10-30-2003

Effect Sizes, Significance Tests, and Confidence Intervals: Assessing the Influence and Impact of Research Reporting Protocol and Practice

Melinda Rae Hess
University of South Florida

Effect Sizes, Significance Tests, and Confidence Intervals:

Assessing the Influence and Impact of Research Reporting Protocol and Practice

by

Melinda Rae Hess

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Measurement and Research
College of Education
University of South Florida

Major Professor: Jeffrey D. Kromrey, Ph.D
Kathryn M. Borman, Ph.D.
John M. Ferron, Ph.D.
Cynthia G. Parshall, Ph.D.

Date of Approval:

October 30, 2003

Keywords: Research Practices, Practical Significance, Statistical Significance,
Educational Research, Confidence Bands

Dedication

It is absolutely inconceivable that I could have hoped to achieve this goal without the support and focus of my parents, Judy and Dean Hess. Regardless of the situations I have found myself in, past and present, they have always encouraged me to pursue my goals, even when things looked bleak. I often thought of quitting, slowing, or delaying this dream due to full time work, financial concerns, classwork, etc., yet their encouragement, support and love provided me with the strength and will power to continue to pursue this goal, regardless of the 'obstacles' present, perceived or otherwise.

I also wish to dedicate this to the memory of my grandfather, Howard Erslev. I have been fortunate to have family on all sides of my family tree who were, and are, good people with strong values. However, my grandfather Erslev has always had a special place in my heart and memory as an individual who knew the value of people as well as hard work. He treated all people well, regardless of race, gender or background during a time and place of great prejudice. The values he instilled and exhibited will never be forgotten and, I believe, were key in shaping my values, even before I knew what 'values' were.

So, in every way possible, thanks Mom and Dad. I dedicate this work to you, my brother Mike, nephew Michael Jr. and my grandparents. There is no way I could have made it through this without you.

Acknowledgements

I am indebted to the entire Measurement and Research Department, for providing the support necessary to successfully complete this program. I was very fortunate to have four extremely knowledgeable and accessible researchers willing to be on my committee. I will always look back on the mentorship of Dr. Jeff Kromrey, Dr. John Ferron, Dr. Cynthia Parshall, and Dr. Kathryn Borman with extreme graditude. Jeff has been a constant inspiration and example of the consummate researcher who, since early in my program, helped me identify my research interests through provision of various research opportunities. John has been a wonderful mentor whose sense of humor provided an additional dimension to this experience, while ensuring I stayed focused on the goal. Cynthia has not only been a wonderful supporter in my research endeavors, but also provided invaluable assistance in obtaining an internship with the Educational Testing Service. Kathryn, who so willing gave of herself, her time and her expertise, has provided me with an appreciation of, the broader implications of methods research.

Of course, without the support and help of my fellow doctoral students, this would have not been the fantastic experience it was. A special thank you to Tom Lang III, Gianna Rendina-Gobioff, Kris Hogarty, Peggy Jones and Freda Watson for making this such an enjoyable experience. It was great to have colleagues with which to laugh, commiserate, and yes, even get some productive work accomplished at times. Thanks so much for everything.

Table of Contents

List of Tables iii

List of Figures iv

Abstract v

Chapter One: Introduction 1
- Statement of Problem 2
- Purpose of the Study 4
- Research Questions 6
- Study Significance 6
- Limitations 7
- Definition of Terms 9

Chapter Two: Review of the Literature 13
- Reporting Research 13
- Disciplinary Norms 16
- Effect Sizes 18
- Statistical vs. Practical Significance 19
- Point Estimates vs. Confidence Intervals 22
- Examples 29
- Summary 32

Chapter Three: Method 34
- Study Type and Description 35
  - Meta-Analysis 35
  - Methodological Research Review 37
- Sample 38
  - Selection of Journals 39
  - Selection of Published Studies 44
- Computations 47
- Confidence Intervals 48
- Data Analysis 49
- Reporting Results and Conclusions 51
- Reliability of Interpretative Results 53
- Recommendations for Reporting Research Results 53

Chapter Four: Results 54
Characteristics of Selected Studies 57
Statistical Significance vs. Practical Significance 68
Potential Impact on Results and Conclusions 73
Examples 75
Summary 79
Point Estimates vs. Confidence Intervals 80
Potential Impact on Results and Conclusions 82
Examples 83
Summary 88

Chapter Five: Conclusions 90
Purpose of Research 90
Overview of Method 91
Impact of Findings 91
Statistical Significance vs. Practical Signficance 92
Point Estimates vs. Confidence Intervals 94
Reporting Results 95
Relevant Issues 97
Future Research 99
Summary 101

References 103

Appendices 116
Appendix A: Coding Sheet for Studies 117
Appendix B: Coding Sheet for Reviewers 119
Appendix C: SAS Code 149
Appendix D: Summary of 42 Analyses 185
Appendix E: Internal Review Board Exemption 212

About the Author End Page

List of Tables

Table 1	Profile of Journals	42
Table 2	Citation Scores and Rankings Compared to all Social Science Journals	43
Table 3	Journal Ranks Relative to Subject Specific Journals	44
Table 4	Types of Analyses Included in Number of Articles	46
Table 5	Effect Sizes and Associated Interpretation	50
Table 6	Types of Analyses Reviewed by Article Number and Journal	59
Table 7	Numbers of Analyses Reporting Statistical Significance Relative to Computed Effect Size	69
Table 8	Number and Percent of Analyses or Sets of Analyses that Warrant Different Degrees of Change When Effect Size is Considered in Addition to Results of Statistical Significance	75

List of Figures

Figure 1.	An illustration of various confidence bandwidths.	25
Figure 2.	Point estimates (Cohen's f^2) of the impact of gender on Mathematics Attitude and Identity.	30
Figure 3.	Point estimates (Cohen's f^2) and confidence intervals on the impact of gender on Mathematics Attitude and Identity at a Type I error rate of .05.	30
Figure 4.	Point estimates (Cohen's d) and confidence intervals on the impact of treatment intensity on gains in students' instructional reading levels at a Type I error rate of .05.	32
Figure 5	Distribution of effect sizes and 95% confidence intervals for all t-tests analyses pooled across journals as effect size increases.	60
Figure 6.	Distribution of effect sizes and 95% confidence intervals for all ANOVA analyses pooled across journals as effect size increases.	61
Figure 7.	Distribution of effect sizes and 95% confidence intervals for all Regression analyses pooled across journals as effect size increases.	62
Figure 8	Distribution of effect sizes and 95% confidence intervals for all t-test analyses as effect size increases by journal type	63
Figure 9	Distribution of effect sizes and 95% confidence intervals for all ANOVA analyses pooled across journals as effect size increases.	63
Figure 10	Distribution of effect sizes and 95% confidence intervals for all Regression analyses as effect size increases by journal.	64
Figure 11	Distribution of effect sizes and 90% confidence intervals for all ANOVA analyses pooled across journals as effect size increases.	65
Figure 12	Distribution of effect sizes and 95% confidence intervals for all ANOVA analyses pooled across journals as effect size increases.	65

Figure 13	Distribution of effect sizes and 99% confidence intervals for all ANOVA analyses pooled across journals as effect size increases.	66
Figure 14	Bandwidth of Cohen's f pooled across journals as total sample size increases for Type I error rates of .01, .05, and .10.	67
Figure 15	Bandwidth of Cohen's f pooled across journals as the ratio of total sample size/number of groups increases for Type I error rates of .01, .05, and .10.	67
Figure 16	Effect sizes of statistically significant findings at an alpha of .05, by journal.	71
Figure 17	Effect sizes of statistically significant findings pooled across journals at an alpha of .05, by analysis type.	71
Figure 18	Effect sizes of non-statistically significant findings pooled across journals at an alpha of .05, by analysis type.	73
Figure 19.	Percent of effect sizes of 95% confidence band endpoints pooled across journals found in statistically significant analyses.	81

Effect Sizes, significance tests, and confidence intervals:

Assessing the influence and impact of research reporting protocol and practice

Melinda Rae Hess

ABSTRACT

This study addresses research reporting practices and protocols by bridging the gap from the theoretical and conceptual debates typically found in the literature with more realistic applications using data from published research. Specifically, the practice of using findings of statistical analysis as the primary, and often only, basis for results and conclusions of research is investigated through computing effect size and confidence intervals and considering how their use might impact the strength of inferences and conclusions reported.

Using a sample of published manuscripts from three peer-reviewed journals, central quantitative findings were expressed as dichotomous hypothesis test results, point estimates of effect sizes and confidence intervals. Studies using three different types of statistical analyses were considered for inclusion: t-tests, regression, and Analysis of Variance (ANOVA). The differences in the substantive interpretations of results from these accomplished and published studies were then examined as a function of these different analytical approaches. Both quantitative and qualitative techniques were used to

examine the findings. General descriptive statistical techniques were employed to capture the magnitude of studies and analyses that might have different interpretations if alternative methods of reporting findings were used in addition to traditional tests of statistical significance. Qualitative methods were then used to gain a sense of the impact on the wording used in the research conclusions of these other forms of reporting findings. It was discovered that tests of non-significant results were more prone to need evidence of effect size than those of significant results. Regardless of tests of significance, the addition of information from confidence intervals tended to heavily impact the findings resulting from significance tests.

The results were interpreted in terms of improving the reporting practices in applied research. Issues that were noted in this study relevant to the primary focus are discussed in general with implications for future research. Recommendations are made regarding editorial and publishing practices, both for primary researchers and editors.

Chapter One

Introduction

The ever-increasing attention and concern about effective educational practices as well as the focus on accountability among educators requires educational research to be as precise and informative as possible. Results of research in education are used in a wide variety of ways, often with potentially critical fiscal, political, and practical implications. As such, current issues in educational research span a wide variety of topics; from decisions on appropriate and critical subjects to be studied and funded (e.g., curriculum effectiveness, student achievement), to how that information should be communicated to key members of the educational community (policy makers, researchers and practitioners).

One of the outcomes of the call for increased accountability in education is an emphasis on science-based research and assessment of educational effectiveness. The recent No Child Left Behind Act (United States Department of Education, n.d.) legislation is but one example of this increased emphasis on educational accountability. Although it is critical that methods used in research be judiciously selected, carefully designed, and fastidiously implemented, the analysis of the data and reporting of the findings must also reflect a rigorous attitude and practice. Discussions on research methods seem commonplace yet the criticality of reporting practices and protocols should

not be overlooked or marginalized. A research study may follow all the tenants of sound design and conduct, but if results are not presented properly and thoroughly, it is possible, and maybe even probable, that consumers of the research may be misled or, even worse, misinformed about the strength of meaning and applicability of the findings. Therefore, researchers must be made aware of, and held accountable for, proper reporting procedures and protocols.

Statement of Problem

The need for awareness of, and compliance with, proper and thorough research reporting practices is the primary inspiration for this study, which focuses on the differences in the strength of inferences that may be drawn as a result of how a researcher chooses to present his or her findings. Through the review and analysis of previously conducted and published research, this study illustrates the impact that reporting practices may have on how results are interpreted and presented by researchers. With a clear demonstration of the differences that may result from how findings are reported, it is anticipated that the appreciation among researchers for the need to approach reporting their results with the same degree of rigor they use when designing their studies and analyzing their data will be enhanced.

Among the vast variety of reporting issues, two in particular have garnered growing interest, and at times conflict, in recent years: (1) how should results be reported to adequately convey their importance and meaning, e.g., significance testing with p-values vs. effect sizes, and (2) how well does the representation of results communicate the precision of the findings, e.g., point estimates vs. confidence intervals (Thompson,

1998; Nix & Barnette, 1998). The last two editions of the American Psychological Association ‘s (APA) Publication Manual (1994, 2001), as well as the 1999 report by Wilkinson and the APA Task Force on Statistical Inference, both recommend and encourage the use of effect size reporting as well as confidence intervals. Fidler and Thompson (2001) provide three very specific recommendations based on the findings of the task force: (1) “Always provide some effect-size estimate when reporting a *p*-value” (p. 599 of the statistical task force report), (2) report confidence intervals as they provide more information than what is available from a decision of yes or no based on a single point estimate, and (3) graphical representations of confidence intervals will aid in data presentation and interpretation. With a variety of factors influencing the most appropriate way(s) of reporting research findings, the debates that result from differing viewpoints about what and how findings should be reported are not likely to be easily resolved. This complexity of influences thus necessitates further exploration of the impact of research reporting practices and protocols.

The growing importance of effect size and confidence interval reporting is further supported not only by a seemingly ever-increasing presence of professional journal articles on the topic but also by a text devoted entirely to the issue of effect sizes (Harlow, Mulaik, & Steiger, 1997). In addition, the summer 2001 publication of an entire issue of *Educational and Psychological Measurement* devoted primarily to these two topics (Vol 61(4), August 2001) further underscores their growing importance in the field. Within this text and journal are numerous articles and papers by a wide range of researchers that cover many aspects of effect size and CI reporting including specific issues with non-

centrality, fixed and random-effects designs as well as statistical power. The recognition of the criticality of reporting effects sizes and using confidence intervals by such recognized authorities as the American Psychological Association as well as professional journals such as *Educational and Psychological Measurement* should leave little doubt about the growing recognition of these two statistical measures as necessary elements in solid research reporting.

Robinson, Fouladi, Williams and Bera (2002) note that "Curiously, no researchers have attempted to determine how the inclusion of effect size information might affect readers' interpretations of research articles" (p. 370). One goal of the proposed study is to address this specific issue, albeit indirectly. It is indirect as this study will be focused on how using these reporting methods might affect conclusions, recommendations, and implications reached by researchers, not a means to empirically assess readers' interpretations.

Purpose of Study

The primary goal of this research is not to advocate the appropriateness of specific statistical tests (ANOVA, t-test, etc) or effect sizes (Cohen's d, Hedges g, Cohen's f, etc.) or methods of computing confidence intervals (bootstrapping, student's t, etc.); rather it is to provide a sense of how reporting results in different ways may affect the strength of inferences that can be obtained from a study and, as a consequence, the potential impact on results, conclusions, implications and recommendations made by researchers. It is anticipated that with clear examples and illustrations of how representing findings can potentially alter the conclusions drawn from specific research studies, educational and

other social science research professionals will gain an even greater appreciation for the importance and criticality of reporting results in a variety of appropriate and meaningful ways to better understand what the data represent. Potential differences that may result from data interpretation using statistical significance approaches compared to practical significance approaches are vital in the understanding of why one or the other alone may not be sufficient. Additionally, the context and purpose of the study underscores interpretation of these two types of measures.

The first issue of interest in this study concerns determining how the significance of results should be reported and interpreted. That is, does one consider statistical significance, as determined by testing a given null hypothesis and focusing on resulting p-values, sufficient? Or should other indices of significance, e.g., effect sizes, such as Cohen's d, be reported instead of, or in addition to, p-values or similar statistical significance measures? Often one finds these two ideas classified, respectively, as statistical significance and practical significance (Fan, 2002; Thompson, 2002a; Robinson & Levin, 1997). Fan presents these two approaches as analogous to two sides of a coin, saying "they complement each other but do not substitute for one another." (2002, p. 275)

The second issue of interest concerns not just what should be reported, but how. Of particular interest is whether a point estimate is sufficient or is it better to use some measure of specificity, such as a confidence interval approach? To complicate this issue even more is determination of an appropriate method for constructing intervals around such measures as effect sizes, which can be much more complex than the more common and accepted practices of constructing confidence intervals around descriptive statistics

such as the mean (Thompson, 2002b).

Research Questions

The main objectives and focus of this research lead to three questions:

1.) To what extent does reporting outcomes of tests of statistical significance vs. tests of practical significance result in different conclusions and/or strengths of inference to be drawn from the results of research?

2.) To what extent does reporting confidence intervals instead of, or in addition to, point estimates affect the conclusions and inferences to be drawn from the results of research?

3.) What method, or combination of methods, is recommended for reporting results in educational studies?

Study Significance

Today's educational atmosphere is highly laden with assessment and accountability issues. Researchers need to be attuned to the need for effectively communicating the practical impact of research results in addition to, or possibly in lieu of, merely reporting findings that are statistically significant. The use of effect sizes and confidence intervals can be key elements in aiding in this communication. Effect sizes provide a means of measuring practical significance and confidence intervals convey the precision of results. The difference between a tight confidence interval and wider confidence interval cannot be underestimated when discussing study implications.

Oft-criticized for substandard practices and products (see, for example, Davis, 2001 and Gall, Borg, & Gall, 1996), educational researchers must increase their

awareness of, and compliance with, sound research methods, including how they report their research. The increased emphasis on accountability in education is not limited to the practitioner. The educational researcher is also likely to be under closer scrutiny as time progresses and resource expenditures for educational program evaluation continue to climb.

When applied to ongoing research in education as well as the other social sciences, the ability to construct effective and efficient confidence intervals that provide precise data summaries will enable decision-makers at all levels of the educational system to make better decisions based on more precise and accurate information about the effectiveness of interventions, curriculum and other aspects of the educational environment. Technology is available to support these enhanced methods and there is not a viable excuse not to pursue and develop the abilities to use confidence intervals instead of point estimates for numerous statistical estimations, including the increasingly critical estimate of effect size.

Limitations

This is an initial investigation into using confidence intervals and effect sizes in addition to, or in lieu of, traditional significance test results beyond the theoretical and conceptual level. It is based on previously reported research and is therefore limited in its ability to predict performance with untested data. That is, it is recognized that reported research is typically research that has shown to have an effect or significant finding. This study, like many meta-analytic studies, is subject to bias due to the exclusion of research studies that may have fallen victim to the 'file drawer' syndrome (Bradley & Gupta,

1997; Rosenthal, 1979; Rosenthal, 1995; Gall, Borg, & Gall, 1996; Riechardt & Gollab, 1997). These studies are likely to have either shown a non-significant result or show evidence in the opposite direction of the hypothesis (Bradly & Gupta, 1997). Therefore, it is possible that studies that haven't been reported because they showed a small or non-significant effect might have a wide confidence band and that if those studies were revisited, using confidence intervals instead of point estimates, it is possible that the null hypothesis might not have been subject to a Fail to Reject (FTR) decision in a definitive fashion, but rather with an awareness that the decision to Fail to Reject may have been a result of a very large confidence band that barely extended to the point of non-significance. Such awareness could provide the researcher with a strong theoretical foundation for his or her alternative hypothesis, but has a weak study design, with enough justification to repeat the study with an improved design (e.g., larger sample sizes).

Cohen's effect sizes (Cohen's *d*, Cohen's *f*, and Cohen's f^2) are just a few of a myriad of effect indices available. They were selected for this study for a variety of reasons, including commonality of use and the oft-desired characteristic of standardization when using multiple studies and scales; however, the use of these statistics does not imply that they are always the most appropriate for a given study. The purpose of a study, nature of the data, and selection of data analysis methods may make the use of different effect sizes more appropriate. Additionally, even when they may be deemed as appropriate statistics to be used in a study, the context and criticality of the study itself is essential for proper interpretation of index values. As the purpose of this study is to investigate how different reporting processes may affect findings and not an

investigation of study method, purpose, and/or strength, this contextual issue, though recognized as a valid and important topic, is not considered to be a primary issue in this study. Likewise, the regular use of Cohen's *d*, Cohen's *f* and Cohen's f^2 throughout the study permits a consistency necessary to make communal decisions and comparisons.

A final limitation of this study pertains to the issue of who is doing the interpretation of the findings. This study is primarily focused on how the researcher(s) of a particular study analyze, interpret, and report the results of their research. Of core interest in this research is an investigation of how different analyses and reporting practices might impact the conclusions and recommendations made by the researcher. Also of interest is how the choice of method of reporting findings may impact the magnitude of strength of the findings. What is not investigated in this study, but is acknowledged as being of fundamental and vital importance, is the consideration of the impact of reporting practices and protocols on the consumer of the research, that is, the practitioner who reads and interprets the findings presented. This type of research question has been addressed to a slight degree (Robinson, Fouladi, and Williams, 2002) and deserves further consideration and investigation external to this study.

Definitions of Terms

The following definitions are provided for clarification. Some of the terms used, e.g., practical significance, have various interpretations depending on the source; the definitions provided were chosen to best reflect how they are intended to be used and interpreted within this study.

Cohen's d: One method of computing an effect size, this measure of effect size is determined by taking the difference of the two sample means and dividing by the pooled standard deviation (Cohen, 1988).

Cohen's f: An effect size often used with ANOVA significance tests, given by:

$$f = \frac{\sigma_m}{\sigma}$$

where σ_m is the mean standard deviation of the means of *k* number of groups around the grand mean and σ the standard deviation of the common population . Values can range from 0, when there is no difference between groups, to, at least theoretically, infinitely large as σ_m increases in magnitude relative to the population mean (Cohen, 1988)

Cohen's f^2: An effect size measure calculated in correlational/multiple regression studies given by:

$$f^2 = \frac{PV_S}{PV_E}$$

where PV_S is the proportion of variance accounted for by the source, or predictor variables, and PV_E is the proportion of variance accounted for by the residuals (Cohen, 1988 and Cohen, Cohen, West, & Aiken, 2003)

Confidence Interval: An interval containing "a range of possible values, so defined that there can be high confidence that the 'true' values, the parameter, lies within this range" (Glass & Hopkins, 1996, p.261). Boundaries are calculated as a function of the level of Type I error designated. Other variables and

characteristics of the study are also taken into account but are dependent on the method of confidence interval estimation used.

Effect Size: An estimate of the magnitude of a difference, a relationship, or other effect in the population represented by a sample (Gall, Borg & Gall, 1996).

Eta-squared (η^2): A measure of association used in ANOVA analyses, this is a measure of variance accounted for by group membership given by:

$$\eta^2 = \frac{SS_B}{SS_T}$$

where SS_B is the Sum of Squares between groups and SS_T is the Sum of Squares Total (Stevens, 1999).

Meta-Analysis: As defined by Hedges and Olkin (1985), Meta-analysis is "the rubric used to describe quantitative methods for combining evidence across studies" (p.13).

Multiple Correlation Coefficient (R^2): A measure of association that provides the proportion of variance of a dependent variable that can be predicted, or accounted for, by the predictors in the model, given by:

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$

where SS_{reg} is the Sum of Squares due to regression and SS_{tot} is the Sum of Squares Total (Stevens, 1999).

Point Estimate: A specific, single quantitative value used to estimate a parameter (Glass & Hopkins, 1996).

Practical Significance: Often a term associated with effect sizes, this is the concept of "evaluating the practical noteworthiness of results" (Thompson, B., 2002a, p.65).

Significance Tests: Statistical tests conducted that lead a researcher to make a decision, either Reject or Fail to Reject. In this study, the Reject-Support approach will be employed (Steiger and Fouladi, 1997) in which a decision to Reject actually supports the researcher's expectations (e.g., that there is a difference in populations) as it is the primary school of thought used in most social science research.

Statistical Significance: A means of using quantitative, probabilistic interpretations to determine whether to Reject (or Fail to Reject) a given null hypothesis (Gall, Borg, & Gall, 1996).

Type I Error: The error that occurs when a researcher incorrectly rejects a True null hypothesis (Glass & Hopkins, 1996, p. 259).

Chapter Two

Review of Literature

This review of the literature is intended to provide a concise yet comprehensive overview of the controversies and explorations relative to significance reporting as well as the use of point estimates compared to confidence intervals. It is divided into five main areas of review. First, an overview of research reporting practices in education, both in general and as a function of study type and method is provided. Second, disciplinary norms and the need to consider their influence when reading research from different disciplines are discussed. Next, a synopsis of effect size uses and characteristics is given. After the discussion on effect sizes, a discourse on the controversy surrounding statistical versus practical significance testing is presented. And finally, there is an overview of the discussions and differences of opinion regarding the use of point estimates compared to confidence intervals.

