The Importance of Drawing Meaningful Conclusions from Data: A Review of the Literature with Meta-Analytic Inquiry

David A. Walker

Using correct statistical concepts is an important component when conducting quantitative research. Ideas such as power, effect size, and confidence intervals need to be addressed appropriately every time a research study is initiated. The intent of this review of the literature is to reacquaint faculty, practitioners, and graduate students with scholarly information pertaining to these important concepts to facilitate improved implementation of quantitative research designs. Practical cases are interwoven within the review to furnish examples of concept importance, and a meta-analysis of concept usage found in articles published in the NASPA Journal is provided as a measure for implications.

Introduction

Often in published research articles, we find that some of the statistical techniques used to measure data and procure inferences are not understood appropriately or are applied improperly, which renders conclusions and suppositions questionable if not ineffectual (Gall,

David A. Walker is an assistant professor of educational research and assessment at Northern Illinois University.
Borg, & Gall, 1996; Hall, Ward, & Comer, 1988; O’Hear & MacDonald, 1995; Onwuegbuzie & Daniel, 2000; Thompson, 2002a, 2002b). Past studies in the field of education indicate that the role and value assigned to the concept of significance testing is often distorted, if not entirely misunderstood, when reporting the prominence of study conclusions such as the effect of a particular treatment under examination (Carver, 1993; Falk & Greenbaum, 1995; Oakes, 1986; Shaver, 1993).

The statistical ideas or postulates covered in this review are not new to research within the field of education. This review is not an exhaustive examination of the areas of discussion, but is meant to serve as a primer. Some of the ensuing information can be found in various statistical methods texts often used in doctoral programs. Reiteration of information emanating from these important texts is necessary to provide a worthy review and also lay the framework for the meta-analytic study presented later.

The purpose of this review of the literature is to reacquaint faculty, practitioners, and graduate students with information pertaining to the salience of related statistical concepts: power, effect size, and confidence intervals. Traditionally, these statistical techniques have been significant components of quality educational research that appear to have been overlooked, or misinterpreted, in the recent past by doctoral programs and researchers (Aiken et al., 1990; Anderson, Burnham, & Thompson, 2000; Meehl, 1990; Pedhazur & Schmelkin, 1991; Posavac, 2002; Thompson, 2002a).

Specifically, in the field of student affairs, Brown (1991) and Malaney (2002) echo this assertion. Malaney finds that “Unfortunately, too many ‘researchers’ today in higher education and student affairs appear to be ill informed methodologically . . . Traditional methodological training is severely needed right now . . .” (pp. 139–140). Thus, it is the aim of this review to inform readers about these techniques and provide examples of concept importance, so that future inferences derived from quantitative research are grounded in sound statistics.
Significance Testing

Typically, significance testing is regarded as a series of prescribed exercises used to test hypotheses, where statistical significance is warranted by $p$ value $< .05$ (i.e., a probability value that forms the boundary between rejecting or not rejecting the null hypothesis, and also addresses an underlying question of whether the observed effect of a study is due to chance). As an example, we are taught that the null hypothesis ($H_0$), or a presumption of no difference, is true. We run a test of significance to nullify this hypothesis (Fisher, 1951) and determine, “Given that $H_0$ is true, what is the probability of these (or more extreme) data?” (Cohen, 1994, p. 998). Thus, if the $p$ value for a particular statistic of study is found to be $< .05$, we decide that the null hypothesis is most likely false and we apply the alternative hypothesis ($H_1$). If the $p$ value is $< .01$, we tend to surmise that the results of the study are exceptionally meaningful, which is not the definition of statistically significant. Ultimately, this line of reasoning reduces significance testing to a twofold choice where we either reject or do not reject the null hypothesis and where the results are deemed important or not meaningful (Falk & Greenbaum, 1995; Kirk, 1996). However, what we should be trying to answer with a significance test is the following question: Is the hypothesis of study appropriate given a particular set of data?

The Null Hypothesis

For years, researchers have addressed the issue of misinterpretation of null hypothesis testing (Cohen, 1994; Kirk, 1996; Meehl, 1967; Oakes, 1986; Rozeboom, 1960; Stewart, 2000; Thompson, 1994). Cohen (1994) found that researchers often apply inattentively and inappropriately nil null hypothesis testing, where the population parameter (i.e., a quantitative measure of the population of interest) is always equivalent to the value of zero. This prevalent phenomenon causes the ubiquitous effect that the nil null has become the default and exclusive hypothesis implemented in many research endeavors, which often contain very few prior studies that support the claim. In fact, Cohen (1994) argued that the overall concept of the null hypothesis might be a false premise because it may not be a precisely true assumption in the population regardless. Hubbard and Ryan (2000) stated “rejection of the null hypothesis is erroneously believed to yield
the probability of the null or research hypotheses being true as well as the probability that a research outcome will replicate” (p. 672). Daniel (1998) added that significance tests did not tell us a great deal about how much, or even if, a sample was concordant with the population of interest.

