistical power analysis of doctoral dissertation research in educational psychology (Doctoral dissertation, Oklahoma State University, 1991). Dissertation Abstracts International, 51(10), 3367A.

Miller, T. K. & Knapp, T. R. (1971). The importance of statistical power in educational research. Occasional Paper 13, Bloomington, IN: Phi Delta Kappa.

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychological and management research. Personnel Psychology, 49, 103-120.

Myrer, J. W., Measom, G., Durrant, E., & Fellingham, G. W. (1997). Cold- and hot-pack contrast therapy: Subcutaneous and intramuscular temperature changes. Journal of Athletic Training, 32(3), 238-241.

National Athletic Trainers Association Board of Certification. (1999). Role Delineation Study (4th ed.). Philadelphia: F. A. Davis.

O'Brien, K. F. & Israel, R. G. (1987). Response to: Why does everyone use the .05 level of significance. Research Quarterly for Exercise and Sport, 58(1), 85-86.

Osternig, L. R. (1988). Research in athletic training: The missing ingredient. Athletic Training, 23(3), 223-225.

Penick, J. E. & Brewer, J. K. (1972). The power of statistical tests in science teaching research. Journal of Research in Science Teaching, 9(4), 377-381.

Perrin, D. H. (1997). Toward a more scholarly journal. Journal of Athletic Training, 32(1), 13-14.

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

Schmelkin, L. (1976, April). Statistical power analysis of research in Exceptional Children. Paper presented at the annual meeting of the Council for Exceptional Children, Chicago, IL.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105(2), 309-316.

Sharpe, S. R., Knapik, J., & Jones, B. (1997). Ankle braces effectively reduce recurrence of ankle sprains in female soccer players. Journal of Athletic Training, 32(1), 21-24.

Shavelson, R. J. (1996). Statistical reasoning for thee behavioral sciences (3rd ed.). Boston: Allyn and Bacon.

Smith, J., Szczerba, J. E., Arnold, B. L., Martin, D. E., & Perrin, D. H. (1997). Role of hyperpronation as a possible risk factor for anterior cruciate ligament injuries. Journal of Athletic Training, 32(1), 25-28.

Stay, J. C., Ricard, M. D., Draper, D. O., Schulthies, S., S., & Durrant, E. (1998). Pulsed ultrasound fails to diminish delayed-onset muscle soreness symptoms. Journal of Athletic Training, 33(4), 341-346.

Teraslinna, P. (1967). On publishable research articles. The Research Quarterly, 38(1), 154-156.

Thomas, J. R., Lochbaum, M. R., Landers, D. M., & He, C. (1997). Planing significant and meaningful research in exercise science: Estimating sample size. Research Quarterly for Exercise and Sport, 68(1), 33-42.

Thomas, J. R. & Nelson, J. K. (1996). Research methods in physical activity (3rd ed.). Champaign: Human Kinetics.

Webster's school dictionary. (1980). Springfield, MA: Merriam-Webster.

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

Woolley, T. W. & Dawson, G. O. (1983). A follow-up analysis of the statistical tests used in the Journal of Research in Science Teaching. Journal of Research in Science Teaching, 20(7), 673-681.