Reporting Research

Appropriate, effective and meaningful reporting practices are critical for communicating research results correctly. Thoughtful interpretation of research and the ability of readers to sift through good and bad research have gone beyond being merely a part of courses in research methodology. Books are now being written to provide readers not only with a sense of interpreting research itself, e.g., Hittleman and Simon's

Interpreting Educational Research: An Introduction for Consumers of Research, 2nd ed. (2002), to entire books about determining the quality of the research (see, for example, *Making Sense of Research. What's Good, What's Not and How to Tell the Difference* (McEwan & McEwan, 2003) and *Evaluating Research Articles from Start to Finish, 2nd Ed* (Girden, 2001)). The mere fact that there is a market for such books is indicative of the lack of trust and/or perceived rigor in research conduct and reporting.

Although poor conduct or design of research must always be a concern, it is also unfortunate that the reporting practices themselves can leave a lot to be desired. The less ethical researcher might alter how they report findings, including only information that supports his or her hypothesis, or present results in such a way as to misinform or mislead the reader. In his book *Statistics as Principled Argument*, Abelson (1995) provides numerous examples of how this might be accomplished. For example, the conduct of numerous types of tests on the same data may be suspect unless clearly justified. As he illustrates on p. 70, "If you look at enough boulders, there is bound to be one that looks like a sculpted human face". Other issues he takes research reporting to task on are those that use rhetoric to justify results not quite meeting the desired conclusion (e.g., p-values of .07 when desired Type I error rate is .05), wording that 'hints' at more in-depth meaning than the data clearly indicate, and findings reached from distributions and/or statistics that are 'strange' (e.g., outliers and/or 'dips' in data distributions, statistics that are logically too small, too large or defy logic). Abelson presents cautions about using statistics (p-values) void of reason, logic, and judgment. While Abelson provides important cautions about interpreting research as well as beneficial guidance on how to

use statistics to support research, his, along with others, concern about the misuse of statistics is not new. Almost half a century ago, a still oft-used book by Huff (1954), *How to Lie With Statistics,* provides the interested reader with numerous examples of how the public had been misled through advertisement and research results during that time-frame. The fact that these types of issues still exist and may even be worse, is a sad and troubling reflection on current research, especially considering the presumably on-going advances in statistical methods, applications, and understanding.

Educational research specifically is often criticized for poor research practices. In their text titled, appropriately enough, *Educational Research*, Gall, Borg, and Gall (1996) advise the reader in their section about studying a research report to "keep in mind that the quality of published studies in education and related disciplines is, unfortunately, not very high" (p. 151). In a review of analytical practices of studies contained in 17 fairly prominent social science journals, Keselman, et al., (1998) noted that 'The present analyses imply that researchers rarely verify that validity assumptions are satisfied and that, accordingly, they typically use analyses that are nonrobust to assumption violations" (p. 350). Tuckman (1990) found that when it came to educational research "much of the work in print ought not to be there" p.22).

The editor of the *Journal of Curriculum and Supervision (*Davis, 2001), provides a succinct yet thoughtful discourse on educational research reporting practices in general. While potentially harsh, the issues discussed in this article provide one with a sense of the impact that poor or inadequate research reporting can have on practice. He states on page 9 that "Educational research inattentive to meanings corrupts the enterprise of inquiry and

fails its obligation to practice." Davis hints at the possibility that ineffective and inappropriate reporting, hopefully a relatively innocent result of unfortunate ignorance of the subject, context, or proper procedure, may also be intentional on the part of the researcher. As such he notes that "Educational research has the moral purpose to inform—not to direct or to control educational practice" (p. 9). Davis also recognizes that the responsibility for good decision-making does not necessarily rely solely on the researcher as the practitioner has a moral duty to be capable enough to discern what the research is telling him or her. However, if the research is not communicated properly and effectively, the practitioner has little, if any, real opportunity to put the research to proper use.

Disciplinary Norms

Understanding that attributes of particular sciences or disciplines differ in many aspects, including written communications, is important to consider when reviewing literature present in various disciplines. Parry (1998) provides a succinct discussion on the importance of disciplinary norms within scholarly writing, including the need to address this issue during the preparation of future academic scholars. She discusses the absence of clear understanding of what disciplinary norms are and attempts to aid the newcomer to this type of knowledge through a vast discussion on previous literature on this aspect of research. Essentially, one might think of disciplinary norms as the conventions, rules, and/or practices, explicit or implicit, that one finds within a certain body of scholarly literature relative to a given discipline.

According to Becher (1987) there are broad disciplinary groupings that encompass a wide variety of disciplinary norms. Furthermore, the conventions of writing and language within different disciplinary norms vary and often are not explicit in nature; rather these norms must be learned through observations within different disciplines and subdisciplines. As such, Gersholm (1990) asserts that many of these norms are implicit and must be learned through tacit means.

Social science research reporting, according to Bazerman (1981), tends to lean toward persuasion due to the potential differences in methodological and theoretical frameworks in the scholarly community. He also identifies six attributes that may be found to contribute to differences in written research as a function of discipline. These attributes include conventions regarding the type of knowledge, traditions, external accessibility of knowledge, degree of technicality, methodological and theoretical considerations, and writing mechanics associated with a given discipline. Becher (1987) asserts that four overlapping domains exist within linguistic preferences and styles in different disciplines: modes of formal scholarly communication; how writers assert field-unique tacit knowledge; guiding conventions of citing and referencing previous research; and traditions of argument structure.

Depending on the discipline umbrella under which research is written, different practices and accepted conventions may be evidenced in different manners depending on the particular field in which the research is conducted and disseminated. As such, it is necessary for consumers of research originating in different disciplines to acknowledge

that underlying differences exist and, at a minimum, be sensitive to those differences when considering the quality, nature, and intention of the research.

Effect Sizes

Effect size has become increasingly recognized as an important statistic that needs to be reported. Numerous field experts have stressed the need for effect size reporting throughout the social sciences, including education (Nix & Barnette, 1998). Both the fourth and fifth editions of the American Psychological Association (1994 and 2001) highly recommend that researchers report effect sizes. Often termed practical significance or, sometimes substantive significance (Robinson & Levin, 1997), effect sizes provide a different, albeit related, piece of information about how a treatment or other variable is impacting the issue of interest.

There are various effect size indices available as well as different terms used when referencing effect sizes. Some of the various descriptors for effect size estimates include percent of variance accounted for, strength of association, and magnitude of effect, among others (Plucker, 1997). Additionally, correlation coefficients such as Spearman rho and the Pearson Product Moment Correlation Coefficient are sometimes considered a type of effect size (Plucker 1997). Hedge's g, Glass's Δ, and Cohen's d are all variations of effect sizes for differences in means between two groups (Rosenthal, 1994 and Cohen, 1988). Effect sizes for studies using statistical methods examining correlational relationships or variance relationships have measures such as eta-squared (η^2), R-squared (R^2), and omega squared (ω^2) available for use (Snyder & Lawson, 1993).

In his book *Statistical Power Analysis for the Behavioral Sciences*, Cohen (1988) provides effect sizes for various types of analyses including those that can be used in t-tests, Chi-square tests, and multivariate tests, just to name a few. Ultimately, of course, the selection of effect size indices is a factor of many considerations, including purpose of the research, data analysis to be employed, and the nature of the data. For example, a decision on whether to use Hedge's g or Glass's Δ, may depend on the disparities between the groups in sample size and variance (Rosenthal, 1988).

Statistical vs. Practical Significance

The literature over the past decade seems inundated with articles and tomes pleading for, as a minimum, inclusion of effect sizes when reporting research results (see, for example: Plucker, 1997; Thompson,1998; Thompson, 1999a; Fan, 2001; and Fouladi & Williams, 2002). In his review of studies reporting effect sizes in gifted education, Plucker describes the relationship between statistical significance and practical significance as analogous to a chasm in the earth. In his illustration, he uses the p-value of a significance test as the indication that the chasm exists, and the effect size reported as the measure of the width of the chasm.

Both of these concepts of significance, as they tend to be thought of today, are products of the last century. During the early 1900s, such groundbreakers of modern statistical concepts such as Karl Pearson, Ronald Fisher, and Jerzy Newman, among others, provided the conceptualization and formal development of null hypothesis based significance testing (Harlow, 1997). However, it wasn't until around the middle of the 20th century that significance tests started taking a dominant role in research literature.

Hubbard and Ryan (2000) reviewed articles in 12 prominent journals of the American Psychological Association and found that until 1940, significance tests only appeared in empirically based research about 40% of the time or less. By 1960, the popularity of using significance tests rose to such a degree that over 90% of empirical research reported findings using some type of significance-based analysis. Interestingly, it is during the rise of publication popularity that the notion of statistical inference testing using a null hypothesis approach began acquiring a vocal set of detractors (Mulaik, Raju, & Harshman, 1995; Rozeboom, (1960). As time has progressed, the popularity of reporting significance tests has continued while at the same time the debates about using other reporting methods, e.g., effect sizes and confidence intervals, has continued to grow stronger and more frequent.

There is a portion of researchers who go so far as to advocate the use of effect sizes in place of, not merely in addition to, the traditional significance tests (Schmidt & Hunter, 1997; and Meehl, 1997). Others are more moderate and take a 'middle of the road approach', arguing that the use of effect sizes and/or tests of significance are both useful, depending on context and purpose of the research. Muliak, Raju, and Harshman (1997) provide arguments for inclusion of indices of practical significance in many cases but also suggest that elimination of significance testing is neither warranted nor desired. They illustrate how influences of factors such as the power of a given study may limit the desirability of relying on significance tests but argue that significance testing has an objective nature that requires the researcher to form an opinion based on theory and/or previous research before conducting the analysis. This required assertion of a formal

hypothesis *a priori* to data analysis helps preserve a certain sanctity of the research by avoiding potentially inappropriate data-driven hypothesizing about effectiveness of a given treatment or study effect.

Regardless of the position held by individual statisticians and researchers, there is little doubt that this topic is one of the 'hot buttons' of debate in educational research today. Within the past few years, an entire text was dedicated to this issue (Harlow, Muliak, & Steiger, 1997) as well as an edition of *Educational and Psychological Measurement* (Vol 61(4), 2001). However, it would be a mistaken notion to consider this to be an issue of recent origin. According to Schmidt and Hunter (1997, p. 58), a discourse by Jones in 1955 was one of the first, if not the first, to argue for the replacement of statistical significance with effect sizes (as well as confidence intervals) in Volume 6 of the *Annual Review of Psychology*. Since then, the topic has ridden a wave of periodic attention, often becoming the topic *du jour* for a period of time before taking a back seat to other topics of interest for a few years and then once again coming back to the forefront of attention. However, over the past decade, this issue has taken on a new and stronger life among researchers, and, rather than waning, appears to be continuing to gather momentum. From the afore-mentioned dedicated text and journal to the stronger stance taken by the APA on reporting requirements, resulting, at least in part, from the findings of the Statistical Task Force of 1999-2001 (Wilkinson, 2001), enhanced attention to the issues of effect size reporting and the use of confidence intervals is evident.

While the stance and beliefs of individual researchers is critical to their personal motivation to report effect size estimates, actual reporting of such estimates is also an

indirect result of what publishers and journal editors demand and expect in submissions. In general, support for effect size reporting is growing as more professional journals across disciplines require such statistics for consideration for publication. At least 17 such journals, spanning areas of interest from careers, education, counseling and agricultural education currently require this information (Fidler & Thompson, 2001). Unfortunately, even though a growing number of journals are requiring effect sizes to be reported, many are not enforcing their own mandates for publication. A review of 13 journals by McMillan, Snyder, and Lewis (2002) that require effect size reporting revealed that most of those journals were not enforcing this particular constraint. Additionally, Devaney (2001) found in a survey of journal editors that while 93% of those surveyed agreed with the importance of effect size reporting, 73% indicated that inclusion of effect size information was not a requirement for consideration of a manuscript. These findings seem to indicate that while there is indeed a perceived need to report effect size information, there is little, if any, enforcement of such reporting. The reasons for this are not clear and it may well be the case that editors and others who make critical decisions on what research is noteworthy require more evidence about how reporting of findings may impact conclusions and the relative significance of findings resulting from a particular study.

Point Estimates vs. Confidence Intervals

Confidence intervals have been accepted for quite some time as a useful method for describing statistical parameter estimates such as sample means and can be traced back at least three decades (Meehl, 1967). The use of statistics to describe population

parameters is an imprecise science and the use of confidence bands around a given statistic allows researchers to gauge the precision of a given statistic and therefore can help determine the strength of conclusions and inferences that can be drawn.

Unfortunately, confidence intervals do not appear as frequently in research as might be desired. Reichardt and Gollob (1997) provide eight reasons for why this might be the case. These reasons, summarized, are: (1) conventional use of statistical test precludes consideration of use of intervals, (2) lack of recognition by researchers of situations conducive to the use of intervals, (3) less frequent production of intervals by computer programs as compared to results of statistical tests, e.g., p-values, (4) diminished size of actual parameter estimate and associated confidence interval is less impressive than reporting statistical significance alone, (5) magnitude of interval width might be large enough to inhibit potential for publication acceptance, (6) some statistical tests, e.g., chi-square test of association for a 2x2 table, do not have a unique parameter defined, thus necessitating additional steps to identify appropriate measures, (7) criticism of statistical tests, sometime themselves incorrect, rather than advocacy of interval strengths, dissuades uses, and (8) the incorrect and inappropriate association of interval use advocacy with statistical testing banning undermines and thus discourages the acceptance and application of confidence intervals.

These reasons for not using confidence intervals seem to fall into three main types of justifications for not using this technique. The first general type of aversion to using confidence intervals is, perhaps, the least alarming. The lack of use resulting from reasons (1), (2), or (3) appear to result more from lack of knowledge and awareness of

either the methods or tools available. These obstacles to using confidence intervals are likely to diminish as awareness increases and computer programs continue to increase in their sophistication. The second broad category of reasons for which one might be reticent to using confidence intervals seems to center around a researcher's concern that his or her research won't get published or recognized because confidence intervals or point estimates might diminish the strength of their findings (reasons (4) and (5)). These types of justifications (and associated ethical issues) seem to be, in some regards, the most insidious of the three and are likely contributors to the skepticism with which research is often viewed. The final broad category encompasses the last two items on the list. The lack of use of intervals due to these concerns have a more philosophical flavor and may be a factor of personal comfort with techniques and tools learned early in one's career (e.g., significance testing) and may be overcome by better communication of the benefits of confidence intervals and less villianization of significance testing.

Although there are issues associated with the lack of universal use of confidence intervals in research reporting, there have been recent advances in using confidence intervals for statistics other than the mean and standard deviation. The use of confidence intervals for other statistical estimates is quickly growing as an improved way of reporting more informative measures of estimates than point estimates. Cumming and Finch (2001) provide four reasons for researchers to give confidence interval estimates when reporting research findings: (1) confidence intervals provide both point and interval information that improves understanding and interpretation, (2) the use of intervals enhances the practice of traditional null hypothesis reporting, it does not negate

it. That is, if a specific null value is being tested and is found to fall outside of the computed interval, it is rejecting the null hypothesis, but with more precision, (3) the use of CIs may serve meta-analytical methods which focus on estimation using many sources of study data, and (4) information about the precision of the study and subsequent findings may be gained through the use of intervals.

In Figure 1, results of four hypothetical studies are illustrated with computed confidence bands around the effect size (Cohen's d, in all cases).

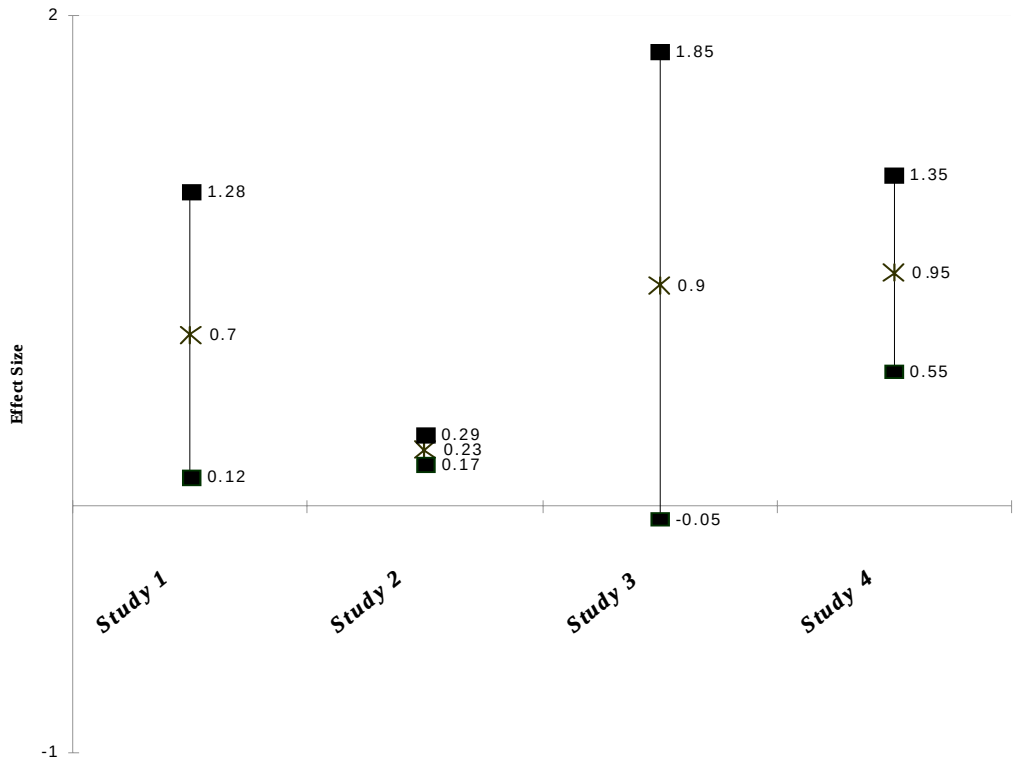

Figure 1. An illustration of various confidence band widths.

In studies, 1,2, and 4, the decision based on statistical significance testing would have been to Reject the null hypothesis. However, this illustration helps demonstrate that the strength of the inference to be drawn from such a conclusion is not consistent. Depending on whether one considers effect size in addition to statistical significance and/or confidence intervals in addition to point estimates can dramatically impact how one interprets the findings and the certainty one places on the associated Reject or Fail to Reject decision.

In study 1, a report of the effect size point estimate only would support the findings of the significance test; however, the lack of precision of the results indicates that the population effect size might be as small as 0.12, a rather minor effect, or as large 1.28, a very large effect. In this case, the reporting of the effect size doesn't really change how one views the results; however, the inclusion of confidence intervals very well might have an impact on interpretation of findings.

In study 2, the opposite phenomenon occurs. In this case, the confidence interval is very tight. A bandwidth of 0.12 indicates high precision of the estimate and one is likely to be confident that there is a statistical difference found in the study. However, an effect size of 0.23 is considered small by Cohen, so although one is likely to have little doubt that there is really a difference, the practicality of the difference is very small. At this point, the context and purpose of the study would be primary determinants in deciding whether such a small measure of practical significance is worthy of pursuing.

In study 4, neither the use of a measure of practical significance and/or confidence interval has the potential for as dramatic an impact on interpretation as the first two

studies did. In this case, although the confidence band still indicates a rather large amount of error in the sample, the effect size is large enough that, at a minimum, the effect is moderately strong (d = .55).

The final study considered, study 3, may illustrate one of the most compelling reasons to use confidence intervals, especially when one Fails to Reject the null. In this case, using statistical significance tests alone would likely result in the unfortunate 'file drawer' syndrome (Bradley & Gupta, 1997; Rosenthal, 1992; and Rosenthal 1979) previously discussed. The researcher would put away this particular line of research inquiry and pursue other endeavors. Using effect sizes and/or confidence intervals, however, the results of the significance test lose quite a bit of credibility. The effect size of 0.9 is large by virtually any standard and the confidence interval clearly indicates that the decision to Fail to Reject was not reached by a large margin. If nothing else, this type of result would indicate that further pursuit of this research is warranted, hopefully with attention paid to increasing power of the study through larger samples, better controls, more potent treatment, etc.

Estimates made prior to conducting a particular study can help guide and inform study design while follow-up of results will provide greater precision about the potential interpretation and inferences that can be drawn from the findings. Confidence intervals provide a measure of precision for statistics and can provide decision makers with yet a better sense of how strong or reliable a reported statistic actually is.

Methods of constructing confidence intervals are as much of a concern as whether to use them or not. Factors such as sample size, distribution shape, variance

heterogeneity, and reliability must be taken into consideration as well as the nature of the parameter to be estimated when deciding on an appropriate method of constructing these intervals. Confidence intervals for descriptive statistics such as the mean and standard deviation are fairly commonplace and have been around for many years. It is only in more recent years that investigation into constructing confidence intervals around statistics such as the multiple correlation coefficient, Cohen's d, Cronbach's Alpha and others have been investigated (see, for example, Steiger & Fouladi, 1997; Fidler & Thompson, 2001, Carpenter & Bithell, 2001, and Fan & Thompson, 2001). Although the argument for effective construction of confidence intervals for a larger variety of statistics have been, at least theoretically, around for many years, it is only within recent years, due, at least in part, to the recent explosion of technology sophistication, that more computationally demanding methods such as Steiger and Fouladi's interval inversion method (1992) have been able to be implemented.

Nine techniques for constructing confidence intervals have recently been examined using Monte Carlo techniques for the indices of practical significance to be used in this study (see Kromrey & Hess, 2001 and Hess & Kromrey, 2003 for details). In general, the Steiger and Fouladi interval inversion method (Steiger & Fouladi, 1992) and Pivotal Bootstrap method (Carpenter & Bithell, 2001) showed the best results, followed by the Normal Z computation for approximately homogeneous samples (Kromrey & Hess 2002 and Hess & Kromrey, 2003). Due to the design of this study, a bootstrap technique such as the Pivotal Bootstrap is not tenable as only summary data and statistics were expected to be available. Therefore, confidence band interval construction was limited to

using the most promising equation based algorithm found in these studies for the type of analyses considered, e.g. the Fisher Z-transformation for R^2, as well as the computer-intensive Steiger and Fouladi methods. Both the hyperbolic sine transformation and student's t show some promise in selected applications; however, they did not add anything to using the simpler computations chosen and therefore were eliminated as unnecessary transformations.

Examples

To illustrate the potential impact of using different reporting practices, or a combination of reporting practices, two studies that reported significant findings were examined. In the first study (Nosek, Banaji, & Greenwald, 2002), the researchers were interested in investigating how an individual's implicit and explicit attitudes and self-identities regarding math and science differed from their implicit and explicit attitudes and self-identities with the arts as a function of their gender. The author's reported significant findings on students' math/arts attitude and identity depending on gender using an alpha of .05. The study did not report effect size information or confidence intervals. Using the data provided, effect sizes for a correlational analysis, f^2, were computed (Figure 2). Using guidance provided by Cohen (1988), these results reflect effect sizes approaching large (attitude $f^2 = 0.32$) and medium (identity $f^2 = 0.14$) measures of practical significance. The magnitude of these effects tend to provide further support for the findings of the researchers that gender has a significant impact on student's academic attitude and identity; however, the strength of these assertions is somewhat diminished when one computes confidence intervals around these effect sizes

(Figure 3). The use of confidence intervals provides more information that should impact the types of conclusions and merit given to these conclusions.