**Type I Error**

Often, statistical significance is a product of sample size, with large sample sizes producing significant results and the rejection of the null hypothesis (Hays, 1981; Thompson, 1989). To determine statistical significance, the probability of a Type I error is commonly established as alpha (a) at the .05 level. A Type I error tells us that the null hypothesis was rejected incorrectly, whereas a Type II error indicates a failure to reject the null hypothesis when it is false.

Type I errors are often viewed as the *faux pas* to avoid. Many Type I errors are committed by repeated application of the same statistical test, such as numerous t tests, which increases the chance of a Type I error. Beasley and Mulvenon (2001) give an example of this situation by demonstrating how the use of a two-way analysis of variance (ANOVA) model, which has three independent effects and therefore three tests, can increase the possibility of at least one Type I error occurring out of the three tests performed. They find that in this situation, to compensate for the multiple tests and to make it more difficult for any one test to be statistically significant, the alpha level should be readjusted from .05 to .017. This is accomplished through the use of the Bonferroni adjustment method where the alpha level is divided by the number of tests performed (i.e., .05 / 3).

**Type II Error**

Power is a provisional probability that is conditioned on the null hypothesis being false in the population. Thus, the power of a test is the probability of correctly rejecting the null hypothesis when it is false or Power = 1 - β, where β is the probability of a Type II error. The acceptable level of power is often considered to be at .80 or higher, which means that there is an 80% probability of achieving statistically significant results (Cohen, 1988). The power of a statistical test can be enhanced by increasing the sample size (Levin, 1998) or by
employing one-tailed instead of two-tailed tests of significance (note: in one-tailed tests, only one tail of the distribution is used for null hypothesis testing in terms of region of rejection. In two-tailed, both tails of a distribution are used in testing a null hypothesis).

Example of Type I and Type II Errors
Suppose that you are the chair of the student disciplinary committee on your campus. A student who has been before your committee in the recent past, due to numerous conduct infractions, is facing expulsion from your institution because of violating his or her “last chance” conditions of conduct. You review the pertinent data of the case and determine the \( H_0 \)—the student is innocent of the current charges—and the \( H_1 \)—the student is guilty of the current charges. You want to make sure that the data used to determine this student’s fate has minimal error. That is, if your committee commits a Type I error, it will find guilty an innocent student, thus resulting in his or her expulsion from the institution. However, if the committee commits a Type II error, a culpable student will go undisciplined and remain at the institution.

Effect Size
Effect sizes show the extent or strength of a relationship. An examination of effect sizes allows researchers to evaluate the importance of the result and not just the probability of the result (Kirk, 1996; Shaver, 1985). Stewart (2000) notes that, “Effect size is generally reported as some proportion of the total variance accounted for by a given effect” (p. 687).

The American Psychological Association (APA) has stated, “Always provide some effect size estimate when reporting a \( p \)-value” (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 599). However, research studies (Henson & Smith, 2000; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000) found that mere urging by the APA has not produced the desired result of journal editors demanding that effect sizes accompany quantitative studies.

According to Levin (1993), significance testing and effect sizes should be understood as analogous ideas. McLean and Ernest (1998) stress
that regardless of statistical significance, or lack thereof, effect sizes should always be reported. In essence, an effect size can be thought of as a standardized difference. When reported with significance testing, effect size can provide information pertaining to the extent of the difference between the null hypothesis and the alternative hypothesis. Thus, the larger the effect size, the greater the power of a test of statistical significance.

\textit{d} Group Effect Sizes
Effect sizes fall into two categories: \textit{d} and \textit{r}. Both groups have many effect size measures, all of which will not be discussed. The \textit{d} group encompasses measures of effect size in terms of mean difference and standardized mean difference (i.e., data in a common metric for comparison purposes). With caution, Cohen (1988) defined the values of effect sizes for this group as small = .20, medium = .50, and large = .80. However, it is at the discretion of the researcher to note the context in which small, medium, and large effects are being defined when using \textit{d}- and \textit{r}-related indices.