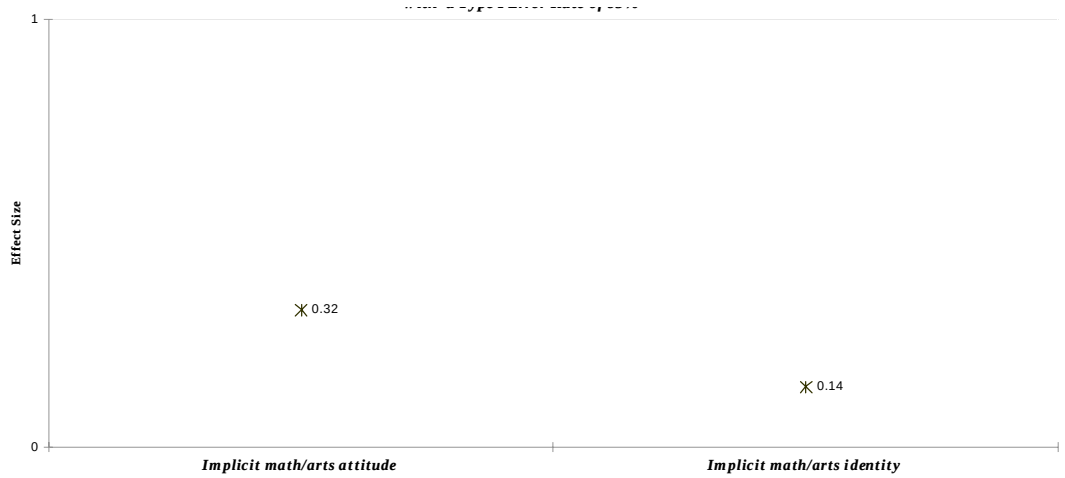

Figure 2. Point estimates (Cohen's f^2) of the impact of gender on Mathematics Attitude and Identity.

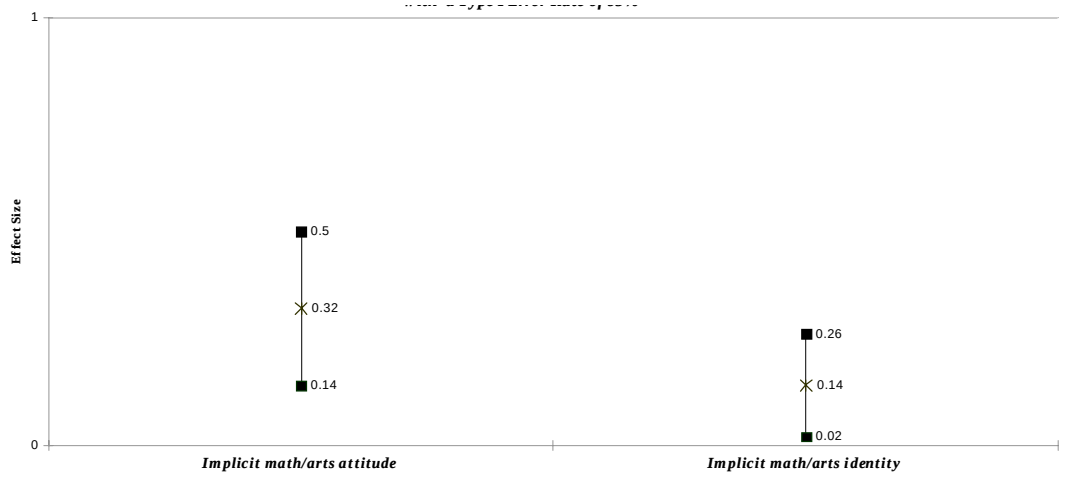

Figure 3. Point estimates (Cohen's f^2) and confidence intervals on the impact of gender on Mathematics Attitude and Identity at a Type I error rate of .05.

While the relatively large width around the attitude measure does not weaken the argument for gender impact on attitude too severely (the lower limit still reflects a medium effect), the confidence interval around the identity variable provides evidence that the impact of gender on a student's math/arts identity may not be very influential after all. The lower limit in this case is 0.02, a very small, almost non-significant effect. In this study, the provision of confidence intervals adds important information necessary to report the findings adequately and comprehensively.

In another study (Fitzgerald, 2000), the researcher investigated the impact of an intervention on a student's reading achievement as, at least in part, a function of the intensity of participation reported significant differences between students who received the treatment for the duration of the program (25 weeks) as compared to those students who were only enrolled in the treatment for a fraction of the program (6-12 weeks). Similar to the first study, the calculation of effect size (Cohen's d) still tended to support the author's conclusion about effectiveness (d = 0.7, a large effect according to Cohen); however, the construction of confidence intervals around the

treatment intensity on gains in students' instructional reading levels at a Type I error rate of .05. effect size (Figure 4) again weakens the definitiveness with which one might regard the results. In this case, the confidence interval is approximately one full standard deviation wide, with a lower limit reflecting a very small effect and an upper limit reflecting a huge effect. The imprecision of the measurement should be clearly

represented in reported findings using a tool such as a confidence interval to fully inform the reader.

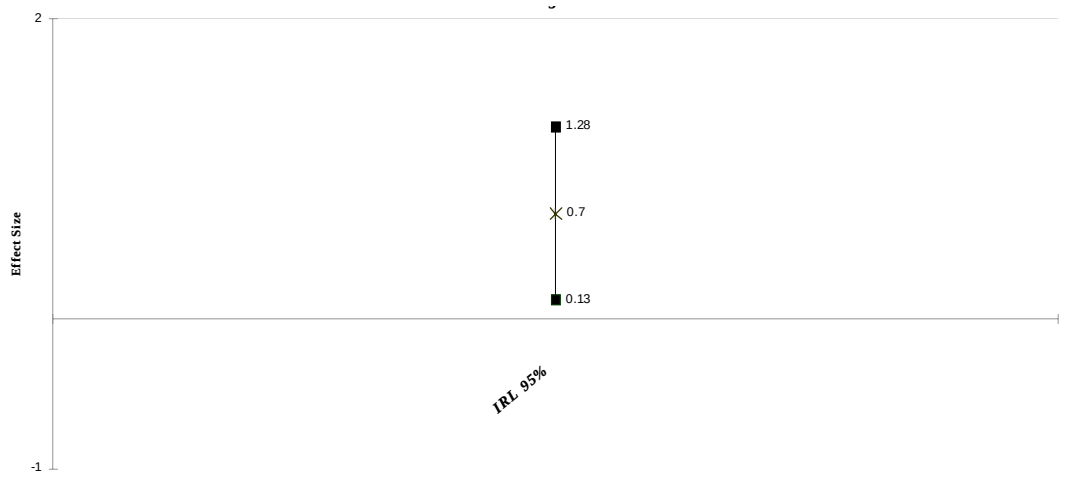

Figure 4. Point estimates (Cohen's d) and confidence intervals on the impact of treatment intensity on gains in students' instructional reading levels at a Type I error rate of .05.

Summary

While the recognition of these two elements of research reporting, effect sizes and confidence intervals, appears to be growing over the last decade, they are not new to debate among statisticians and researchers. The theoretical knowledge and conceptual basis of effect sizes can be traced back to early in the 20th century (Harlow, 1995). The use of confidence intervals as they are currently applied can be traced back at least three

decades (Meehl, 1967). However, it is only due to the recent advances in technology and availability of high-powered computers to the average researcher that has enabled the use of more advanced and precise techniques. Statistical software packages available commercially in the past few years readily report and compute different statistics that used to require extensive programming and calculations by the researcher (Fidler & Thompson, 2001). These computations, probably taken for granted by many researchers in the past few years are only recent when one considers the historical evolution of tools.

Given the fact that these reporting issues have relative longevity as issues in the statistical and research world, an attempt to at least broach the issue from an applied setting is called for. Additionally, since the lack of appropriate mechanisms and necessary technology is no longer a barrier to conducting this type of research, it is imperative that beginning steps be taken to start to bridge the conceptual and theoretical world of research to connect with the realistic and applied world of research. This study is intended to begin building such a bridge.

Chapter Three

Method

The general purpose of this study, to investigate the impact of reporting practices on the types of conclusions reached by researchers, is supported by three questions:

1.) To what extent does reporting outcomes of tests of statistical significance vs. tests of practical significance result in different conclusions and/or strengths of inference to be drawn from the results of research?

2.) To what extent does reporting confidence intervals instead of, or in addition to, point estimates affect the conclusions and inferences to be drawn from the results of research?

3.) What method, or combination of methods, is recommended for reporting results in educational studies?

To address this purpose and associated questions, this study goes beyond the rhetoric and philosophical arguments currently found in most of the literature published regarding this issue. Rather, actual studies already deemed worthy of professional consideration and use by others in the field, as evidenced by publication in peer-reviewed journals that are well-known and used throughout professional circles, were examined to determine if alternative conclusions, and/or differences in inferential strength might have resulted from different analysis and reporting procedures.

Study type and description

The nature and objective of this study are such that it does not cleanly fit into one classification or type of study. It uses techniques that are both qualitative and quantitative in nature but is not one or the other explicitly. As such, it takes on a mixed method approach and might reflect the type of study that Tashakkori and Teddlie (1998) call a mixed model design with multilevel uses of data, using different types of analyses and methods of analysis at different levels of the study. Summary data, not original raw data, are used so it cannot be considered a secondary data analysis. Probably the closest description of this study would be to consider it a mixed method design with a blending of meta-analytic methods (Hedges & Olkin, 1985) and a methodological research review (Keselman, et al., 1998).

Meta-Analysis. While there is evidence of research synthesis across studies as far back as 1904 (Cooper & Hedges, 1994, p. 5), the now common term, meta-analysis, debuted courtesy of Glass (1976). He defined meta-analysis as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p.3). Over the past three decades, the use of meta-analysis has increased at a tremendous rate not only in the social sciences, but also in other fields such as medical research. According to Cooper and Hedges (1994), only four books and two major papers emerged in the first half of the 1980s. This rather limited number of resources has expanded virtually exponentially over the last decade and a half. A cursory search of literature reveals a much more detailed list of resources dedicated to meta-analysis, its techniques, uses, and applications. Cooper and Hedges also discuss how studies using

meta-analytical techniques have increased in conjunction with resource materials. A search of three major databases (PsycINFO, ERIC, and Social Scisearch) over a 15 year time period (1974 to 1989) revealed almost a non-existence of meta-analytic studies the first four years considered, 1974 to 1977. Beginning in 1978 an approximately exponential growth was seen with about 18 studies reported across the three databases in 1978 to almost 300 meta-analytic designed studies in 1989. It is highly likely that this type of growth in using meta-analysis in social science research has continued.

Traditionally, meta-analysis is used to synthesize findings across studies with a common theme or substantive research question, e.g., gender difference impact on mathematics or effectiveness of new medications for members of different populations. In traditional meta-analytic studies, researchers gather primary research studies pertinent to their topic of interest (often with varied and disparate findings and conclusions), code articles to determine relative strengths and weaknesses, and perform statistical calculations, typically in the form of effect sizes, to determine effectiveness of a treatment, magnitude of difference between groups, etc. There are a myriad of forms these different steps can take, any one of which would likely be worthy of further investigation. However, this study, while using a meta-analytic approach through the synthesis of findings from different studies, has a slightly different research focus. Rather than targeting a specific topic or applied research question, a meta-analytic philosophy was used to examine effect of chosen statistical analysis and chosen reporting method(s) on interpretation of findings using various studies found in published research that has potential implications for educators.

Methodological Research Review. Since this study is more focused on method and practice, one might also consider it to be, at least in part, a methodological research review. According to Keselman, et al. (1998, p. 351) these types of reviews tend to have two main purposes: (a) to form a basis for recommending improvements in research practice, and (b) to use as a guide for procedures to be taught in methods courses.

The American Educational Research Association offers the following definition of a methodological review to be considered when submitting an article to their journal *Review of Educational Research*: "descriptions of research design, methods, and procedures that can be employed in literature reviews or research in general. The articles should highlight the strengths and weaknesses of methodological tools and explore how methods constrain or open up opportunities for learning about educational problems. They should be written in a style that is accessible to researchers in education rather than methodologists." (AERA, 2003).

A review of some of the studies that used this phrase, Methodological Research Review, or some derivation of it such as Methodological Review, Research Review, etc. finds a rather wide umbrella of study goals and design. In Barnett, Docherty, and Frommelt (1991), the authors' reviewed 43 studies published since 1963 for a broad range of types of methodological flaws in a very specific topic of study, that of child and adolescent psychotherapy. Other studies are more specific about the method type they are interested in and less concerned about the substantive topic at hand. For example, Morgan (1996) investigated appropriate methods for a specific strategy of data collection: focus groups, across academic and applied research areas. Other studies may have a mix

of specificity regarding both method of interest as well as topic, or domain of interest. In DiPrete and Forristal's (1994) study, they reviewed a fairly specific family of methods, multilevel models, used within a broad yet focused area of study, sociology, over a more restrictive span of time, 10 years.

Similar to the issues about classifying this as a meta-analysis in the traditional sense, this study cannot be considered a pure methodological review either. Rather, statistical methods are being augmented within each published study to determine the potential impact of such changes in reporting.

Sample

Previously conducted social science research studies with either a direct or indirect educational implication were gathered and reviewed. Studies were drawn from a limited number of education and social science journals in order to restrict variation of research rigor that may be influenced by publication source as well as targeted audience. Obviously, within the journals selected, there was the influence of publication source; however by limiting the number of journals used in this study, it is hoped that this publication bias was minimized. Additionally, consideration must be given to the idea of disciplinary norms. Disciplinary norms address the differences in which professionals within various different disciplines communicate, including conventions regarding the conduct and reporting of their research. As such the sampling for this research addressed research contained within the broad umbrella of specific disciplines within *Social Sciences*. Although articles selected for inclusion were required to have either a direct or indirect educationally-oriented focus, at least a portion of the journals in the sample were

written for audiences that included not only educators but also psychologists and other social scientists.

The specific number of studies from each journal varied slightly, due to differences in the frequency of publication and number of articles per publication; however, the goal of a minimum of ten studies to be extracted from each journal was met (see Table 6). Additionally, in order to attain a representative sample of current research reporting practices, only studies that were published within a five-year time frame were considered for inclusion (July 1998-June 2003).

Selection of Journals

Considerations leading toward journal selection included a review of journals sponsored by professional organizations such as the American Psychological Association, the National Council of Teacher's of Mathematics, and the American Educational Research Association. Characteristics of the types of studies was of key importance as many of those reviewed were primarily methodologically based, e.g., the *Review of Educational Research*, or possessed a majority of studies that were not of a nature conducive to inclusion such as those using many qualitative types of studies, e.g., the *Journal of Research in Mathematic Education*. Other considerations for selection included whether or not journals utilized a peer-review process as well as their longevity in the field. A final consideration was frequency of use and consultation of the selected journals as evidenced by their availability in libraries and frequency of citations by other journals. These criteria have been identified to maintain some degree of similarity both in

expected research rigor, as well as exposure to more recent advances in research methods and philosophy.

Based on a preliminary review of journals currently in use in the social sciences, three journals were identified as the primary sources for studies to be reviewed. After a preliminary screening of the recent five-year collection of studies within each journal, it was determined that a sufficient number of studies were available with the required data within each of the three journals. The journals included in this study as the sources of research studies analyzed are: (1) *Reading Research Quarterly*, (2) *Journal of Educational Research*, and (3) *Journal of Personality and Social Psychology*. These three were selected after a review of journals used in the social sciences and consultation with individuals familiar with research-based professional journals using the criteria and considerations previously discussed. All three have a national or international research audience and contain empirically-based research with educational consequences. Additionally, the three represent journals that have audiences that vary in scope. The first, *Reading Research Quarterly*, the flagship journal of the International Reading Association, is of primary interest to educators with a focus on literacy issues. The *Journal of Educational Research* has a more broad scope of audience, including educators of various academic disciplines as well as roles, e.g., administrators. The final journal, the *Journal of Personality and Social Psychology* reaches beyond the educational community and encompasses the entirety of social science professionals.

The difference in aspects of disciplinary norms associated with the different primary target audiences of these journals must be taken into consideration. While there

is some concern that the research contained in these journals may contain differences regarding type of knowledge as well as the technical depth of the research, the fact that all three journals fall within the realm of *Social Science* research is likely to minimize the impact of such differences. To some degree, the audience of the smaller scope journal might include those readers of the other two, and the audience of the *Journal of Educational Research* might include readers of the third; however, this is not a reciprocal relationship. This difference in scope may be of potential importance regarding the impact of research regarding the rigor and reporting methods relative to the type and size of the intended audience.

All three journals are disseminated worldwide and were thus readily accessible. Table 1 contains a brief profile of each journal regarding the source and frequency of publication, as well as a summary of the number of libraries currently subscribing to each journal (University of South Florida Virtual Library, n.d.). This table illustrates the diversity of the types of journals contained within the broad context of educational research, not only in scope of topic but also in sponsoring organization and frequency of publication.

A review of the Journal Citation Reports—Social Sciences Edition (Institute for Scientific Information, 2002) indicated varying degrees of strength of use as evidenced by the frequency of citations in other journals (Table 2). The Impact Score is intended to provide an indication of a journal's relative importance to the field and is calculated by dividing the number of citations during a given year, in this case 2001, by the number of articles published during the preceding two years (1999 and 2000). The Immediacy

Score, a measure intended to provide an indication of how timely the journal is cited, is calculated by dividing the number of citations of the journal in a given year from articles that were published in that same year. The *Journal of Educational Research* (JER), for example, was cited 29 times in articles published in 1999 and 2000. During that time (1999 and 2000) JER published 71 articles. To calculate the Impact Factor, we divide 29 by 71, which provides the ratio 0.408. Likewise, 1 article was cited in 2001 from the 29 published during that year, resulting in an Immediacy Index of 0. 034.

Table 1.

Profile of Journals

Journal Name	Sponsoring Organization	Frequency of Publication	Number of Libraries Subscribing
Journal of Personality and Social Psychology	American Psychological Association	Monthly	1683
Journal of Educational Research	Heldref Publications	Bi-Monthly	1661
Reading Research Quarterly	International Reading Organization	Quarterly	1190

The differences in number of citations and the other indices is not, for the purposes of this study, considered problematic due to the substantive differences in the target audience of each journal as well as the differences in the frequency of publication.

Table 2.

Citation Scores and Rankings Compared to All Social Science Journals

Journal Name	Impact Score (rank)	Immediacy Score (rank)	Number of Citations in 2001 (rank)
Journal of Personality and Social Psychology	3.61 (24)	0.48 (142)	23,565 (3)
Journal of Educational Research	0.41 (1075)	0.034 (1142)	395 (606)
Reading Research Quarterly	1.87 (139)	0.15 (560)	922 (280)

Total number of journals in Social Sciences Journal Citation Report = 1682

The journals were ranked relative to the entire body of social science journals as well as to those found in their specific discipline. Although the journals were ranked at widely disparate levels when considering their overall rank compared to other social science journals (Table 2), the strength of their ranking was enhanced when compared to other journals in their discipline (Table 3). The only one of the three that was not in one of the first two ranks in their discipline was the *Journal of Educational Research*. However, the 47 journals preceding *JER* in the Education and Educational Research category showed a lack of fit for this study in either focus, content, or scope. Only eight of the higher ranking journals were research focused and of those, five were subject specific, e.g., *Health Education Research* (Rank: 11), and three were methodological or review oriented, e.g., R*eview of Educational Research* (Rank: 1). The highest ranked

subject specific research journal, *Reading Research Quarterly* was selected for this study as the subject specific journal. The *Journal of Educational Research* was the highest ranked research journal with a general educational focus that contained primarily empirically based research. As such, it was considered the most acceptable for use in this study when all factors were taken into consideration.

Table 3.

Journal Ranks Relative to Subject-Specific Journals

Journal Name	JCR Subject Category	Number of Journals in Category	Rank
Journal of Personality and Social Psychology	Psychology, Social	43	1
Journal of Educational Research	Education and Educational Research	92	48
Reading Research Quarterly	Education and Educational Research	92	2

Selection of Published Studies

Studies were considered for inclusion which, to the extent possible, meet the following selection criteria: (1) availability of all necessary statistical estimates to permit calculation of appropriate effect size (if the effect size is not reported in the published report) and confidence intervals, including, but not limited to means, standard deviation, and sample size, (2) studies that used the analyses of interest as a primary basis for reported results, conclusions, and recommendations, and (3) studies that were of a nature

conducive to the purposes of this research, e.g., the research is examining differences between two or more groups (t-tests or ANOVA designs) and those employing regression/correlational designs. It was determined that although it would be ideal if other key information such as reliability indices and data distribution information were included to help ascertain the soundness of a given study, it was anticipated, and was proved to be true, that this information was not available for many studies and was therefore not considered to be a requirement for inclusion. These criteria permitted a certain degree of commonality between studies selected based on design type and group similarity, thus limiting comparisons to only three general types of studies with groups that are reasonably homogeneous. Additionally, in the case of studies from the *Journal of Personality and Social Psychology*, only studies with a direct or indirect educational relevance (e.g., studies on the attention span of children or other behavior that could have impact in a classroom) were considered to maintain an educational focus.

The selection process of the final sample had multiple stages. Once the journals were identified, all studies within the three journals covering the time span of interest (July 1998-June 2003) were scanned to determine if the types of analyses included and statistics reported warranted consideration for inclusion. Additionally, the topic of each article was considered relative to the direct or indirect relationship to educational issues. From this initial review, 79 articles were selected as potential studies to include for the study. Each of these were then reviewed more in-depth to determine the level of data available. That is, were standard deviations, group sizes and other critical information clearly reported relative to the analysis employed? At this point, the context of how the

analyses that were to be addressed in this study were being employed in the article was considered to determine the impact the analyses had on the overall findings and purpose of the study. For example, some studies might only have used t-tests to examine preexisting differences between groups without any significant or direct impact on the goal of the study. The final sample of articles and the types of analyses represented within articles (N=33), by journal and analysis type, is provided in Table 4.

Table 4.

Types of Analyses Included in Number of Articles

	Journal of Personality and Social Psychology	*Journal of Educational Research*	*Reading Research Quarterly*	Total
Two Group Comparisons (t-tests)	4	7	1	12
More than Two Group Comparisons (ANOVA)	9	4	9	22
Regression Analyses	1	0	3	4

Note: In some cases, studies used more than one analyses of interest, thus the different total than that reported in the text.

The types of analyses used in different articles was fairly diverse when considering the number within a specific journal as well as across journals. For example, ANOVA applications tended to dominate the literature with 22 articles using this type of analysis.

Comparatively, only four studies incorporated regression analyses with two group comparison using t-tests falling almost halfway between these two extremes, used in 22 studies.

Computations

Using the reported information, the following statistics were computed, if not already reported in the published study:

1. Test of statistical significance (t-values, etc), including associated p-value.
2. Confidence interval for the statistic of interest. For studies comparing differences between two groups, the CI for the difference of means were constructed. For studies comparing differences between more than two groups, e.g., in an ANOVA context, CIs were constructed around η^2, a measure of degree of variance attributable to group membership. For studies examining a correlational relationship, the CI around the squared multiple correlation coefficient, R^2, a measure of explained variance, was constructed.
3. Statistic of practical significance. Depending on the study design and analysis, one of three effect sizes were computed.

   a. For studies comparing differences between two groups, Cohen's d was used, given by:

$$d = \frac{\overline{X_1} - \overline{X_2}}{\hat{\sigma}_p}$$

   where $\overline{X_1}$, $\overline{X_2}$ are the means of the two groups and $\hat{\sigma}_p$ is the pooled standard deviation.

b. For studies that are comparing more than two groups, e.g., ANOVA analyses, Cohen's *f* effect size were computed, given by:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$$

c. For studies that examine a correlational relationship, e.g., those using a regression analyses, Cohen's signal-to-noise ratio, f^2, was used, given by:

$$f^2 = \frac{R^2}{1-R^2}$$

4. Confidence intervals for the statistic of practical significance were constructed using the Normal Z transformation and the Steiger and Fouladi interval inversion method.

Confidence Intervals

Confidence intervals were constructed using Type I error rates of 0.01, 0.05, and 0.10, using both the normal Z-transformation as well as the Steiger and Fouladi interval inversion method. Based on previous studies (Hess & Kromrey, 2003 and Kromrey & Hess 2002), it was anticipated that the results of these two methods would not differ to a substantial degree, an expectation that was fulfilled. The only issue relative to CI construction was limited to a very small portion of the studies analyzed. In a this small portion of cases (less than 2%) the values were so extreme (due primarily to inordinately large sample sizes combined with either very large or very small effect sizes) that the

Steiger and Fouladi interval inversion method would not function due to the limitations of the SAS software system on probability computations in the extreme tails of the t and F distribution. However, in all cases, the other calculations used (e.g., student's t, Fisher z-transformation or z distribution, as appropriate) were used if necessary.

The width of the intervals were then examined at each of the three levels for general distributional characteristics. To the extent possible, studies used were analyzed with consideration given to the strength of the study design as well as variables considered and types of related information reported (e.g., was there specific mention of the type I error rate that significance tests were conducted at). The strength of the conclusions that could be drawn using a confidence band instead of a point estimate were examined and discussed. All computational aspects of the analysis were conducted using SAS version 8.2 run on the Windows XP operational system. The data were then imported into Microsoft Excel for the purposes of constructing visual displays of the findings in tables and figures.

Data Analysis.

The selected studies were coded to collect information on the characteristics of the study such as distributional information, impact of missing data, etc. as well as the statistics reported, e.g., ANOVA F values, Regressions R^2 values (Appendix A). The purpose of the coding was not to report a clear measure of study strength or rigor, rather it was intended to gather relevant information about the study as well as provide a sense of the type of information typically reported.

Effect sizes were calculated, regardless of whether they had been reported based on the data provided by the author(s), e.g., reported means, sample sizes, degree of variability. This computation external to the study was necessary to preclude the potential of the author(s) using an effect size calculation other than the three identified for this study. For the purposes of this study, effect size magnitudes were classified using Cohen's criteria (Cohen, 1988) without attention to contextual issues. Table 5 contains a summary of the three effect sizes and Cohen's classification, albeit reluctant, as small, medium, or large.

Table 5.

Effect Sizes and Associated Interpretation

	Effect Size Index		
	Cohen's *d*	Cohen's *f*	Cohen's f^2
Small Effect	0.20	0.10	0.02
Medium Effect	0.50	0.25	0.15
Large Effect	0.80	0.40	0.35

The consideration of context when interpreting effect sizes is vital for applied purposes; however, this is not a direct consideration in this study and will therefore not be included.

Confidence band widths were calculated at three Type I error rates: .01, .05 and .10 using appropriate techniques for the analysis of interest. Confidence intervals for comparisons of two-groups were constructed using the Student's t distribution for the

differences between means and the z-distribution as well as the Steiger and Fouladi interval inversion method for the Cohen's d measures (see Hess & Kromrey, 2002 for details). Similar approaches were used for Regression and ANOVA analyses, using a logarithmic transformation of Z, similar to the Fisher transformation, as well as the Steiger and Fouladi interval inversion approach. Details of the effectiveness of these techniques can be found in Hess and Kromrey (2002) and Kromrey and Hess (2000). Intervals were examined to determine if there were noticeable differences in the research rigor found in different journals or in the impact of precision based on the type of study and method of analysis chosen.

Reporting Results and Conclusions

The discussion sections of the published studies were reviewed to determine if findings or conclusions might have been affected or altered by different reporting practices. Specific discussions and statements relative to the statistical analysis conducted were culled from the study and reviewed with the intent to determine if additional information, e.g., effect sizes and/or confidence intervals, should have impacted the strength of the wording used in results and conclusions. A determination was made if inclusion of effect sizes and/or confidence intervals would:

1. have no impact on how the results and conclusions were reported, that is, No changes needed.
2. have some impact on how the results and conclusions were reported, that is, slight changes needed.

3. have substantial impact on how the results and conclusions were reported, that is, drastic changes needed.
4. have a major impact on how the results and conclusions were reported, that is, a complete revision required.

A copy of the instrument used for this determination as well as a sample study and analysis summary is included in Appendix C. A total of 42 analyses or sets of analyses were extracted from the 33 studies for this portion of the study. These analyses or sets of analyses were identified upon review of the results and conclusions provided. If a statement was clearly based on a single analysis, then the statistics associated with that analyses were used. If a statement was based on a group of analysis, then they were reviewed conjointly. This typically happened when an ANOVA test was conducted with follow-up t-tests. A large majority of the analyses conducted within the broad scope of this research did not lend themselves to inclusion in this part of the study. The reasons for this varied, with the most dominant reason being that although results of statistical significance tests might have been reported numerically either in the text or a table, the impact of these specific analyses were not uniquely identifiable within the results and/or discussion of the results. Multiple t-tests may have been run for a written conclusion within a larger context. Other examples that were not investigated relative to interpretation aspects of the study included those analyses that were run for preexisting differences (typically not a focus of results or discussions of implications of findings) and those that addressed non-focal points of the study, e.g., analyses of demographic data that were not addressed relative to conclusions or impact.

Reliability of Interpretative Results

Twenty of the analyses or sets of analyses were independently reviewed by measurement specialists well versed in educational research to determine if the decisions reached by this researcher would likely to be representative of members of the research world in general. One of the twenty analyses had to be discarded due to a problem noted in the summary information provided to the reviewers. Thus the percent agreement was based on 19 analyses or sets of analyses. This was not considered to be a major problem as 43.54% of the sample was used as a basis for verification and a measure of reliability of this researcher's recommendations for change.

Prior to the independent reviews, the researcher coded all the analyses (or sets of analyses) using the '1' (*No Change Needed*) to '4' (*Complete Revision Needed*) scale described previously. The analyses (or sets of analyses) coded by the independent reviewers were selected to be representative of the 42 used in the analysis. The subset of analyses used for this reliability check included analyses from all three studied in this research (t-tests, ANOVAs, and Regression) as well as analyses (or sets of analyses) from all three journals selected. Additionally, the subset included analyses that had been determined by the researcher to need varying degrees of interpretative adjustment when effect size and confidence interval information was included. That is, a range of analyses were provided to the independent coders, previously rated by the researcher as needing *No Change, Slight Change, Much Change,* or *Complete Revision*.

Once the subset of analyses had been selected, the researcher conducted a training session with the reviewers. Each of the reviewers were provided with an instruction

sheet, coding sheet for each analysis (or set of analyses), and a summary of each of the studies that they were to review with the appropriate statistics (see Appendix C). The researcher read the instructions aloud while the reviewers read the instruction sheet. The reviewers were given the opportunity to ask questions and provide input. At that point, one analysis was reviewed and coded independently by each individual and the results discussed among the group. There were some initial differences in how much to consider information such as study strength (some reviewers had taken sample size, deducted from degrees of freedom information) into consideration of their ratings. They were instructed to concentrate primarily on the statistics themselves and not take into consideration other elements of the study. After the training, practice, and discussion, the reviewers were given all their materials to conduct the rest of their reviews independently. Coding sheets were then returned to the researcher (one reviewer emailed their results) and ratings were input into an Excel spreadsheet.

The decisions reached by these independent reviewers were then compared to those reached by this researcher and the percent agreement, both by item and overall, was computed. In general, agreement was strong. Overall agreement was 83% with the highest agreement resulting from the impact of confidence intervals on the degree to which results and conclusions might be affected (89%). Interestingly, the lowest agreement (79%) was the degree to which the results and conclusions might be altered based on the results of the significance tests conducted, and reported, within the original study. The percent agreement regarding the degree to which reporting effect sizes might impact revisions of results and conclusions was in between the other two at 82%.

Recommendations for Reporting Research Results

Finally, the results of this study were considered holistically to provide recommendations for reporting research results. The use of illustrations from actual results is anticipated to provide yet another piece of justification for researchers to more thoroughly report their findings and for editors of journals to demand such reporting. Just as educators in the field are being held accountable for their methods, so should the methods and work of educational researchers, including their reporting practices and protocols.

Chapter Four

Results

The purpose of this research was to examine the potential impact of different methods of reporting research results on the conclusions that could, and should, be made from these findings. Specifically, this study investigated how the use of practical significance as measured by effect sizes in addition to measures of statistical significance might impact the degree to which one should interpret results. Additionally, the use of confidence intervals around point estimates was examined in order to determine the precision of measurements obtained in studies and how that degree of precision might impact conclusion drawn from findings.

Previously conducted research deemed worthy of publication that contained one of three rather traditional and oft-used statistical analyses, t-tests, Analysis of Variance (ANOVA), and/or regression were reviewed and results reanalyzed using not only the significance test results provided in the study, but also using the appropriate measures of practical significance (Cohen's *d*, Cohen's *f*, and Cohen's f^2 respectively). Further, confidence intervals for all point estimates, including measures of statistical as well as practical significance were constructed. Results and conclusions relative to specific statistical analyses were then examined with consideration given to the additional information provided by the calculated effect size and confidence intervals. The degree

to which the results and conclusions that were presented might be adjusted or reconsidered was estimated.

The three questions investigated in this research were:

1.) To what extent does reporting outcomes of tests of statistical significance vs. tests of practical significance result in different conclusions and/or strengths of inference to be drawn from the results of research?
2.) To what extent does reporting confidence intervals instead of, or in addition to, point estimates affect the conclusions and inferences to be drawn from the results of research?
3.) What method, or combination of methods, is recommended for reporting results in educational studies?

Characteristics of Selected Studies

For the most part, researchers did not report either effect sizes or confidence intervals in their results. Only one article of the 79 studies considered for final inclusion during the screening steps of study selection reported results of significance tests, effect sizes and confidence intervals (Baumann, Edwards, Font, Terehinski, Kameenui, & Olejnik, 2000). No other studies reviewed reported confidence intervals and few reported effect sizes and none did so consistently. Of the final sample of 33 articles, 393 ANOVA analyses, 108 regression analyses, and 149 t-test analyses were reviewed. The types of analyses within specific articles as well as different journals varied widely (see Table 6 for specifics). For example, the *Journal of Educational Research* tended to have fewer

analyses within a given study and reported more analyses of two-group comparisons than the other two journals. ANOVA applications seemed to dominate studies in both *Reading Research Quarterly* as well as the *Journal of Personality and Social Psychology.*

During the initial screening, numerous articles were excluded from inclusion due to nonreporting of statistics required for this study such as sample size or standard deviation. For example, two group comparisons using t-test analyses were evident in the *Journal of Personality and Social Psychology* a little more often than is obvious in this study; however, there tended to be a dearth of sufficient information to permit inclusion of those studies within this study. It was possible, in limited cases, to derive some of that information from other data provided, e.g., degrees of freedom, but this was only done in limited situations where the derived information could be safely relied on.

The contribution of regression analyses to this study was limited. Only four articles were found that contained appropriate information to include in this analysis. In many cases, studies that had regression applications reported weights and coefficients only, with no indication of explained variances. Of the four regression studies, two had results that do not seem typical of regression analyses in general and thus may be responsible for the distribution of the results to be highly skewed toward very large effect sizes. For example, Sutton and Soderstrom (1999), reported R^2 values that were atypically large, e.g., 0.80, 0.76.

Not all of the analyses contained within the 33 studies were considered as appropriate to include in the interpretation of results and conclusions part of this study, as they examined such things as preexisting differences between groups,

Table 6.

Types of Analyses Reviewed by Article Number and Journal

Article No.	Journal	t-Test	ANOVA	Reg	TOTAL
1	JER	12			12
2	JER			22	22
3	JER	33			33
4	JER	10		8	18
5	JER	10			10
6	JER			6	6
7	JER		14		14
8	JER	1			1
9	JER		4		4
10	RRQ		26	1	27
11	RRQ		45		45
12	RRQ		3		3
13	RRQ		6		6
14	RRQ		38		38
15	RRQ		1		1
16	RRQ	3			3
17	RRQ			38	38
18	JPSP	58	38		96
19	JPSP		32		32
20	JPSP		15		15
21	JPSP		9		9
22	JPSP			21	21
23	JPSP	5	4		9
24	JPSP		6		6
25	JPSP	3	27		30
26	JPSP		20		20
27	JPSP	6	11		17
28	JER	2			2
29	JER		25		25
30	JER	6	3		9
31	RRQ		11		11
32	RRQ		32	12	44
33	RRQ		23		23
	Total	**149**	**393**	**108**	**640**

provided evidence of known differences, or were not an evident or specific contributor to the results and conclusions discussed. These analyses were included when examining the general behavior of the statistics as a function of study

Regardless of the type of analyses conducted, the general distribution of effect sizes revealed extremes at either end, with most effect sizes spanning Cohen's small to large range (Figures 5 and 6) for group comparison studies (ANOVAs and t-tests).

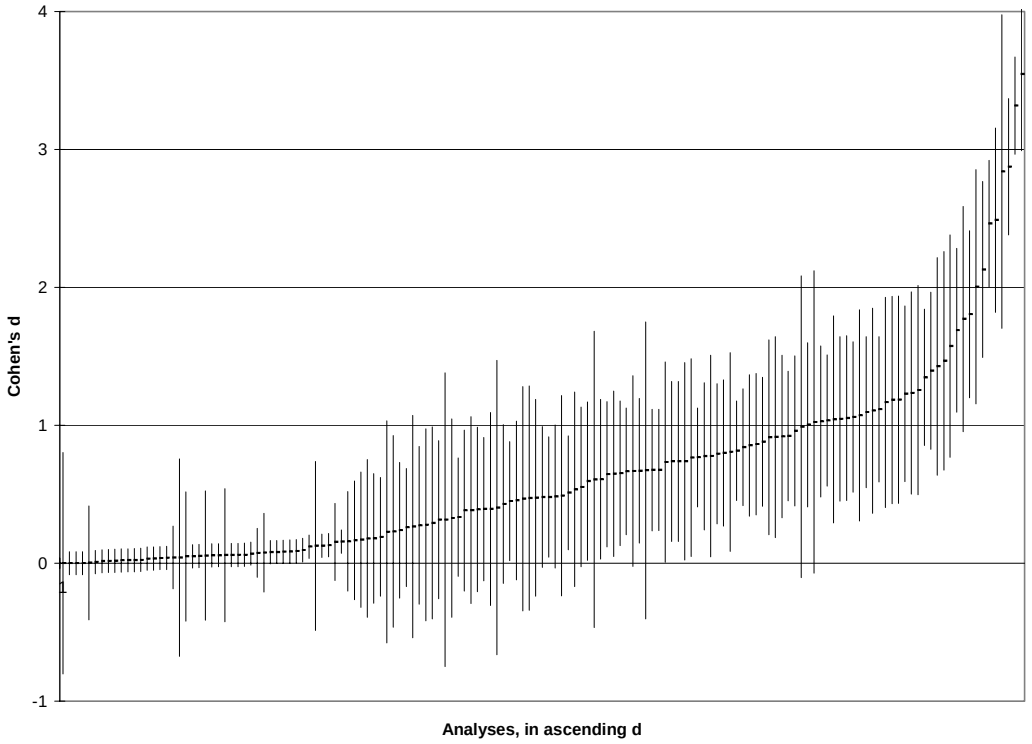

Figure 5. Distribution of effect sizes and 95% confidence intervals for all t-test Analyses pooled across journals as effect size increases.

The studies with extreme values were further examined and found to primarily reflect unique comparisons that, upon review, seemed to provide understandable conditions for

the extremeness of the result. For example, many of the large effect sizes in the ANOVA applications came from one study that examined differences in text composition in different literary genre. The only exception was the distribution of the results of Cohen's f^2 (Figure 7 and Figure 10) which shows a tendency toward rather large effect sizes. This may be due, at least in part, to the limited availability of regression-based studies available for inclusion in this study (n = 4).

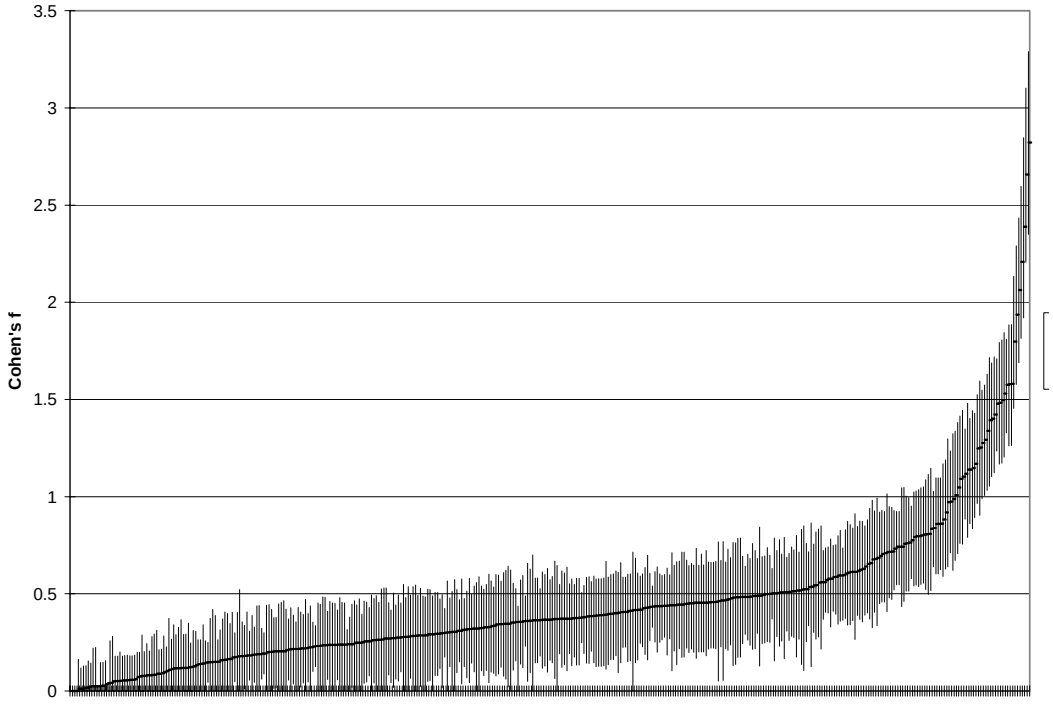

Figure 6. Distribution of effect sizes and 95% confidence intervals for all ANOVA Analyses pooled across journals as effect size increases.

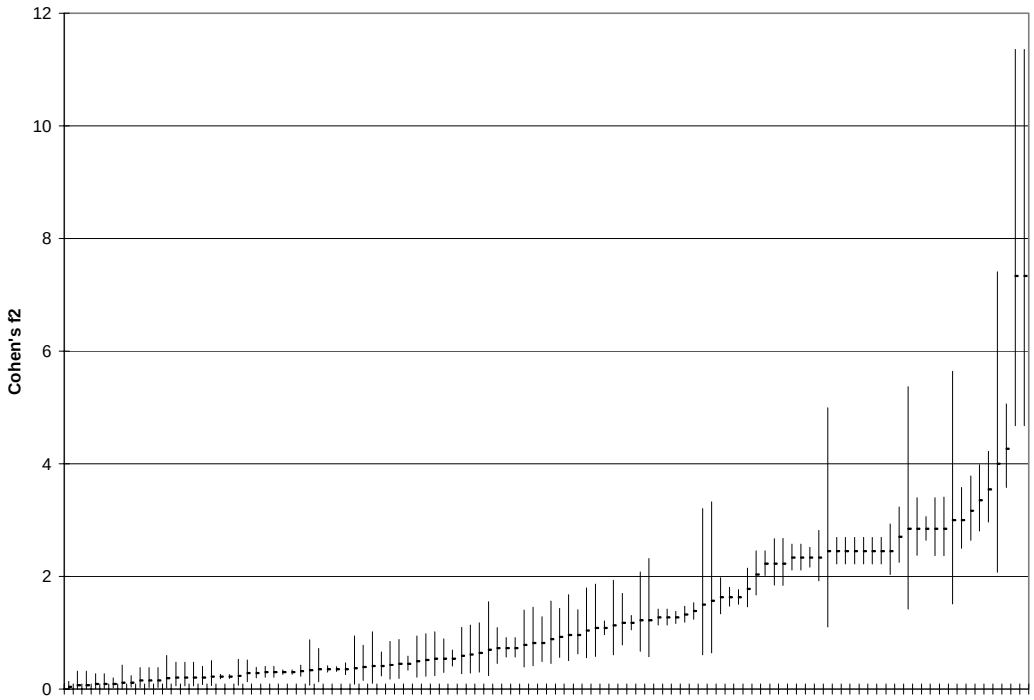

Figure 7. Distribution of effect sizes and 95% confidence intervals for all Regression analyses pooled across journals as effect size increases.

Additionally, the distribution of effect sizes was relatively similar across journals (see Figures 8, 9, and 10), although the frequency of different types of analyses varied from journal to journal. Although the number of published studies that contain t-tests was largest in the *Journal of Educational Research*, the actual number of t-tests conducted within those studies was largest within the *Journal of Personality and Social Psychology*.

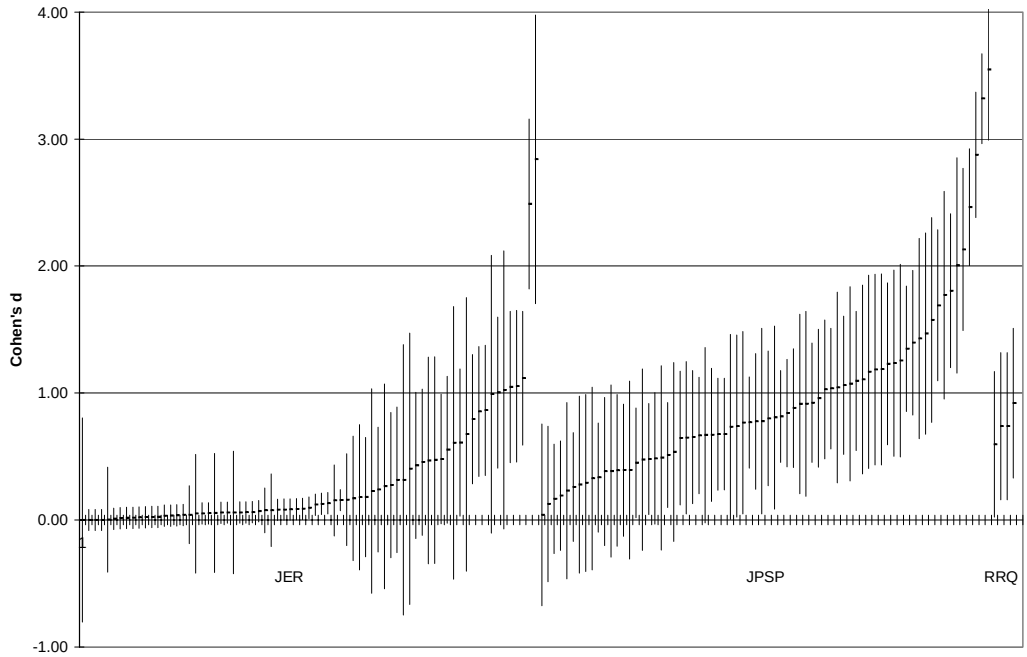

Figure 8. Distribution of effect sizes and 95% confidence intervals for all t-test analyses as effect size increases by journal type.