\textit{r} Group Effect Sizes
The \textit{r} group can be considered as based on the correlation between treatment and result (Levin, 1994). It is used with bivariate correlations, such as the Pearson Product Moment correlation coefficient (\textit{r}), which serves as an appropriate effect size estimate when both variables are interval (i.e., data measured on ordered categories that have equal distances between each other). When both variables are dichotomous (i.e., data measured where the variable has two possible classifications), the phi coefficient (\textit{\phi}) is appropriate. When one variable is interval and the other variable is dichotomous, the point biserial correlation coefficient (\textit{r}_{\text{pbs}}) is a useful effect size index. Finally, when both variables are ordinal or ranked (i.e., data measured with ordered classifications), the Spearman’s Rank-Order correlation (\textit{r}_{s}) is apropos (Gall et al., 1996).

For the chi-square statistic, as a measure of association between two variables, a regularly used effect size is based on the coefficient of contingency (\textit{C}) for nominal variables (i.e., data measured without ordered classifications) (Sprinthall, 2000). As a caveat with the use of
C, it has been noted that its highest value cannot attain 1.00, as is common with other effect sizes, which makes concordance with akin effect sizes arduous. In fact, C has a maximum approaching 1.0 only for large tables. In tables smaller than 5 x 5, C may underestimate the level of association (Cohen, 1988; Ferguson, 1966). As an alternative to C, Sakoda's Adjusted C (C*) may be used, which varies from 0 to 1 regardless of table size, or Cramér's V. For chi-square-related effect sizes, Cohen recommended that .10, .30, .50 represent small, medium, and large effects.

When conducting regression analysis, the coefficient of determination ($R^2$) serves as an effect size measure by providing the percentage of variance in the dependent variable explained by the linear combination of the independent variables. However, there is a caveat with the use of $R^2$. It has been found to overestimate model effect sizes, which means that it tends to miscalculate, in a bias upward, the population value (Agresti & Finlay, 1997; Pedhazur, 1997). Due to amending for overestimation, researchers prefer the adjusted $R^2$ as a better method than $R^2$ in estimating effect sizes. This adaptation produces an adjusted effect size corrected for shrinkage, which is the propensity for correlations to decrease when a regression equation is replicated in another research study (Gall et al., 1996; Synder & Lawson, 1993).

In terms of ANOVA and Analysis of Covariance (ANCOVA), if we find that a result is statistically significant at the .05 level, for example, the effect size eta-squared ($\eta^2$) will inform us percentage-wise regarding the amount of real difference present in the sample between the null hypothesis and the alternative hypothesis. In general, effect sizes derived from ANOVA methods with values of .02, .15, and .35 are considered to represent small, medium, and large effects (Cohen, 1988).

Confidence Intervals

Confidence intervals can be calculated for any statistic, including effect sizes, and should be included in all research findings. Confidence intervals show the size and direction of an observed effect, the extent of error in a measure, and if the effect is large enough to be deemed useful in the research (Kirk, 2001). Further, confidence inter-
vals are comparable across samples (Schmidt, 1996) and afford continuous data about the accuracy of scores derived from numerous implementations of an instrument. Being interval estimates, they provide a range of probable values for parameter estimates and should be used as assessments of effect size (Oakes, 1986). As the APA (2001) has indicated, “Because confidence intervals combine information on location and precision and can often be directly used to infer significance levels . . . The use of confidence intervals is therefore strongly recommended” (p. 22).

A Confidence Interval Example

Usually, researchers establish confidence intervals as having either a probability of .99, .95, or .90. For instance, this means that if numerous samples were taken and an estimated mean (i.e., a point estimate) and corresponding 95% confidence interval were constructed for each sample, then 95% of these confidence intervals would contain the true population parameter (Huck & Cormier, 1996).

It should be noted that due to sampling error, called the standard error, the population parameter may be smaller or larger than the given sample point estimate. The width of a confidence interval is influenced by the standard error and the sample size. For a normalized sample of 100 students, a 95% confidence interval, with 5% chance of error, will have a smaller width than a 99% interval. The 95% interval will give us a claim to “tighter knowledge;” that is, “We must pay some chance of error [e.g., 5% versus 1%] to extract knowledge or belief from data” (Tukey, 1991, p. 101).