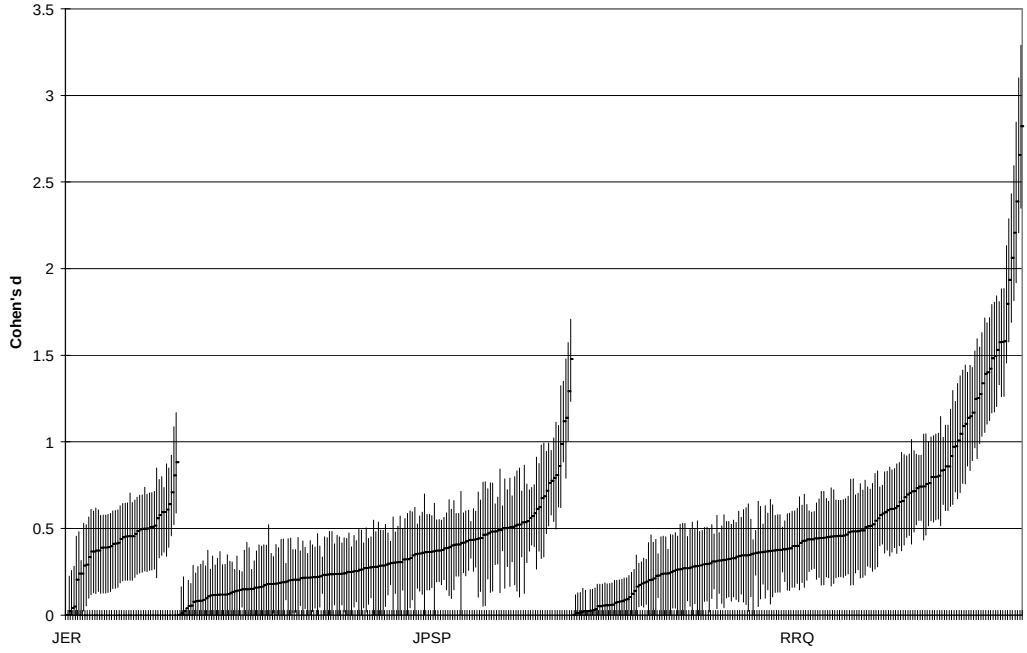

Figure 9. Distribution of effect sizes and 95% confidence intervals for all ANOVA analyses as effect size increases by journal type.

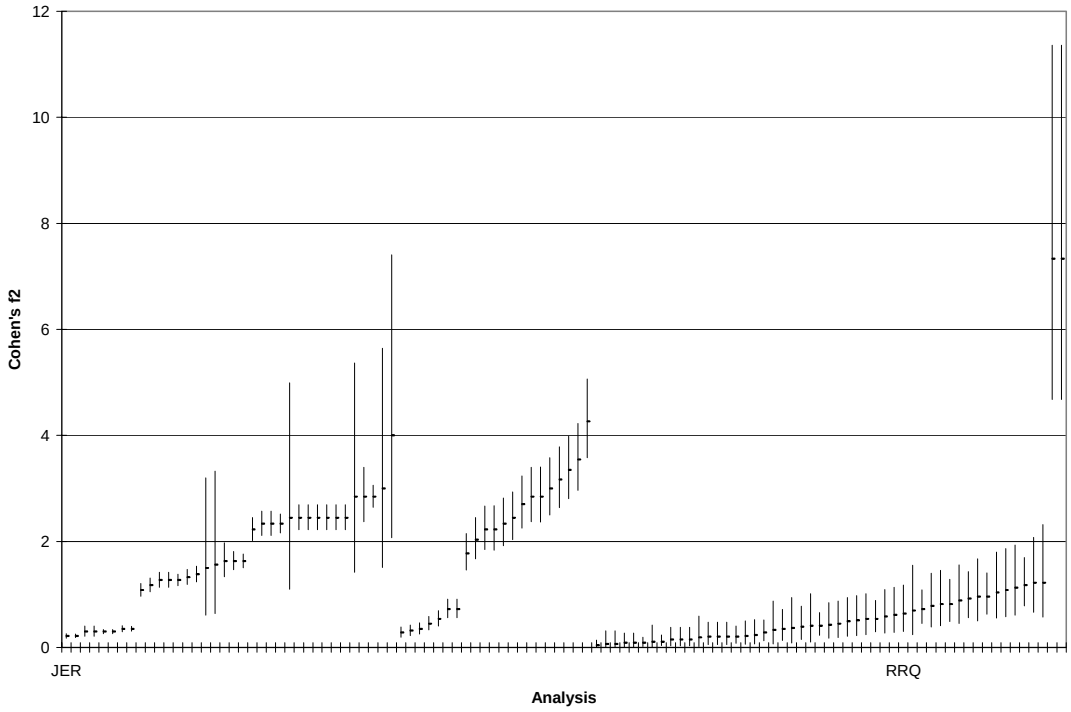

Figure 10. Distribution of effect sizes and 95% confidence intervals for all Regression analyses as effect size increases by journal.

Relative to the Type I error rate of interest, the bandwidth noticeably increases as alpha decreases as would be expected. Figures 11, 12, and 13 provide an illustration of this using the results of the ANOVA analyses.

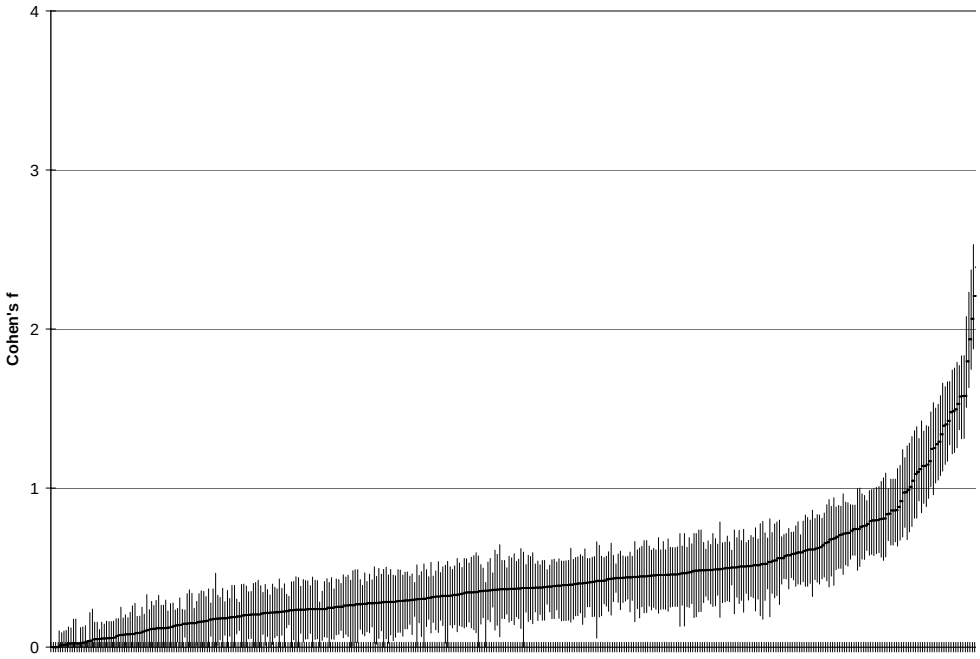

Figure 11. Distribution of effect sizes and 90% confidence intervals for all ANOVA analyses pooled across journals as effect size increases.

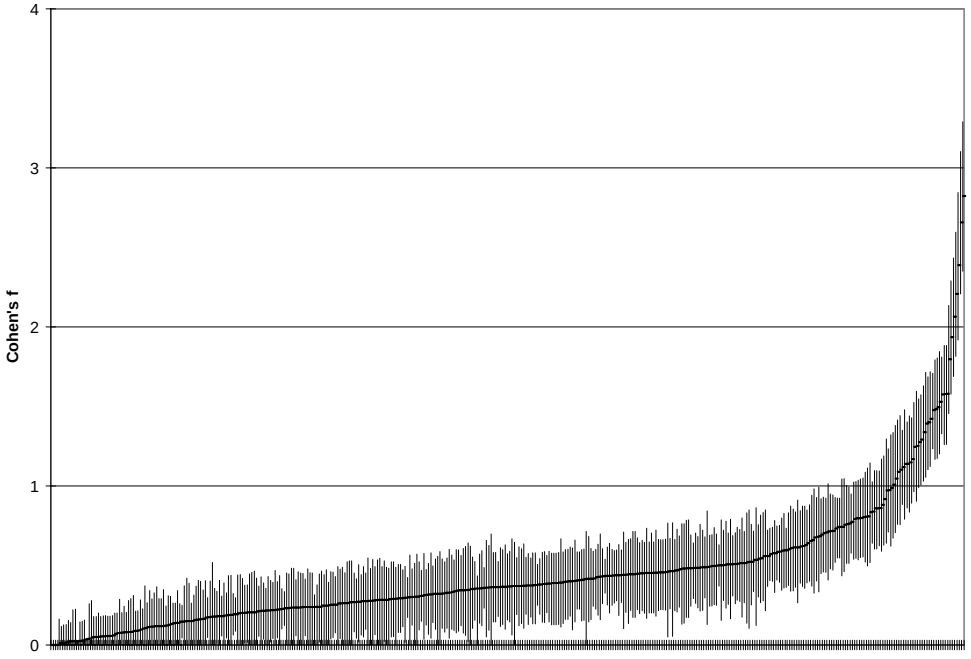

Figure 12. Distribution of effect sizes and 95% confidence intervals for all ANOVA Analyses pooled across journals as effect size increases.

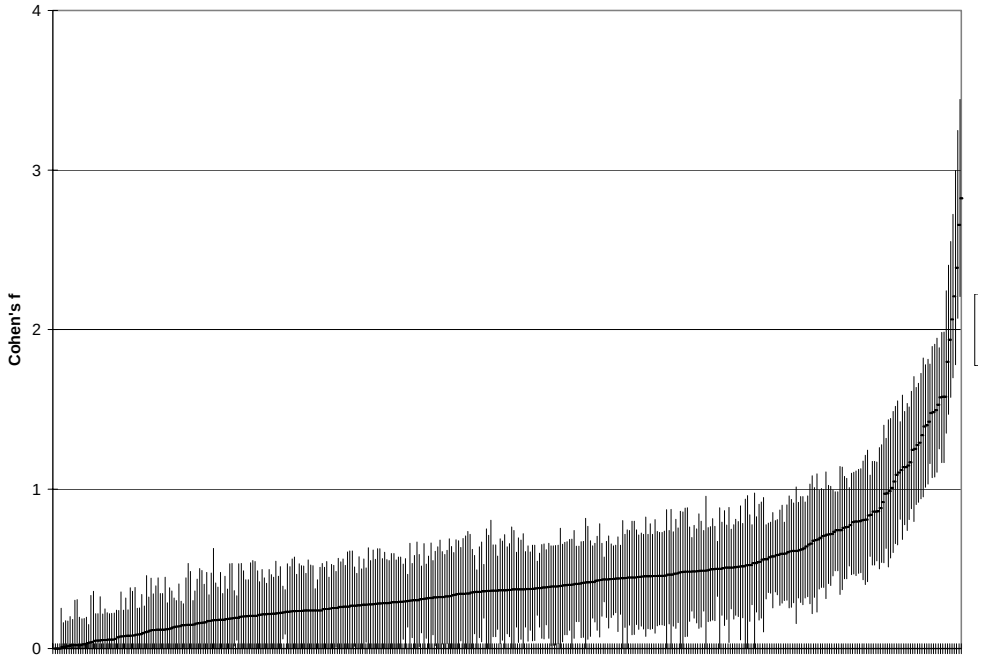

Figure 13. Distribution of effect sizes and 99% confidence intervals for all ANOVA Analyses pooled across journals as effect size increases.

Sample size, as one might expect, had a notable impact on the results of bandwidth. In Figure 14, bandwidths for ANOVA analyses are illustrated as a function of increasing total sample size for the three type I error rates examined.

A similar trend was noted for studies using t-tests and Regression analyses. Additionally, as the ratio of sample size to the number of groups in ANOVA studies increased (that is, increased average sample size within each group), bandwidths also tended to decrease (Figure 15)

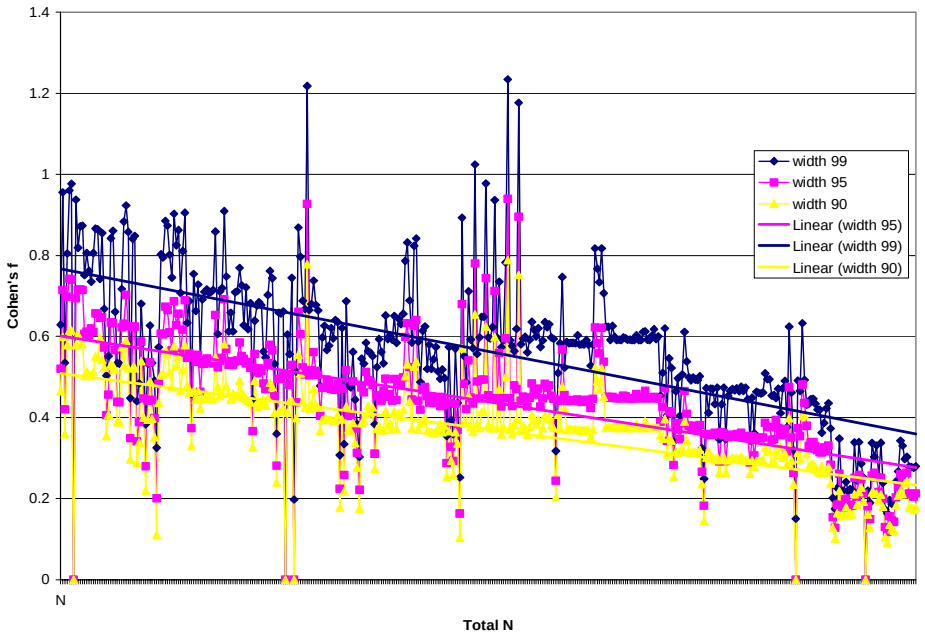

Figure 14. Bandwidth of Cohen's *f* pooled across journals as total sample size increases for Type I error rates of .01, .05, and .10.

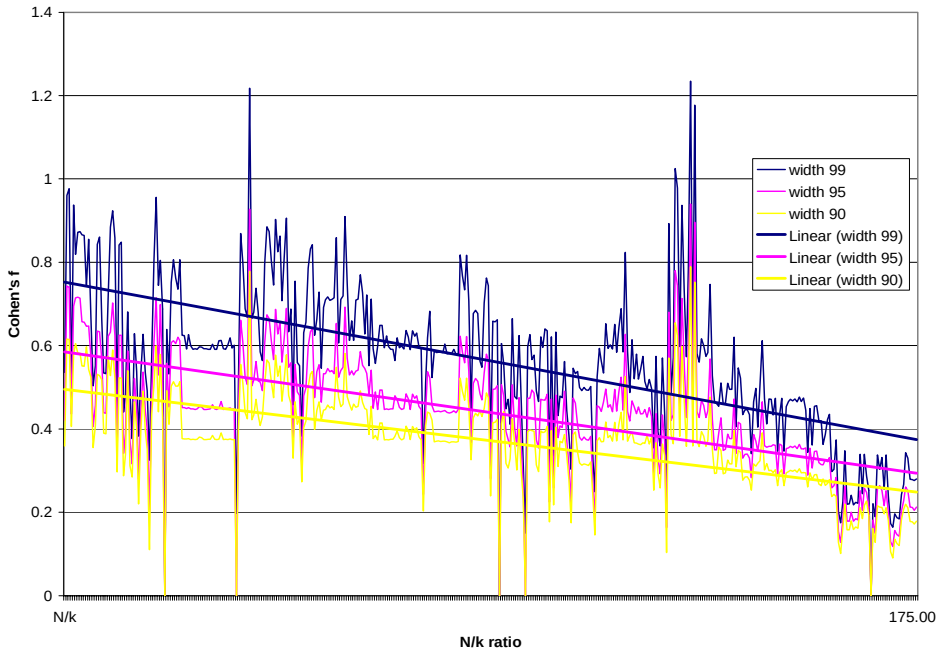

Figure 15. Bandwidth of Cohen's *f* pooled across journals as the ratio of total sample size/number of groups increases for Type I error rates of .01, .05, and .10.

Additionally, there was a notable lack of what some might consider basic but critical information regarding a research study. Noticeably lacking in most studies, were measures of, and information regarding, reliability and validity, distributional characteristics of the data (including presence or absence of outliers), missing data, dependence/independence of observations, etc. The propensity to leave this type of information out was alarming.

Statistical Significance vs. Practical Significance

Of the 640 individual analyses used in this study, the degree to which they were reported as statistically significant varied as a function of the type of analysis conducted (see Table 7). There were a total of 149 two group comparisons that used t-tests as their analysis of choice. Of those 149, slight less than half (n=70), 47%, reported statistically significant findings. Contrast this to the reported regression analyses, 88% (n = 95) of which reported statistically significant findings and the ANOVA analyses, 81% (n = 319) reporting significant findings. Although not explicitly stated in virtually all studies examined, it seemed evident that most, if not all, significance testing was done using a Type I error rate of 0.05. This inference is made based on the fact that most findings that were not contained in a table and asterisked (*) to imply various levels significance (a fairly common, and lamentable, practice in reporting research results) were reported using notation such as '$p<.xx$' and in the studies reviewed, this number did not exceed 0.05. Rather the letters *ns*, implying non-statistical findings were reported. Additionally, this initial examination of effect sizes with regard to significance testing, absent of context,

revealed that in many cases, results that were reported to be statistically significant (original author's interpretations) had various degrees of practical significance (see Table 7).

Table 7.

Numbers and Percent of Analyses Reporting Statistical Significance Relative to Computed Effect Size

Type of Test	Total	No Effect (Cohen's: *d*: <.1 *f*: <.05 f^2: <.01)	Small Effect (Cohen's: *d*: .1-.34 *f*: ..05 - .16 f^2: .01-.08)	Medium Effect (Cohen's: *d*: .35-.64 *f*: .17-.32 f^2: .09-.25)	Large Effect (Cohen's: *d*: .65+ *f*: .33+ f^2: .25+)
T-Test	149	37 (24.83%)	24 (16.11%)	28 (18.79%)	60 (40.27%)
Significant	70	0 (0%)	3 (4.29%)	13 (18.57%)	54 (77.14%)
Non-significant	79	37 (46.84%)	21 (26.58%)	15 (18.99%)	6 (7.59%)
ANOVA	393	18 (4.58%)	45 (11.45%)	98 (24.94%)	232 (59.03%)
Significant	319	0 (0%)	9 (2.82%)	82 (25.71%)	228 (71.47%)
Non-significant	74	18 (24.32%)	36 (48.65%)	16 (21.62%)	4 (5.41%)
Regression	108	0 (0%)	3 (2.78%)	16 (14.81%)	89 (82.41%)
Significant	95	0 (0%)	1 (1.05%)	15 (15.79%)	79 (83.16%)
Non-significant	13	0 (0%)	2 (15.38%)	1 (7.69%)	10 (76.92%)

As might be expected from the dominating reporting of statistically significant findings in regression analyses, this type of analysis reported the greater number of large effect sizes.

The degree to which effect sizes varied based on whether or not tests showed statistical significance was further investigated.

Of the 640 analyses investigated, a total of 484 were reported to be statistically significant. The magnitude of effect sizes associated with these analyses were reviewed both as a function of journal and type of analysis. Figure 16 contains a summary of the effect sizes for the different analyses by journal. As might be expected, no statistically significant analyses reported effect sizes that indicated complete absence of effect and only a small number indicated small effects. When considering whether or not a medium or large effect size associated with significant findings varied by journal type, the *Journal of Educational Research* exhibited a greater preponderance of studies reported containing large effect sizes (73 of 78, 93.6%, studies included) as compared to either the *Journal of Personality* and *Social Psychology* (139 out of 208, 66.8%, of the analyses) or *Reading Research Quarterly* (157 out of 209, 75.1%, of the analyses).

When effect sizes of statistically significant analyses were reviewed based on type of analyses, again, as expected, there were not any instances in which no effect was present and only a limited number revealed small effects. Depending on the analysis, there were some differences regarding evidence of a large or medium effect. Regression analyses tended to have large effects with statistically significant results (95 out of 108 total). The results of statistically significant ANOVA tests revealed a little over a quarter of the analyses had medium effects or less, with a slightly smaller proportion of Regression and t-test analyses indicating medium effect sizes or less.

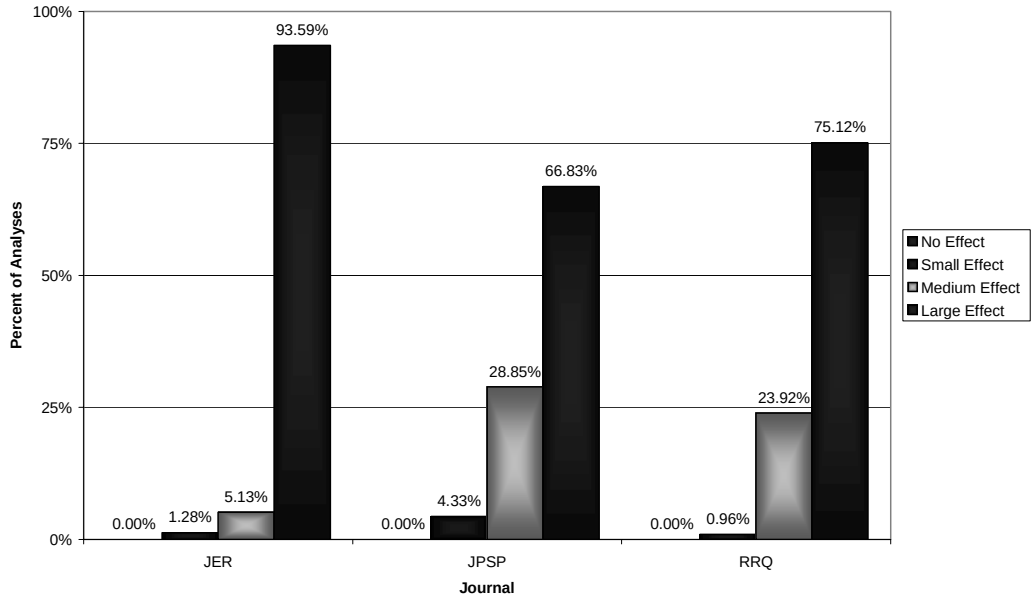

Figure 16. Effect sizes of statistically significant findings at an alpha of .05, by journal.

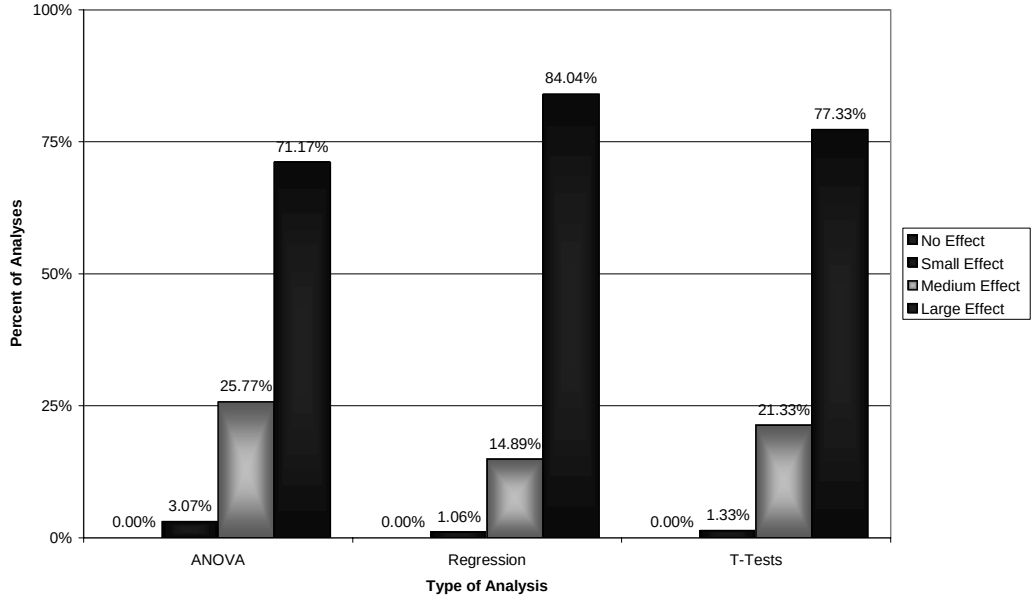

Figure 17. Effect sizes of statistically significant findings pooled across journals at an alpha of .05, by analysis type.

Results of non-statistically significant analyses were not as plentiful due to the nature of publishing preferences toward statistically significant findings. Only 166 analyses reporting non-significant findings (a little less than about one-third of that for statistically significant findings) contained enough information to calculate effect sizes. Additionally, the 166 that were available were predominantly from studies using ANOVA and/or t-tests. Only four of the non-significant findings used regression analyses. While it is not reasonable to offer a definitive explanation for this seeming disparity, it may result from the nature of the tests themselves. Multiple regression models using the same variables in various combinations often are tested and only the ones performing successfully may have been included in the final analysis. Additionally, the comparative nature of t-tests and ANOVA using multiple variables of interest might make it less likely for researchers to exclude non-significant findings when reporting significant ones.

Effect sizes of significance tests were examined as a function of analysis type for non-significant findings. Regardless of the direction, Cohen's *d* of around 0.2 or more indicates some degree of difference, so it was not considered problematic to consider evidence of effect within these analyses compared to the other two analyses considered. Figure 18 contains the results of considering point estimates for non-significant findings. It is important to note that the regression results only include four cases so the generalization of the likelihood of this distribution is very limited.

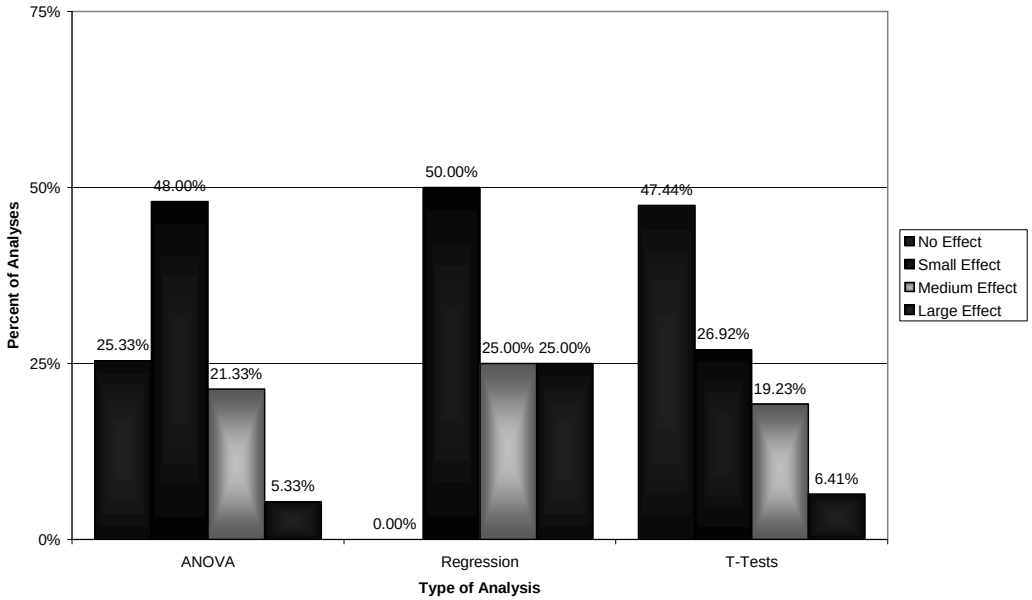

Figure 18. Effect sizes of non-statistically significant findings pooled across journals at an alpha of .05, by analysis type.

Of the other two analyses reviewed, data for 74 ANOVA tests and 79 t-tests were available. Of note in these results is the evidence of at least a small effect in most of the analyses. Over half of the t-tests indicated the presence of at least a small measure of practical difference between the two groups examined, with either a medium or large effect evident in approximately a quarter of the cases (25.64%) ANOVA had a similar proportion with medium or large effects (26.66%) and only a quarter of the analyses indicated the absence of a practical difference (25.33%). The sparse regression representatives all indicated some effect with two analyses having a small effect size, one a medium effect size, and the fourth a large effect size.

Potential Impact on Results and Conclusions

The final piece of this analysis was reviewing the results and conclusions reported that were based on the tests of statistical significance. The 42 analyses or groups of analyses included in this portion of the study were examined considering the computed effect size(s) in addition to the statistical significance tests. The results and conclusions were then determined to need varying degrees of adjustments based on the information provided by effect sizes:

1. have no impact on how the results and conclusions were reported, that is, No changes needed.
2. have some impact on how the results and conclusions were reported, that is, slight changes needed.
3. have substantial impact on how the results and conclusions were reported, that is, drastic changes needed.
4. have a major impact on how the results and conclusions were reported, that is, a complete revision required.

Only 26.19 % (n = 11) of the studies were determined to have results and conclusions that did not need any revision based on the addition of effect size information. About a quarter of the sample analyses were determined to need substantial changes (n = 12, 28.57%) with relatively few being recommended for complete revisions (n = 2, 4.76%). The largest relative proportion of studies, 40.48% (n=17), were identified as needing slight changes when the magnitude of effect size was considered in addition to tests of statistical significance.

Table 8. Number and Percent of Analyses or Sets of Analyses that Warrant Different Degrees of Change when Effect Size or Confidence Interval is Considered in Addition to Results of Statistical Significance Tests

	No Change Needed	Slight Change Needed	Much Change Needed	Complete Revision Needed
When Effect Size is Considered	11 (26.19%)	17 (40.48%)	12 (28.57%)	2 (4.76%)
When 95% Confidence Interval is Considered	3 (7.14%)	8 (19.05%)	13 (30.95%)	18 (42.86%)

These findings are fairly comparable to those found by other coders. When the results of the 19 sets of analyses reviewed by other researcher specialists, there was an adequate percent agreement with the decisions of the researcher of this study. The percent agreement when effect size was considered was 82% and the percent agreement when confidence intervals were considered was 88%.

Examples

Four analyses or sets of analyses were extracted from the sample to illustrate examples resulting in various levels of the four decisions possible, (1) *No Change Needed*, (2) *Slight Change Needed*, (3) *Much Change Needed*, and (4) *Complete Revision Needed.*

For the first example, the study investigated the degree to which college student's believed that their admission was based, at least in part, on their race/ethnicity (Brown, Charnsangavej, Keough, Newman, & Renfrow, 2000). Students were classified as members of a 'stigmatized' race/ethnicity if they were African American or Latino; Conversely, they were classified as members of a 'non-stigmatized' race/ethnicity if they were White or Asian American. The results of the statistical significance test, ANOVA, indicated the presence of a statistically significant difference: $F(1,369) = 69.89$, $p<.001$. The authors reported that:

> "When we compared stigmatized and non-stigmatized students in the degree to which they suspected that their race or ethnicity might have helped them gain admission to college, we also found a significant difference, as expected. Stigmatized students suspected that their admission to the University of Texas at Austin had been influenced by their race or ethnicity to a greater extent than did non-stigmatized students." (p. 254)

The computed effect size, Cohen $f = .4043$, tends to support the author's conclusion. As such, the inclusion of effect size is not likely to have added any further information that would have suggested different results or necessitated alterations to the conclusions drawn. The rating received by this analysis was a (1), *No Change Needed.*

In the second example, it was determined that while the stated results and conclusions were supported by consideration of the effect size in general, the effect size magnitude was sufficient to suggest slight modifications to the statement made in the conclusions. The researcher in this study (Fitzgerald, 2001) was investigating the degree

to which student's participation in a tutoring program (part time vs. full time) impacted their achievement in reading. The results of an ANOVA conducted on a measure of post-participation reading level found statistically significant differences, $F(1,76) = 4.72$, $p = .03$. The associated concluding comment by the author was:

> "There was a statistically significant treatment effect. Overall, high level treatment children outperformed low-level treatment children in instructional reading level." ($p = .45$)
>
> Cohen *f:* .2385

In general, the effect size supported the author's conclusion; however, a rating of (2), *Slight Change Needed*, was assigned due to the rather strong wording associated with what may be, at most, a medium practical effect. It would be recommended that the term 'outperformed' be replaced or conditionally qualified to slightly lessen the strength with which these findings were reported.

In many studies, the results were found to need more attention to qualifying the wording when one included effect size information in addition to statistical significance. In this example, the results were agreed with in principle but were considered to need some revamping in order to reflect appropriate strength of inference. In this study, high school student's indicated a preference for morning or afternoon academic work (Callan, 1999). These students were then randomly assigned to different groups which were administered an Algebra exam in the morning and in the afternoon. The groups contained a mix of student's with different preferences. In this set of analyses, the question being investigated was whether or not students with different time preferences (morning or

afternoon) perform differently if they take a test in the morning. Statistical significance was found between the performance of students with different preferences, $F(1,64) = 5.44$, $p<.05$. The author's concluded that,

> "There was a significant difference between afternoon-preferenced students and morning-preferenced students taking the test in the morning." (p.296)
>
> and,
>
> "The results indicate clearly that the time-of-day element in learning style may play a significant part in the instructional environment. When time preference and testing environment were matched, significant differences emerged between test results—but only for the morning test." (p. 298)

The measure of practical significance found a medium effect present, Cohen *f*: .2849. It was determined that the author's should alter the severity of strength reflected in their comments. Using words and phrases such as 'clearly indicate' and 'play a significant part' are very strong and considered not to be appropriate for the potential presence of a medium effect and are thus potentially misleading. As such, this was assigned a rating of (3) *Much Change Needed.*

Finally, there were a few studies for which inclusion of effect size tended to negate or inappropriately represent the results. That is, the results, after inclusion of effect size information were considered to be in need of complete revision. One such study addressed how different types of praise impacted childrens' judgment of their performance on tasks (Mueller & Dweck, 1998). Student's were put into three groups, one in which the children were praised for their ability (also referred to as praise for

intelligence), another in which the children were praised for effort, and a third in which no praise was provided. Based on the results of the significance tests, $F(2, 48) = 2.04$, *ns*, the authors reported that:

> "These results indicate that effort praise and intelligence praise do not lead children to judge their performance differently." (p.42)

This finding, as written, indicates a rather definitive decision about the lack of differences between the three groups of children on how harshly they judge their performance. However, when one considers the associated effect size, Cohen's $f = 0.2828$ which indicates, according to Cohen, the potential presence of at least a medium effect, the certainty with which one decides that there is no difference should be impacted. Due to the definitiveness of the statement regarding the findings of this part of the study, this example was considered to warrant a (4): *Complete Revision Needed*. The results of the practical significance indicates the possible presence of a medium effect size between the groups that should be addressed in the discussion. It would be advisable to at least discuss the possible existence of an effect and that further research into this issue might be warranted and avoid making a definite statement or judgment.

Summary

Reporting effect sizes in addition to measures of statistical significance appears to add valuable information to at least a small proportion of tests that have statistically significant results. The utility of a measure of effect appears to be enhanced when statistical tests result in non-significant findings. Over 75% of the non-statistically significant results had indications of at least a small to moderate effect. This type of

information might be valuable to researchers who believe, based on theory, previous research, or experience that a true difference does exist, however other factors might have impacted significance findings (e.g., research design, rigor).

Point Estimates vs. Confidence Intervals

The use of confidence intervals tended to be scant in the literature. Only one article was found during the initial review of journals that possessed information on confidence intervals. However, when confidence intervals were constructed around the statistics of interest in this study, including effect sizes, it became fairly obvious that they added an important element of information regarding the strength with which one should rely on the findings. Figure 19 contains a summary of the percent of analyses that had lower limit and upper limit effect sizes (using a 95% confidence band) of either no effect, little effect, medium effect, or large effect, as defined by Cohen (Cohen, 1988). This bar chart provides representation of the proportion of confidence bands, by analysis type, that contained varying levels of effect size. The left half of the chart shows the percent of analyses that had a lower band limit that had a magnitude that indicated no effect, little effect, moderate effect or large effect. The right half of the chart shows the percent of analyses that had an upper band limit with a magnitude indicating no effect, little effect, moderate effect or large effect.

With the exception of the regression analyses, confidence bands tended to include effect sizes of little or no effect in a substantial amount of the analyses (39% for ANOVA analyses and 43% for t-test analyses). Only 12 % of t-tests contained a large effect for both the lower and upper limits. Consideration of these confidence intervals leads to

clear evidence of a lack of precision in many of these studies. For example, in at least 15% of the ANOVA analyses, the lower band included effect sizes indicating lack of any effect and 28% contained small effect sizes. As such, in at least 15% of the ANOVA based studies found to be statistically significant, one cannot determine with certainty that there is a true difference between the groups of interest. Additionally, only 57% of those found to be statistically significant at an alpha of .05 had confidence bands that included only medium to large effects.

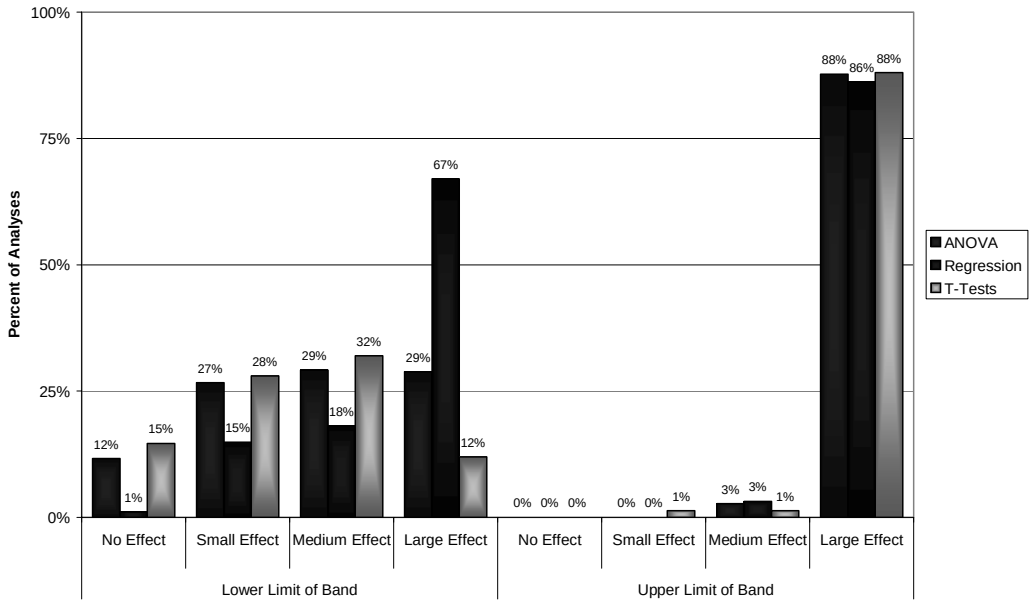

Figure 19. Percent of effect sizes of 95% confidence band endpoints pooled across journals found in statistically significant analyses.

Using a more stringent Type I error rate, e.g., an alpha of 0.01 further dilutes the ability to determine if there is a substantiated finding in the research such as a true difference between groups or impact of a treatment. For example, when 99% confidence intervals were constructed around effect sizes, the percent of ANOVA analyses that

included effects sizes that indicated no-effect jumped to 19 % (n=74) of the analyses and bands containing small effects (not including those that had lack of any effect present) went to 152 (39%). Thus, less than half (42%) of the statistically significant analyses could say with any degree of confidence at an alpha level of .01 that the findings were indicative of a medium or large effect.

Potential Impact on Results and Conclusions

The final piece of this analysis was reviewing the results and conclusions reported that were based on the tests of statistical significance. The 42 analyses or groups of analyses included in this portion of the study were examined considering the computed confidence intervals around effect sizes in addition to effect size(s) and statistical significance tests. The results and conclusions were then reviewed to determine the possible need for varying degrees of adjustments based on the information provided by effect sizes:

1. have no impact on how the results and conclusions were reported, that is, No changes needed.
2. have some impact on how the results and conclusions were reported, that is, slight changes needed.
3. have substantial impact on how the results and conclusions were reported, that is, drastic changes needed.
4. have a major impact on how the results and conclusions were reported, that is, a complete revision required.

The inclusion of bandwidth information had a rather dramatic impact on the degree to which one could agree with the results and conclusions reported in the study. Of the 42 results and conclusions examined in light of specific analyses, only three (7.14%) were considered adequate when confidence intervals were considered (see Table 8). A slightly larger amount were determined to need some changes (n = 8, 19.05%) with a greater number possibly needing more substantial changes to the wording (n = 13, 30.95%). The relative majority were considered to need complete revision (n=18, 42.86%) of wording to better reflect appropriate strength of inferences as evidenced in results and conclusions relative to the analysis. The overall findings of this portion of the study are quite comparable to those found by other researchers, as evidenced by a review of randomly selected analyses used in this study. When the recommendations for changes in strength of wording of reported results and conclusions of the 19 sets of analyses reviewed by other researcher specialists were compared with those reached by the researcher conducting this study, there was an strong level of percent agreement (89%).

Examples

Four examples were extracted from the sample to illustrate the basis for reaching each of the four decisions possible, (1) *No Change Needed*, (2) *Slight Change Needed*, (3) *Much Change Needed*, and (4) *Complete Revision Needed.*

One of the few analyses reviewed that had results and/or conclusions that were not considered to be impacted by the reporting of confidence intervals was a study conducted by Sutton and Soderstrom (1999). In this study, the researchers were investigating the impact of variables within the control of a school system such as class size, teacher

experience, and expenditure per pupil as well as those variables considered outside the control of the school system, e.g., mobility, attendance, and low income, on the impact of student achievement. They built regression models to determine the relationship of these variables in combination into two models. One model contained the 'Can Control' variables and the other model contained the 'Cannot Control' variables. The outcome of the regression model for the Cannot Control model indicated statistical significance, with $R^2 = .70$, $p<.001$ for reading achievement and $R^2=.56$, $p<.001$ for math achievement. The author's reported that:

> "In contrast to the low model R^2 values obtained for the can control regression models, the R^2 values obtained for the cannot control regression models were considerably higher. We therefore concluded that the cannot control models accounted more accurately for variance in Grade 3 achievement scores than did the can control variables."

The calculation of confidence intervals around the estimated effect sizes, $2.1149<f^2<2.5706$ for reading and $1.1397< f^2<2.4176$ for math, supports the author's conclusions as the strength of the lower and upper limits of the band are inordinately large. As such, it was determined that *No Change* was necessary, a rating of (1), based on the inclusion of confidence band information.

In a few cases, the results were considered to need only a slight adjustment in wording to reflect the additional information that might be gleaned about the strength of the inference through the use of confidence intervals. In the study by Helwig, Rozek-Tedesco, Tindal, & Heath (1999), researchers were interested in determining if students

would do better on a math test that was augmented with video as compared to the more traditional written test. The general concern was an investigation into how reading level might impact math performance and could be minimized through the use of a video-based delivery of the test as an accommodation. The findings did not reach statistical significance at a .05 Type I error rate (p=.08, no t-value reported) and the author's concluded:

> "Students taking the video version of the test scored slightly higher than those taking the standard version, although that difference was not statistically significant." (p. 121)
>
> and,
>
> "As our results indicate, accommodations are unnecessary for the majority of students." (p. 123)

Based on the confidence interval around the associated effect size which contained an upper limit of close to a small effect, .1012<d<.251, it was determined that the wording might be slightly altered to reflect at least an indication of the potential for an impact of the accommodation, thus being rated a (2) for *Slight Change Needed*.

The use of confidence intervals had more impact on some studies without going as far as requiring a complete revision. Stangor, Carr, and King (1998) conducted a study on whether or not someone's belief that they were chosen for a leadership role based on merit or on group membership (in this case gender) impacted performance. Women were paired with a male individual to perform certain performance tasks. One group was told they were selected based on merit to the be the leader of the pair, the other group was told

they were selected merely based on their gender and not merit. The research team found statistical significance $F(1, 75) = 4.75$, $p<.04$ between the performance of the women, depending on which group they were assigned to. The author's concluded that:

> "As predicted, participants in the gender-only conditioned performed worse than participants in the control and gender+merit conditions." (p. 1191)
>
> and,
>
> "The data were conceptually consistent with prior research in demonstrating that the belief that one has been selected for a task on the basis of gender alone." (p. 1195)

Based on the results of the significance test and point estimate of effect size (Cohen f = .2484) these statements do not appear to be too strong. However, when one considers the confidence interval, $.0225<f<.4867$, with a lower limit close to no effect, then the results seem to be too strongly worded. It would seem that while there does appear that a true difference exists, there is also a possibility that any difference that exists is very small. As such, this case earned a rating of (3), *Much Change Needed.*

Finally, in many cases, the use of confidence intervals impacted the results/conclusions that were written quite strongly, resulting in a recommendation for complete revision. Using an example from one of the studies cited in the previous set of examples, (Mueller & Dweck, 1998) in which children were studied for their response to different types of praise, either for intelligence or effort, as well as the absence of praise, we can also see the potential impact of confidence intervals on findings, albeit from a different perspective statistical significance. In this example, a different group of children

were studied, grouped into the same three categories as before. This part of the study examined the differences regarding how children in the three groups differed in how much they reported enjoying tasks. Unlike the previous example from this study, the findings were statistically significant, $F(2, 120) = 7.73$, $p < .005$ (with three supporting t-tests, all showing statistical significance). Based on the results of these tests, the authors reported that:

> "Children praised for intelligence enjoyed the tasks less than did children praised for effort; again, children in the control conditions fell in between the other two groups. Children praised for intelligence were significantly less likely to enjoy the problems than were children in the effort and control conditions. Further, children in the control condition were less likely to enjoy the problems than those praised for effort" (p. 37)

and,

> "Indictment of ability also led children praised for intelligence to display more negative responses in terms of lower levels of task enjoyment than their counterparts" (p.48).

The results of both the statistical significance tests and practical significance tests supported these assertions to a fair extent with resulting p-values less than .05 on both ANOVA and t-tests and effect sizes ranging from moderate to large point estimates. However, when confidence bands were constructed around the effect sizes, two of the three two-group comparisons included values indicating no effect. Only t-tests between the group of children praised for ability and effort had a confidence band that ranged from

moderate to very strong differences between the two groups ($0.4136 < d < 1.3495$). The bandwidth around the effect size for the differences between children praised for intelligence and those receiving no praise was almost a full standard deviation wide, including a lower band of almost zero ($0.0175 < d < 0.8814$) and the band around the practical effect size between student's receiving praise for effort and those not receiving praise was similar ($0.0043 < d < 0.9158$. This lack of precision in the estimate is alarmingly large and does not support the strength of the author's allegations. As such, it would have been appropriate for the authors to report their findings with indications of the limitations of the inferences that could be drawn between the control group, the effort group and the ability group. The rating received for this analysis regarding change was a (4) for *Complete Revision Needed*.