An Empirical Example of Overall Concept(s) Importance

The following example can be conducted on most major statistical software programs. The analysis being conducted is a test of two independent sample proportions. You are the director of health services at a small liberal arts college with an enrollment of 600 students (i.e., 150 students per class). Three years previously, you noted a trend in student alcohol consumption data that indicated by sophomore year, a large percentage of both male and female students appeared to be
consuming five or more drinks per occasion. Your professional training suggested that many factors play a role in an increase in alcohol consumption among students within this group, but you ascertained that the most prevalent determinant was a lack of education about the issue. Due to this information, you implemented an alcohol awareness education program on the campus to assist with ameliorating this problem. Three years have passed since the implementation of your program, and you would like to know if a decreased proportion of sophomores at your college consume five or more drinks per occasion. The $H_0$ is: The two proportions of sophomore groups are the same. The $H_1$ is: The two proportions of sophomore groups are not the same. A 2 x 2 table will help us determine if there is a difference between these two independent groups of sophomores.

**Table 1**

Health Services Example

<table>
<thead>
<tr>
<th></th>
<th>Consumed 5 or more drinks per occasion</th>
<th>Consumed less than 5 drinks per occasion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sophomore group 1</td>
<td>33 (44%)</td>
<td>42 (56%)</td>
</tr>
<tr>
<td>Sophomore group 2</td>
<td>25 (33%)</td>
<td>50 (67%)</td>
</tr>
</tbody>
</table>

Upon initial review of Table 1, there is an 11% difference between the two normally distributed proportions (i.e., following within reason the distribution of the normal curve). The effect size of use in this instance is Cohen’s $h$, which examines the difference between population proportions. Our $h = .23$ and represents a small effect size. Our 95% confidence interval for the estimated difference between proportion 1 and proportion 2 = (−.0483, .2616). This very large interval tells us that there is a lot of error in this measure and, because it subsumes 0, statistical significance is not present. Finally, the $p$ value = .177. Our preliminary conclusion is that there is not a statistically significant difference in proportions, which means that alcohol consumption of 5 or more drinks per occasion from group 1 to group 2 was essentially the same. However, we forgot to run a power analysis on the data to determine confidently if the hypothesis of study is appropriate given a particular set of data.
From the aforementioned data, should we conclude that groups 1 and 2 had equal proportions? The data indicated that there was a difference of 11%, nevertheless. We need to understand the power of this test and our ability to make a Type II error or to accept $H_0$ when $H_1$ is actually true. Power = .80 is desired. Assuming that both sample sizes are equal (e.g., $n_1 = 75$ and $n_2 = 75$ in this instance), we have power = .28. This means that if the population proportions really differ by 11%, our chance of detecting this 11% difference is only 28% or an effect will be discovered 28 out of 100 times. Conversely, we have a 72% chance of making a Type II error. Thus, if power is low, there may be little chance of demonstrating an effect. Essentially, our ability to accept a large risk of error is indefensible and we cannot begin to ascertain a difference in population proportions with ample high probability.

Our definitive conclusion is that there really was not much of a discernable difference in a decreased proportion of sophomores who consumed five or more drinks per occasion from group 1 to group 2. We failed to reject the $H_0$, which is bolstered by the finding that our test has very low power at 28%, a small effect size at .23, and a large confidence interval with substantial error that does not appear to provide a range of values that have a considerable probability of including the true parameter of interest.

Notice that we have separated the aforementioned methodology from the usual, single criterion of reject/do not reject, which affords limited understanding of the results. Instead, we have examined not only the $p$ value, but also effect size, confidence intervals, and power to provide the reader with more information that will be beneficial in deciding if the data support the hypothesis of study to answer the previously-posed question: Is the hypothesis of study appropriate given a particular set of data?

A Meta-Analysis of NASPA Articles: Fall 1996 to Fall 2002

To determine if all or some of these concepts were being employed by authors of published research articles in the field of student affairs, a meta-analysis, using NASPA data via its Web site, was conducted on
articles published from fall 1996 to fall 2002. This 6-year time span covered by the meta-analysis coincides with the APA’s call for authors to report most of the same statistical concepts discussed in this article (Wilkinson & The APA Task Force on Statistical Inference, 1999). Further, the meta-analysis, as well as the empirical example provided previously, is intended to serve as indicants of some of the specific aspects affiliated with power, effect size, and confidence intervals that should be included in quantitative research.