Summary

The results of this portion of the study provide strong evidence that the inclusion of confidence intervals in reporting research findings may, in fact, severely impact the strength with which one interprets their results. In the majority of the analyses in this study, the width of confidence intervals and their propensity to include measures of a lack of effect or small effect is of concern. Conversely, the ability to report that a confidence interval contains only medium to large effects serves to enhance the strength with which a researcher can draw conclusions. Unfortunately, this latter situation was not the typical situation in the studies found. The use of confidence intervals in approximately 74% of the analyses reviewed resulted in a recommendation that results and conclusions be changed to a large extent, even though they may reflect the findings of significance

testing to a slight degree, or they needed to be completely revamped as they did not substantiate the results and conclusions made based on the significance testing.

Chapter Five

Conclusions

Purpose of Research

The purpose of this research was to examine the potential impact of different methods of reporting research results on the conclusions that could, and should, be made from these findings. Specifically, this study investigated how the use of practical significance as measured by effect sizes in addition to tests of statistical significance might impact the degree to which one should interpret results. Additionally, the use of confidence intervals around point estimates was examined in order to determine the precision of measurements obtained in studies and how that degree of precision might impact conclusions drawn from findings. The three questions investigated in this research were:

1.) To what extent does reporting outcomes of tests of statistical significance vs. tests of practical significance result in different conclusions and/or strengths of inference to be drawn from the results of research?

2.) To what extent does reporting confidence intervals instead of, or in addition to, point estimates affect the conclusions and inferences to be drawn from the results of research?

3.) What method, or combination of methods, is recommended for reporting results in educational studies?

Overview of Method

Journals used in the social sciences were reviewed for inclusion and three rather prominent journals were selected for consideration: *Reading Research Quarterly*, *Journal of Personality and Social Psychology*, and the *Journal of Educational Research*. Previously conducted research deemed worthy of publication that contained one of three rather traditional and oft-used statistical analyses, t-tests, Analysis of Variance (ANOVA), and/or regression were reviewed and results reanalyzed using not only the significance test results provided in the study, but also using the appropriate measures of practical significance (Cohen's *d*, Cohen's *f*, and Cohen's f^2 respectively). Further, confidence intervals for all point estimates, including measures of statistical as well as practical significance were constructed. Results and conclusions relative to specific statistical analyses were then examined with consideration given to the additional information provided by the calculated effect size and confidence intervals. The degree to which the results and conclusions that were presented might be adjusted or reconsidered was estimated.

Impact of Findings

The criticality of thorough and appropriate reporting of research results should be of primary importance to researchers, policy-makers, funding agencies, publishing entities, and practioners alike. The propensity of the current research-based literature to rely almost exclusively on the results of tests of statistical significance has the potential to

rob the consumer of researcher, including fellow researchers and practioners, of important information regarding the strength of the findings of the research. The findings of this study provide evidence that supports the APA's Task Force (Wilkinson, 2001) recommendations to include measures of practical significance as well as confidence intervals when reporting findings of quantitative research.

The additional reporting of measures of practical significance, e.g., effect sizes, had a limited, though often informative, impact on the strength of inferences drawn in the articles examined in this study. However, the inclusion of confidence bands in analyses appears to have the potential for drastic impact on the types and strengths of results and conclusions drawn by researchers. Admittedly, this is one of the reasons that has been suggested regarding the resistance to using intervals as reporting intervals might have the consequence of weakening the strength of conclusions drawn from a study, a rationale at least partially substantiated by the results of this study. While this might be highly likely, it is not, obviously, an ethically sound reason to avoid including this information in results and should be stridently opposed. It is incumbent upon consumers of research to expect inclusion of this type of information if research is to contribute to practice effectively. In the end, it does not benefit the education populace to allow potentially substandard reporting practices to continue.

Statistical Significance vs. Practical Significance

When considering the results of this study, question one, "To what extent does reporting outcomes of tests of statistical significance vs. tests of practical significance result in different conclusions and/or strengths of inference to be drawn from the results

of research?", is addressed with caution. While there were clear indications that effect size reporting did impact a select number of studies, especially those found not to be statistically significant, effect sizes did not, for the most part, drastically alter how one considered the results of studies shown to have statistically significant results. Overall, only 30.57% (n= 14) of the results/conclusions examined were considered to require major or complete revision when considering measures of practical significance in addition to findings of statistical significance.

Although this researcher continues to maintain that the reporting of effect sizes is a reasonable expectation of researchers as it provides a different yet complementary interpretation of results, it does not appear, based on these findings, to have a substantial impact on how one views the results of a large portion of studies reporting statistically significant results found in this type of literature. It is important to note, however, that the vast majority of the studies reviewed in this research contained sample sizes that might be considered small to moderate. Only six of the studies contained samples sizes that exceeded 100 participants, and three of those were from the four regression analyses. This limitation made it somewhat unlikely to see the relationship between statistical significance and practical significance when sample sizes are large. One of the ongoing arguments for reporting measures of practical significance addresses the concern that the likelihood of finding statistically significant results increases as sample size increases. As such, with larger sample sizes, which typically provide enhanced precision of the estimate, there is possibly a greater potential for statistically significant results to have

smaller measures of practical significance that would have further impact on how strongly one can interpret the results of a given study.

The consideration of practical significance measures in analyses containing non-statistically significant results had a slightly greater impact on the findings of this study. The fact that evidence of at least a small effect was present in the majority of analyses reporting the lack of statistical significance, 111 of 166 (66.87%) is quite notable. It may be that the need to consider effect sizes in research is more critical for those finding non-significance, especially if the design of the study is not rigorous. The potential that there exists a true difference between groups as evidenced by an effect size measure that was not found through statistical significance testing may provide enough of a foundational rationale to pursue a particular line of research with enhanced study design.

Point Estimates vs. Confidence Intervals

The results of this study provide a much stronger basis for answering question two: "To what extent does reporting confidence intervals instead of, or in addition to, point estimates affect the conclusions and inferences to be drawn from the results of research?." Clearly the results of both the analytic review of the disparity of confidence band limits in conjunction with the interpretive review of results supports the contention that confidence bands are critical to ensuring that results are interpreted and reported appropriately. Very few bands indicated any strong degree of measurement precision in the findings and this lack of precision weakens the strength with which one should interpret the results. Only 7.15% of the results and conclusions considered were determined to adequately reflect the strength of inference that should be drawn when

confidence interval information was included in addition to results of statistical significance as well as point estimates of effect sizes.

The failure to include measures such as confidence bands is a disservice to the consumer of research. The degree to which one is able to interpret the strength of inference present in any study is key to ensuring that the information is presented properly and thoroughly. The lack of including this type of information is likely to result in conclusions that are, at best, misleading, and at worst, incorrect.

Reporting Results

In order to address question three, "What method, or combination of methods, is recommended for reporting results in educational studies?", many elements of the nature of the research to be conducted and study design need to be taken into account. It doesn't seem reasonable to consider that the reporting of all three types of information, statistical significance, practical significance, or confidence bands, should ever be discouraged on considered as unacceptable due to such things as limits on manuscript length for publication purposes. One of the studies used in the examples provided earlier (Mueller & Dweck, 1998) clearly illustrated how the use of both practical significance and confidence intervals can impact different aspects of findings and conclusions in different ways within one study. The strength of non-significant findings were found to be questionable when considering measures of practical effect and the strength of statistically significant findings were weakened when considering confidence intervals. However, it is important to realize that the criticality of including such measures may vary by study. Practical significance measures in statistically significant analyses

provides additional information that can contribute to interpretation of results but may have limited substantive contribution to changes in overall conclusions and findings, especially when sample sizes are small to moderate. One of the concerns about the limitations of statistical significance tests is the tendency to find statistical significance as sample size increases. This research, due to the limitations inherent in it, did not possess many studies that had very large samples. As such, it is quite possible that the importance of including measures of practical significance in studies with statistically significant results increases as sample size increases. In studies that have do not have statistically significant results, the importance of including effect sizes appears to have more impact as it may be a key piece of information that may or may not help researchers determine whether or not to pursue a given line of research.

While the recommendations about whether or not to include measures of practical significance are somewhat murky, the same cannot be said regarding confidence intervals. The results of this study clearly indicate that the importance of including such a measure to assist with determining the precision of research results. To not include this information is to withhold critical information for consumers of research and should not only be encouraged, but, increasingly be made an expectation.

When considering recommendations for what to include in research reporting, a critical element guiding decisions must be the intended use of the findings. If research findings will impact decisions on such things as funding, policy-making, or choice of curriculum, the importance of providing all relevant information about effectiveness and significance of research reports cannot be underestimated. The more critical a decision is,

the more information should be provided. To that end, the information gleaned from practices such as effect size reporting and confidence intervals should always be reported.

Relevant Issues

In addition to the findings that were a direct goal and consequence of this research, other issues were identified that impact the overall integrity of research reporting. Few studies reported what many might consider to be highly important information regarding research design and data characteristics (e.g., distributional information, reliability and validity data). Of particular note was the dearth of information about the Type I error rate at which a given study was being conducted. Related to this issue, studies that used more than one t-test did not indicate that they had performed any special analyses, e.g., Bonferroni adjustments, to compensate for the possibility of inflated type I error rates due to multiple comparisons. Relative to this study, this issue requires further investigation into how one thinks about constructing CIs under these conditions. That is, do the algorithms for constructing confidence intervals need to be adjusted under situations that have multiple comparison tests?

In most cases, one had to make assumptions of the alpha level based on what they reported as significant. The infamous 'asterisks in the table' did not dominate all the studies but was a notable contributor to the inability to determine what Type I error rate was of true interest. This seems to indicate an underlying violation of one of the basic tenants of good research taught in most beginning research courses: the need for the researcher to make a decision, based on criticality of the research and knowledge of their field, regarding the alpha level that he or she is going to conduct significance testing at *a*

priori to actual conduct of research. The obvious absence of the communication of this rather foundational aspect of a research design is just one possible reason that the ethics of research is sometimes called into question.

The findings of this study also impacts how one thinks of the disciplinary norms associated with the reporting of research contained within the disciplines within the social sciences. Perhaps the community as a whole needs to consider the accepted practices of reporting research in such disciplines as education and psychology regarding their current expectation and what, perhaps, might be changed to make the research available less open to criticism or alternative interpretations. Even within a given discipline, the roles of different professionals within that discipline will influence how they think about, interpret, and apply results and conclusions of research. Within this research itself, this issue is evident. For example, other methodologists with similar backgrounds and training to the researcher conducting this study conducted the review of the interpretative results. As such, the rather strong level of interrater reliability can only be used to support the contention that other methodological researchers would draw the same types of conclusions. In cannot be used to support a claim that other consumers of researchers, e.g., practitioners, theorists, etc., would have similar interpretations regarding the impact that effect size and/or confidence interval information might impact their view of the results and conclusions.

A final element that should be considered if there is to be any potential for changing the reporting practices of researchers is preparation of future scholars, researchers, and practitioners. Members entering into a given profession engage in the

practices for which they have been trained and instructed on. As such, in addition to trying to reach those currently active in the engagement, dissemination and consumption of research, it seem critical to be properly training and educating those entering the field on appropriate reporting practices. New researchers should be made aware of both the frailties and merits of various options of reporting results. The type of information provided by effect size estimates as well as confidence intervals should be an important element of that training.

Future Research

The findings of this study strongly support the need for further investigation of the impact of research reporting practices on the integrity and interpretability of published research. This study was an initial foray into the practical implications of using effect size information as well as confidence intervals in addition to measures of statistical analyses. Future studies might benefit the research community by selecting a more specific genre of research literature to review in order to assess impact on specific fields, e.g., subject specific research such as mathematics, administrative based research such as policy analyses, or different levels of development such as specific school levels. Additionally, similar studies within a given field but with respect to varying professional roles and responsibilities within those fields, e.g., practitioner vs. statistician, would provide yet another way of considering how different individuals and professionals perceive results based on how they are reported.

One might also consider an extended examination of the impact of publication source on how much measures of practical significance and confidence intervals are

either reported, or impact published findings. In general, the findings of this study did not indicate any strongly notable differences between the three somewhat diverse journals used, with the exception of the types of statistical analyses typically used; however, other explorations with a focus on this as a primary question might have different results.

Additionally, the relationship of the importance of measures such as practical significance and confidence intervals with the design of research studies is likely to be vital to determining the true utility of these measures in research reporting under certain conditions. Research into more definitive impacts of design characteristics such as sample size, heterogeneity of samples, etc. in applied research studies, along with an evaluation of their impact on effect sizes and confidence intervals, would be very beneficial to researchers throughout the social sciences.

The other element of this type of issue is the need for research from the point of view of the consumers of research. One of the issues that became evident when measurement specialists were used to determine possible changes in the results reported was their tendency to use all aspects of the research design in consideration of their ratings. How this might change when the reader is less likely to well-versed in measurement, statistical analyses and research design is an important distinction that might further guide refinements making determinations and judgments about appropriate practices in reporting research.

A final consideration for future research taps into the preparation of researchers. It could be quite enlightening to investigate the extent to which graduate students are trained and instructed on the use of various reporting methods and practices when

conducting research studies. This type of inquiry could take on many forms, from course content reviews, e.g., syllabi, textbook reviews, to a methodological review of dissertations and thesis'. An examination of how often effect sizes and confidence interval information is provided in new scholars work would provide some evidence regarding the extent to which new researchers are entering the field prepared to report findings above and beyond the results of significance testing.

Summary

The findings of this research reinforce the need for increased emphasis on appropriate and thorough research reporting practices. Individuals in leadership positions that have critical decision-making power in the research world, e.g., administrators, policy-makers, journal editors, funding sources, etc. need to require enactment and enforcement of more in-depth research reporting practices and protocols. Without substantial requirements of such guiding forces in research as well as enforcement of these requirements, the quality of research reported in the social sciences is not likely to see any substantial change or improvement.

The degree of quality of research in any field does not merely impact the research community. Poor research has the potential to damage the leadership of a professional community, the policy and guidelines constructed for that community, and ultimately, the consumers or customers within that community. In education, this translates to damage to the learner. As a society that values education and understands that a strong educational foundation is necessary to keep society strong, we cannot afford to overlook the importance of insuring that sound research practices are in place for all aspects of

research conduct, including study design, method, conduct and reporting. The idea that there is a problem with the quality of educational and social science research is not new and it is incumbent upon leaders in the field that guide policy and funding to take strong actions to improve the situation. It is often suggested that research should guide practice. What benefit is that if the research is poorly conceived, designed or reported?

References

Abelson, R. P. (1995). *Statistics as Principled Argument.* Hillsdale NJ: Lawrence Erlbaum Associations.

American Educational Research Association (n.d.). Call for Manuscripts 2003-2006, Review of Educational Research. Retrieved May 18, 2003 from http://www.aera.net/pubs/rer/recall.htm

* Alspaugh, J. W. (1998). Achievement loss associated with the transiton to middle school and high school, *Journal of Educational Research, 92*(1), 20-25.

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Barnett, R. J., Docherty, J. P., & Frommelt, G. M. (1991). A review of child psychotherapy research since 1963. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*(1), 1-14.

* Baumeister, R. F., Twenge, J. M., & Nuss, C. K. (2002). Effects of social exclusion on cognitive processes: Anticipated aloneness reduces intelligent thought, *Journal of Personality and Social Psychology, 83*(4), 817-827.

Bazerman, C. (1981). What written knowledge does, *Philosophy of the Social Sciences, 2*, 361-387.

Becher, T. (1987). Disciplinary discourse, *Studies in Higher Education, 12*, 261-274.

* Blanton, J., Buunk, B. P., Gibbons, F. X., & Kuyper, H. (1999). When better-than-others compare upward: Choice of comparison and comparative evaluation as independent predictors of academic performance, *Journal of Personality and Social Psychology, 76*(3), 520-430.

* Bowman, C. L. & McCormick, S. (2000). Comparison of peer coaching versus traditional supervision effects, *Journal of Educational Research, 93*(4), 256-382.

Bradley, M. T. & Gupta, R. D. (1997). Estimating the effect of the file drawer problem in meta-analysis. *Perceptual and Motor Skills, 85*, 719-722.

* Brown, R. P., Charnsangavej, T., Keough, K. A., Newman, M. L., & Renfrow, P. J. (2000). Putting the "Affirm" into affirmative action; preferential selection and academic performance, *Journal of Personality and Social Psychology, 79*(5), 736-747.

* Brown, R. P. & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance, *Journal of Personality and Social Psychology, 76*(2), 246-257.

* Callan, R. J. (1999). Effects of matching and mismatching students' time-of-day preferences, *Journal of Educational Research, 92*(5), 295-299.

Carpenter J. W. & Bithell C. (2001). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine, 19*, 1141-1164.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Academic Press.

Cohen, J., Cohen P., West, S. G., & Aiken, L. S. (2003). *Applied Mulitple Regression/Correlational Analysis for the Behavioral Sciences (3rd ed.).* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cooper H. & Hedges, L. (1994). *The Handbook of Research Synthesis*. New York: Russel Sage Foundation.

Cumming G. & Finch S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentrl distributions. *Educational and Psychological Measurement, 61(4),* 532-74.

* Davies, M. F. (1998). Dogmatism and belief formation: Output interference in the processing of supporting and contradictory cognitions, *Journal of Personality and Social Psychology, 75*(2), 456-466.

Davis, O.L. (2001). So what? *Journal of Curriculum and Supervision, 16*(2), 91-94.

DiPrete T. A. & Forristal, J. D. (1994). Multilevel models: methods and substance. *Annual Review of Sociology, 20,* 331-359.

* Evertson, C. M. & Smithey, M. W. (2000). Mentoring effects on proteges classroom practice: An experimental field study, *Journal of Educational Research, 93(*5), 294-204.

Fan, X. (2001). Statistical significance and effect size in education research: two sides of a coin. *The Journal of Educational Research, 94*(5), 275-282.

Fan, X. & Thompson, B. (2001). Confidence intervals about score reliability coefficients. Please: An EPM guidelines editorial. *Educational and Psychological Measurement 61*(4), 517-531.

Fidler F. & Thompson B. (2001). Computing correct confidence intervals for ANOVA fixed-and random-effects effect sizes. *Educational and Psychological Measurement, 61(4),* 575-604.

* Fitzgerald, J. (2001). Can minimally trained college student volunteers help young at-risk children to read better? *Reading Research Quarterly 36*(1), 28-47.

* Galassi, J. P., White, K. P., Vesilind, E. M. & Bryan, M. E. (2001). Perceptions of research from a second-year, multisite professional development schools partnership, *Journal of Educational Research, 95*(2), 75-83.

Gall M. D, Borg W. R, & Gall J. P. (1996). *Educational Research: An Introduction (6th ed.),* New York: Longman Publishers.

Gerholm, T. (1990). On tacit knowledge in acadamia, *European Journal of Education, 25,* 263-271.

Girden, E. R. (2001). *Evaluating Research Articles from Start to Finish (2nd ed.).* Thousand Oaks, CA: Sage Publications.

Glass, G. V. & Hopkins, K. D. (1996). *Statistical Methods in Education and Psychology (3rd ed.).* Needham Heights, MA: Allyn and Bacon.

Grissom R. J. & Kim J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, 6(2), p. 135-146.

* Hancock, D. R. (2000). Impact of verbal praise on college students' time spent on homework, *Journal of Educational Research, 93*(6), 384-389l

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

Hedges L. V. & Olkin I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

* Helwig, R., Rozek-Tedesco, M. A., Tindal, G., & Heath, B. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students, *Journal of Educational Research, 93*(21), 113-125.

Hess, M. H. & Kromrey J. D. (2001, November). *Confidence Intervals around Standardized Mean Differences: An empirical comparison of methods for constructing confidence bands around standardized mean differences.* Paper presented at the annual meeting of the Florida Educational Research Association, Marco Island FL.

Hess M. H. & Kromrey J. D. (2002, November). *Variations on the Bootstrap: A comparison of confidence band coverage for the standardized mean difference.* Paper presented at the annual meeting of the Florida Educational Research Association, Gainesville, FL.

Hittleman, D. R. & Simon, A. J. (2002). *Interpreting Educational Research: An Introduction for Consumers of Research (3rd ed.).* Upper Saddle River, New Jersy: Merrill Prentice Hall.

Hogarty K. Y. & Kromrey, J. D. (1999, August). *Traditional and robust effect size estimates: Power and Type I error control in meta-analystic tests of homogeneity.* Paper presented at the Joint Statistical Meetings, Baltimore.

Hubbard, R. & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology-and its future prospects. *Educational and Psychological Measurement, 60*(5), 661-681.

Huff, D. (1954). *How to Lie with Statistics*. New York: Norton.

Institute for Scientific Information (2002). Social Sciences Citation Index Journal Citation Reports for 2001 [Microform]. Philadelphia, PA: Institute for Scientific Reform.

* Jenkins, E., Queen, A., & Algozzine, B. (2002). To block or not to block: That's not the question, *Journal of Educational Research, 95*(4), 196-202.

* Jordan, G. E., Snow, C. E., & Porche, M. V. (2000). Project EASE: The effect of a family literacy project on kindergarten students' early literacy skills. *Reading Research Quarterly, 35*(4), 524-546.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses, *Review of Educational Research*, 68(3), 350-386.

* Kim, H. S. (2002). We talk, therefore we think? A cultural analysis of the effect of talking on thinking, *Journal of Personality and Social Psychology, 83*(4), 828-842.

Knapp, T. R., Sawilowsky, S. S. (2001). Constructive criticisms of methodlogical and editorial practices, *The Journal of Experimental Education, 70*(1), 65-79.

Kromrey, J. D. & Hess, M. H. (2001, April). *Interval Estimates of R^2: An empirical comparison of accuracy and precision under violations of the normality assumption*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Kromrey, J. D. & Hess, M. H. (2002, November). *Constructing Confidence Bands Around Mean Differences Between Non-homogeneous Groups: An Empirical Comparison of Methods.* Paper presented at the annual meeting of the Florida Educational Research Association, Gainesville, FL.

* Lepore, S. J., Ragan, J. D., & Jones, S. (2000). Talking facilitates cognitive-emotional processes of adaptation to an acute stressor, *Journal of Personality and Social Psychology, 78*(3), 499-508.

* Leseman, P. M. & de Jong, P. F. (1998). Home literacy: Opportunity, instruction, cooperation and social-emotional quality predicting early reading achievement. *Reading Research Quarterly, 33*(3), 294-318.

McEwan, E. K. & McEwan, P. J. (2003). *Making Sense of Research. What's Good, What's Not, and How to Tell the Difference.* Thousand Oaks, CA: Corwin Press.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds.), *What if There Were No Signficance Tests?* (pp. 393-425). Mahway, NJ: Lawrence Erlbaum.

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

Morgan, D. L. (1996). Focus Groups (using focus groups in research). *Annual Review of Sociology, 22*, 129-153.

* Mori, Y. & Nagy, W. (1999). Integration of information from context and word elements in interpreting novel kanji compounds. *Reading Research Quarterly, 34*(1), 80-101.

* Mueller, C. M. & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance, *Journal of Personality and Social Psychology, 75*(1), 33-52.

Muliak, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds.), *What if There Were No Significance Tests?* (pp. 65-116). Mahway, NJ: Lawrence Erlbaum.

Nix, T. W. & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2), p. 3-14.

* Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math=Male, Me=Female, therefore Math ≠ Me. *Journal of Personality and Social Psychology* 83(1), 44-59

* Paris, A. H. & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly, 38*(1), 36-76.

Parry, S. (1998). Disciplinary discourse in doctoral theses. Higher Education, 36, 273-299.

Plucker, J. A, (1997). Debunking the myth of the "highly significant result: effect sizes in gifted education research, *Roeper Review, 20*, 122-126.

Reichardt, C. S. & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds.), *What if There Were No Significance Tests?* (pp. 260-284). Mahway, NJ: Lawrence Erlbaum.

Robinson, D. H., Fouladi, R. T., & Williams, N. J. (2002). Some effects of including effect size and "what if" information. *The Journal of Experimental Education,* 70(4), 365-382.