Excluding book reviews, there were 153 articles published in the NASPA Journal during this time period. Of those 153 articles, only empirically based, quantitative articles were used for this analysis (e.g., articles that presented tests such as chi-square, t tests, ANOVA, regression, or exemplified some form of quantitative data arrived at via inferential techniques). Literature reviews, theoretical pieces, or position papers were not used in the analysis. Thus, 50 articles exhibited singular or multiple use of the concepts discussed earlier (Table 2). The key identifiers for the analysis came from the literature and have all been discussed previously in this article. Specifically, this meta-analysis reviewed the following three areas: (1) power, (2) effect sizes, (3) and confidence intervals. Table 2 presents the results of the analysis.

As Table 2 indicates, the two most prevalent effect sizes reported in NASPA Journal articles were \( r \) and \( R^2 \), which were mostly connected with factor analyses or regression models. Unfortunately, effect sizes affiliated with data derived from t tests and ANOVAs, which were used the most frequently, were only mentioned twice (i.e., both \( d \) and \( \eta^2 \) were reported once). Further, many authors applied the chi-square test, but no effect sizes were reported with its usage. In the area of power, Type II errors were not discussed, but some authors did talk about controlling for Type I errors via a correction method or increasing sample size. Finally, confidence intervals were only reported twice and a small amount of authors, although not reporting actual confidence intervals, noted the standard errors affiliated with their tests.

As a review of the literature indicated previously, and the current meta-analytic results reiterate these important concepts are being underreported, if at all, which presents the potential for yielding deficient research designs and unsound inferences. Certainly, further stud-
Table 2

<table>
<thead>
<tr>
<th>ES</th>
<th>d</th>
<th>r</th>
<th>pphi</th>
<th>rpho</th>
<th>C</th>
<th>C*</th>
<th>v</th>
<th>R²</th>
<th>Adj R²</th>
<th>εta²</th>
<th>General</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Power</td>
<td>Type I</td>
<td>Type II</td>
<td>Correction</td>
<td>Sample Size</td>
<td>Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td>CI</td>
<td>Error</td>
<td>Only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:
1. For ES, "Other" included Kendall’s tau-b and the odds ratio.
2. For ES, "General" is related to a basic description of effect size, but with no numerical indication.
3. For Power, the alpha correction methods used were Bonferroni and Siddak.
4. For Power, "Sample Size" is related to a description of how a small sample size influences power.
ies of this nature should be conducted on other leading student affairs journals to determine if a trend exists within the field. The publication of results from future meta-analyses could inform student affairs of the effectiveness and completeness of the research being published in the field. Moreover, understanding how effect sizes, power analyses, and confidence intervals could be, and are, used may assist other researchers with interpretation of their data.

Conclusion

This article has added to the important topics addressed recently in a 2002 special issue of the NASPA Journal pertaining to scholarship in student affairs. Most of us who have studied quantitative research designs and protocols in master’s or doctoral programs understand that it is imperative to return to the scholarly literature regularly to learn about new research methods or reacquaint ourselves with proven concepts. Thus, upon review, it is apparent from past and recent literature that statistical clarity remains an important area within research that should be nearly perfect. Concepts such as power, effect size, and confidence intervals need to be addressed correctly every time a research study is initiated.

The following is a brief list of publications from the field of student affairs that embody some, if not all, of the concepts discussed in this article. These publications, along with the study’s meta-analytic results, provide directives of what to look for when evaluating an article, in terms of power, effect size, and confidence intervals, regardless of the type of data used in a study. Furthermore, the subsequent articles are pertinent because they give the reader an indication of how to implement these ideas into a research design, how to describe and discuss their prevalence in terms of research implications, and, overall, demonstrate why these ideas matter. For articles that incorporate the concept of confidence intervals, see Guth, Lopez, & Fisher, 2002; Walker, 2003; and Zheng, Whalen, Ciccone, & Shelley, 2001. For an article that uses power, see Schuh & Shelley, 2001. Finally, for articles that describe effect sizes, see Mohr & Sedlacek, 2000; Pascarella, Flowers, & Whitt, 2001; Reason, Walker, & Robinson, 2002; Taylor & Miller, 2002; VanZile-Tamsen, 2001; and Walker, 2001.
These articles epitomize the essence of this review. To make reliable, justifiable decisions when conducting quantitative research entails applying and describing the effects associated with power, effect size, and confidence intervals correctly to a study. These concepts can assist researchers in determining if the hypothesis of study was appropriate given the set of data under review. As the review of the literature points out, as well as the meta-analysis and empirical example, a lack of attention to reporting these statistical concepts can lead to false impressions concerning the data and hypothesis of study, which may lead to inappropriate conclusions.

References


Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? Research in Schools, 5, 43–53.