Robinson, D. H. & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26 (5),* 21-27.

Rosentahl, R. (1992). Effect size stimation, significance testing, and the file-drawer problem. Journal of Parapsychology, 56, 57-58.

Rosenthal, R. (1988). Parametric measures of effect size. In H. Cooper and L.V. Hedges (Ed.) *The Handbook of Research Synthesis* (pp. 231-244). New York, NY: Russel Sage.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

* Roth, F. P., Speece, D. L., & Cooper, E. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *Journal of Educational Research, 96*(5), 259-271.

* Santa, C. M. & Hoien, T. (1999). An assessment of Early Steps: A program for early intervention of reading problems. *Reading Research Quarterly, 34*(1), 54-79.

Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L.

Harlow, S.A. Muliak & J.H. Steiger (Eds.), *What if There Were No Signficance Tests?* (pp. 37-64). Mahway, NJ: Lawrence Erlbaum.

* Spooner, F., Jordan, L., Algozzine, B., & Spooner, M. (1999). Student ratings of instruction in distance learning and on-campus classes, *Journal of Educational Research, 92*(3), 132-140.

* Stangor, C., Carr, C., & King, L. (1998). Activating stereotypes undermines task performance expectations, *Journal of Personality and Social Psychology, 75*(5), 1191-1197.

Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds.), *What if There Were No Signficance Tests?* (pp. 221-257). Mahway, NJ: Lawrence Erlbaum.

Steiger, J. H. & Fouladi, R. T. (1992). R^2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research, Methods, Instruments, and Computers, 4,* 581-582.

Stevens, J. (1999). Intermediate Statistics: A Modern Approach. Mahway, N.J.: Lawrence Earlbaum.

Stine, R. (1990). An introduction to bootstrap methods. *Sociological Methods and Research, 18 (2&3),* p. 243-291.

* Sutton, A. & Soderstrom, I. (1999). Predicting elemtary and secondary school achievement with school-related and demographic factors. Journal of Educational Research, 29(6), 330-338.

Tashakkori, A. & Teddlie, C. (1998). *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* Thousand Oaks California: Sage Publications.

* Thompkins, A. C. & Binder, K. S. (2003). A comparison of factors affecting reading performance of functionally illiterate adults and children matched by reading level. Reading Research Quarterly, 38(2), 236-258.

Thompson, B. (2002a). "Statistical," "practical," and "clinical": how many kinds of significance to counselors need to consider?. *Journal of Counseling and Devlopment, 80*(1), 64-71.

Thompson, B. (2002b). What future quantitative social science research could look like: confidence intervals for effect sizes, *Educational Researcher,* 25-32.

Thompson, B. (1999a). Improving research clarity and usefulness with effect size indices as supplements to statistical tests, *Exceptional Children, 65*(3), 329-337.

Thompson, B. (1999b). Why "encouraging" effect size reporting is not working: the etiology of researcher resistance to changing practices. *The Journal of Psychology, 133*(2), 133-40.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*(2), p. 33-38.

* Tiedens, L. Z. & Linton, S. (2001). Judgment under emotional certainty and uncertainty: The effects of specific emotions on information processing. *Journal of Personality and Social Psychology, 81*(6), 973-988.

Tuckman, B. W. (1990). A proposal for improving the quality of published educational research. *Educational Researcher, 19*(9), 22-24.

United States Department of Education (n.d.). Inside No Child Left Behind. Retrieved May 29, 2003 from http://www.ed.gov/legislation/ESEA02

United States Department of Education (n.d). No Child Left Behind Act Factsheet. Retreived May 27, 2003 from http://www.ed.gov/offices/OESE/esea/factsheet.html

University of South Florida Virtual Library (n.d.). . Retrieved June 5, 2003 from http://www.usf.virtuallibrary.edu

* Van den Branden, K. (2000). Does negotiation of meaning promote reading comprehension? A study of multilingual primary school classes. *Reading Research Quarterly, 35*(3), 426-443.

Wilkinson & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Appendices

Appendix A

Article Coding Sheet

*

Title of Article: _____

Authors:
_____

Website, date accessed (if applicable): _____

Journal Name: _____ Vol (No): _____ Date: _____ Pgs: _____

Preliminary Screening Information:

Which of the three analyses of interest are used in this study: _____ T-tests _____ Regression _____ ANOVA

Is one of the 'analyses of interest' the primary analysis used for this study? ____yes ____no

If 'no', explain relationship of analysis to be focused on to other analyses in the study. (ex. T-tests are used to provide supportive and/or additional information in a study that uses SEM as the primary analysis. _____
_____
_____
_____
_____
_____
_____

General Description of Study:

<table>
<tr><td>Date(s) of Study:
_____</td><td>Conducted by:
_____</td><td>Description of participants (Age, grade, school, etc.):
_____
_
_____
_
_____
_</td><td>Where was study conducted (classroom, school, lab)
_____
_</td><td colspan="3">Purpose of Study: _____
_____
_____
_____</td></tr>
<tr><td rowspan="2">All Method(s) used:</td><td>____Regression</td><td>____ Qual.</td><td>____ t-tests</td><td>____ ANOVA</td><td>____ ANCOVA</td><td>___ MANOVA</td></tr>
<tr><td>____ HLM</td><td>____ SEM</td><td>____ Other _____</td><td colspan="2">_____ Other _____</td></tr>
<tr><td colspan="7">How was missing data handled? (not discussed, listwise deletion, imputation, etc.): _____
_____</td></tr>
<tr><td colspan="7">Was power discussed? If so, briefly describe: _____
_____</td></tr>
<tr><td colspan="7">Were validity and reliability discussed? If so, briefly describe: _____
_____</td></tr>
<tr><td colspan="7"></td></tr>
<tr><td rowspan="2">No of groups:
_____</td><td rowspan="2">Demographics:</td><td>____ Race/Ethnicy</td><td>_____ Gender</td><td>_____ Age</td><td colspan="2">____ SES</td></tr>
<tr><td>____ Other _____</td><td colspan="3">____ Other _____</td></tr>
<tr><td colspan="7">Other characheristics/issues of study: _____
_____
_____
_____
_____
_____</td></tr>
</table>

Appendix A

Article Coding Sheet

Two groups						
N1	N2	Mean1	Mean2	Sd1	Sd2	Sig rep?
Regression						
N	K	R2				
ANOVA						
N	K	F				

Appendix B

Article Reviewer Instructions and Cases

Reviewer Instructions
(to be provided verbally as well as written):

You have received a collection of analyses pulled from published research. Each analysis contains a synopsis of the study with relevant statistical information provided as well as results and conclusions reported by the author(s) of the study. The synopsis is not necessarily a direct quote from the study investigated, rather it is a summary; However, all statistical information and related results/conclusions are directly from the article of interest and words are direct quotes pertaining to the statistical information provided.

1. Please read the synopsis and analysis reported. Then, review the author's words regarding their interpretation and application of that statistical analysis. Once you have reviewed the analysis and results, decide whether or not you concur with the findings/results of the author as reported and to what degree, and then complete item A on the review sheet.

2. After completing item A, consider the calculated effect size provided. Using Cohen's definitions of effect size, decide whether or not you concur with the findings/results of the author as reported and to what degree, and then complete item B on the review sheet.

	Effect Size Index		
	Cohen's *d*	Cohen's *f*	Cohen's f^2
Small Effect	0.20	0.10	0.02
Medium Effect	0.50	0.25	0.15
Large Effect	0.80	0.40	0.35

3. Finally, consider the confidence interval calculated at a Type I error rate of 0.05 which indicates we are 95% confident that 'truth' resides somewhere within that band, although where we do not know. When considering the interval and related results/conclusions reported, take into account such characteristics of the interval such as lower and upper limits, width, etc. Using this information, again decide whether or not you concur with the findings/results of the author as reported and to what degree, and then complete item B on the review sheet.

Appendix B

Article Reviewer Instructions and Cases

Study Number: ____ Analysis: ____ Coder: _____

A. Based on the information provided **by the author** regarding **statistical significance**, I:

_____ Agree completely with the results/conclusions drawn. ***No changes needed.***

_____ Agree in essence with the results/conclusions provided; However, ***wording of results/conclusions should be changed slightly*** to better reflect appropriate strength of inferences, generalizability, etc.

_____ Agree a little bit with the results/conclusions provided; However, ***wording of results/conclusions should be changed drastically to*** better reflect appropriate strength of inferences, generalizability, etc.

_____ Disagree completely with the results/conclusions drawn. ***Complete revision needed.***

B. Based on the information provided **by the researcher** regarding **practical significance**, I:

_____ Agree completely with the results/conclusions drawn. ***No changes needed.***

_____ Agree in essence with the results/conclusions provided; However, ***wording of results/conclusions should be changed slightly*** to better reflect appropriate strength of inferences, generalizability, etc.

_____ Agree a little bit with the results/conclusions provided; However, ***wording of results/conclusions should be changed drastically to*** better reflect appropriate strength of inferences, generalizability, etc.

_____ Disagree completely with the results/conclusions drawn. ***Complete revision needed.***

C. Based on the information provided **by the researcher** regarding **95% confidence intervals**, I:

_____ Agree completely with the results/conclusions drawn. ***No changes needed.***

_____ Agree in essence with the results/conclusions provided; However, ***wording of results/conclusions should be changed slightly*** to better reflect appropriate strength of inferences, generalizability, etc.

_____ Agree a little bit with the results/conclusions provided; However, ***wording of results/conclusions should be changed drastically to*** better reflect appropriate strength of inferences, generalizability, etc.

_____ Disagree completely with the results/conclusions drawn. ***Complete revision needed.***

Appendix B

Article Reviewer Instructions and Cases

Study Number: 68

Synopsis of Study: The purpose of this study was to determine whether or not the prediction of impending misfortune and/or aloneness (emphasis was on aloneness) impacted perseverance and/or cognitive abilities. Three groups were assembled. Based on results of assessments administered, they were told that they would either: 1) Spend the rest of their life surrounded by people who care about them, 2) be accident prone the rest of their life, or 3) become increasingly alone in life (lose friends over time, not replaced). Participants were then administered an intelligence test. Measurement were taken regarding number of items attempted and total score.

Statistical Significance Reported with associated results and conclusions:

Analysis 1:

Issue addressed: Difference between groups regarding correctness of answers

Statistical Signficance Information: $F(2, 37) = 5.44$, $p < .01$

Relevant Results/Conclusions:

Participants in the future alone condition answered significantly fewer questions correctly, as compared with participants in the future belonging and misfortune condition (p. 819)

Thus, hearing that one was likely to be alone later in life affected performance on a timed cognitive test. (p. 819-820)

A diagnostic forecast of future social exclusion caused a significant drop in intelligent performance (p. 820)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.5215

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.1372

Upper Cohen's f: 0.8318

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 68

Synopsis of Study: The purpose of this study was to determine whether or not the prediction of impending misfortune and/or aloneness (emphasis was on aloneness) impacted perseverance and/or cognitive abilities. Three groups were assembled. Based on results of assessments administered, they were told that they would either: 1) Spend the rest of their life surrounded by people who care about them, 2) be accident prone the rest of their life, or 3) become increasingly alone in life (lose friends over time, not replaced). Participants were then administered an intelligence test. Measurement were taken regarding number of items attempted and total score.

Statistical Significance Reported with associated results and conclusions:

Analysis 2:

Issue addressed: Difference between groups in effort, as measured by number of items attempted.

Statistical Signficance Information: $F(2, 37) = 3.46$, $p < .05$

Relevant Results/Conclusions:

> This analysis again showed significant variation among the three conditions. Participants in the future alone condition attempted the fewest problems. Again, the deficit was specific to feedback about social exclusion, insofar as participants in the misfortune control condition attempted as many problems (if not more) than the people in the future belonging condition (p. 820)
>
> The decline in performance reflected both a higher rate of errors and reduced number of problems attempted (p. 820)
>
> A diagnostic forecast of future social exclusion caused a significant drop in intelligent performance (p. 820)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.4159

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.000

Upper Cohen's f: 0.7149

Appendix B

Article Reviewer Instructions and Cases

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 53

Synopsis of Study: The purpose of this study was to determine whether or not the degree to which someone was considered dogmatic impact such things as their confidence and tendency to be judgmental. This study also investigated the degree to which dogmatism impacted an individual's ability to provide reason behind decisions and judgments and the nature of those reasons. Faced with two possible outcomes to given scenarios (e.g., likelihood of persons stopping to help an injured person with blood present vs no blood present), participants selected their prediction of the outcome and then indicated how confident they were in their decision. They then listed reasons why they thought their outcome was most likely (pro decisions) as well as reasons why the other outcome might occur (con decisions)

Analysis 1:

Issue addressed: Difference in confidence between individuals classified as high or low in dogmatism.

Statistical Signficance Information: $F(1, 61) = 3.46$, $p < .01$

Relevant Results/Conclusions:

> Individuals high in dogmatism were much more confident in their judgments (M=7.17) than individuals low in dogmatism (M=6.19). (p.458)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.2905

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.0505

Upper Cohen's f: 0.5238

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 53

Synopsis of Study: The purpose of this study was to determine whether or not the degree to which someone was considered dogmatic impact such things as their confidence and tendency to be judgmental. This study also investigated the degree to which dogmatism impacted an individual's ability to provide reason behind decisions and judgments and the nature of those reasons. Faced with two possible outcomes to given scenarios (e.g., likelihood of persons stopping to help an injured person with blood present vs no blood present), participants selected their prediction of the outcome and then indicated how confident they were in their decision. They then listed reasons why they thought their outcome was most likely (pro decisions) as well as reasons why the other outcome might occur (con decisions)

Statistical Significance Reported with associated results and conclusions:

Analysis 2:

Issue addressed: Are there differences in the types of reasons provided for outcomes that support an individuals opinion (pro decisions) as compared to the reasons that oppose an individual's opinion (con decisions) resulting from how dogmatic an individual is?

Statistical Signficance Information:

Due to the nature of the issue and statistics provided to support results and conclusion, consideration of data from two main effects and an interaction effect are necessary for this analyses. Please use all relevant information when deciding on how you will answer the review sheet.

Main effect of dogmatism on generation of 'pro' reasons.

$F(1, 61) = 3.47, \quad p < .07$

Main effect of dogmatism on generation of 'con' reasons:

$F(1,61) = 3.07, \quad p < .08$

Interaction of level of dogmatism and type of reason generated

$F(1,61) = 10.03, \quad p < .01$

Relevant Results/Conclusions:

There was a significant interactions of dogmatism with type of reason generated (see interaction information). Individuals high in dogmatism produced more pro reasons than individuals low in dogmatism (see main effect 1). Also, they produced fewer con reasons than individuals low in dogmatism (see main effect 2). (p. 458)

The results (of the experiment) show that individuals high in dogmatism are more likely to generate cognitions supporting their newly created

beliefs and are less likely to generate cognitions contradicting them. (p.459)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.2347
Calculated Effect Size (Cohen's f) 0.2207
Calculated Effect Size (Cohen's f) 0.4049

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)
Lower Cohen's f: 0.0
Upper Cohen's f: 0.4842

Calculated Confidence Interval (95%)
Lower Cohen's f: 0.0
Upper Cohen's f: 0.4699

Calculated Confidence Interval (95%)
Lower Cohen's f: 0.1462
Upper Cohen's f: 0.6605

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 52

Synopsis of Study: The purpose of this study was to determine if the types of praise given to children impacted their motivation and performance. Children were placed in three groups. In the two experimental groups, children were given different types of praise for accomplishments. The first group was praised on ability and children wer told 'You must be smart at these problems' and the second group was praised on effort, 'You must have worked hard at these problems.' The third group was controlled and given no feedback. Students were subsequently given measures that rated persistence, enjoyment, quality of performance and failure attributions. Additionally, they were administered a second assessment (similar to the one that they had received praise on) of similar difficulty.

Analysis 1:

Issue addressed: Do children who receive different types of praise (ability, effort, or none) differ in what they attribute their performance (effort or intelligence) to on performance measures?

Statistical Significance Information:

Two main effects reported, no interactions:
Effect of 'low effort' on performance: $F(2,120) = 8.64$, $p < .001$
Effect of 'low intelligence' on performance: $F(2, 120) = 4.63$, $p < .05$

Relevant Results/Conclusions:

Children differed in their endorsements of low effort and low ability as causes of their failure (p.37)

Overall, the findings (of the study) support our hypothesis that children who are praised for intelligence when they succeed are the ones leastlikely to attribute their performance to low effort, a factor over which they have some amount of control. (p.39)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.3748
Calculated Effect Size (Cohen's f): 0.2744

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.1750
Upper Cohen's f: 0.5482

Article Reviewer Instructions and Cases

Calculated Confidence Interval (95%)
Lower Cohen's f: 0.0621
Upper Cohen's f: 0.4423
Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 52

Synopsis of Study: The purpose of this study was to determine if the types of praise given to children impacted their motivation and performance. Children were placed in three groups. In the two experimental groups, children were given different types of praise for accomplishments. The first group was praised on ability and children wer told 'You must be smart at these problems' and the second group was praised on effort, 'You must have worked hard at these problems.' The third group was controlled and given no feedback. Students were subsequently given measures that rated persistence, enjoyment, quality of performance and failure attributions. Additionally, they were administered a second assessment (similar to the one that they had received praise on) of similar difficulty.

Analysis 2:

Issue addressed: Do children who receive different types of praise (ability, effort, or none) differ in how they rate their enjoyment of tasks?

Statistical Significance Information:

Difference between groups:

$F(2, 120) = 7.73,\ p<.005$

Follow up groups comparisons:

Ability vs. Effort, $t(81) = -3.81$, $p<.001$

Ability vs. Control, $t(83) = -2.03$, $p<.05$

Control vs. Effort, $t(82) = 2.16$, $p< .05$

Relevant Results/Conclusions:

Children praised for intelligence (M= 4.11) enjoyed the tasks less than did children praised for effort (M=4.89); again, children in the control condition fell in between the other two groups (M=4.52) Children praised for intelligence were significantly less likely to enjoy the problems than were children in the effort and control conditions. Further, children in the control condition were less likely to enjoy the problems than those praised for effort. (p.37)

Indictment of ability also led children praised for intelligence to display mor negative responses in terms of lower levels of task enjoyment than their counterparts, who received commendations for effort. (p.48)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.3545

Calculated Effect Size (Cohen's t): -0.8816
Calculated Effect Size (Cohen's t): -0.4495
Calculated Effect Size (Cohen's t): -0.4801

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)
Lower Cohen's f: 0.1358
Upper Cohen's f: 0.5269
Calculated Confidence Interval (95%)
Lower Cohen's t: -0.4136
Upper Cohen's t: -1.3495
Calculated Confidence Interval (95%)
Lower Cohen's t: -0.0175
Upper Cohen's t: -0.8814
Calculated Confidence Interval (95%)
Lower Cohen's t: 0.0043
Upper Cohen's t: 0.9158

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 52

Synopsis of Study: The purpose of this study was to determine if the types of praise given to children impacted their motivation and performance. Children were placed in three groups. In the two experimental groups, children were given different types of praise for accomplishments. The first group was praised on ability and children wer told 'You must be smart at these problems' and the second group was praised on effort, 'You must have worked hard at these problems.' The third group was controlled and given no feedback. Students were subsequently given measures that rated persistence, enjoyment, quality of performance and failure attributions. Additionally, they were administered a second assessment (similar to the one that they had received praise on) of similar difficulty.

Analysis 3:

Issue addressed: Do children who receive different types of praise (ability, effort, or none) differ regarding their future expectations of their performance?

Statistical Significance Information:

Difference between groups:

$F(2, 48) = 1.01$, ns

Relevant Results/Conclusions:

No significant differences were noted for children's expectations; children in the intelligence, effort, and control conditions displayed equivalent expectations. (p.40)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.199

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.0

Upper Cohen's f: 0.4419

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 52

Synopsis of Study: The purpose of this study was to determine if the types of praise given to children impacted their motivation and performance. Children were placed in three groups. In the two experimental groups, children were given different types of praise for accomplishments. The first group was praised on ability and children wer told 'You must be smart at these problems' and the second group was praised on effort, 'You must have worked hard at these problems.' The third group was controlled and given no feedback. Students were subsequently given measures that rated persistence, enjoyment, quality of performance and failure attributions. Additionally, they were administered a second assessment (similar to the one that they had received praise on) of similar difficulty.

Analysis 4:

Issue addressed: Do children who receive different types of praise (ability, effort, or none) differ in how harshly they judge their performance?

Statistical Significance Information:

Difference between groups:

$F(2, 48) = 2.04$, ns

Relevant Results/Conclusions:

No significant differences were noted for children's expectations; children in the intelligence, effort, and control conditions displayed equivalent expectations. (p.40)

These results indicate that effort praise and intelligence praise do not lead children to judge their performance differently

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.2828

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.0

Upper Cohen's f: 0.5366

Appendix B

Article Reviewer Instructions and Cases

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 52

Synopsis of Study: The purpose of this study was to determine if the types of praise given to children impacted their motivation and performance. Children were placed in three groups. In the two experimental groups, children were given different types of praise for accomplishments. The first group was praised on ability and children wer told 'You must be smart at these problems' and the second group was praised on effort, 'You must have worked hard at these problems.' The third group was controlled and given no feedback. Students were subsequently given measures that rated persistence, enjoyment, quality of performance and failure attributions. Additionally, they were administered a second assessment (similar to the one that they had received praise on) of similar difficulty.

Analysis 5:

Issue addressed: Do children who receive different types of praise (ability, effort, or none) differ regarding persistence?

Statistical Significance Information:

Difference between groups:

$F(2, 45) = 3.16, p = .05$

Follow up groups comparisons:

Ability vs. Effort, $t(30) = -2.09, p<.05$

Ability vs. Control, $t(30) = -2.22, p<.05$

Control vs. Effort, $t(30) = -0.12$, ns

Relevant Results/Conclusions:

Children praised for intelligence were less likely to want to persist on the problems after setbacks than were children praised for effort; children in the control condition closely resembled those in the effort conditions. Follow-up t-tests revealed significant differences between the intelligence condition and the effort and control conditions but no difference between the effort and control conditions. (p.46)

Indictment of ability also led children praised for intelligence to display mor negative responses in terms of lower levels of task persistence than their counterparts, who received commendations for effort. (p.48)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.3707

Calculated Effect Size (Cohen's t): -0.7332

Calculated Effect Size (Cohen's t): -0.7777
Calculated Effect Size (Cohen's t): -0.0412

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)
Lower Cohen's f: 0.0
Upper Cohen's f: 0.6462
Calculated Confidence Interval (95%)
Lower Cohen's t: -0.0055
Upper Cohen's t: -1.4609
Calculated Confidence Interval (95%)
Lower Cohen's t: -0.0472
Upper Cohen's t: -1.0582
Calculated Confidence Interval (95%)
Lower Cohen's t: 0.7570
Upper Cohen's t: 0.-.6746

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 1

Synopsis of Study: The purpose of this study was to determine if differences in course delivery mode (on-campus vs distance learning) of college courses impacted student perceptions/satisfaction of the course in aspects of instructor, organization, teaching, and communication. Student in two graduate level special education courses delivered in both modes responded to surveys administered measuring satisfaction with course

Analysis 1:

Issue addressed: Do ?

Statistical Significance Information:

Overall Satisfaction:

$t(25) = -0.81, p>.01$, ns

Relevant Results/Conclusions:

No differences were evident in overall ratings. Students' overall perceptions of the course were similar when the course was taught on campus or off campus with distance education technologies. (p.46)

As evidenced by this research, data on outcomes of distance learning experiences are favorable. Within the context expanded by data on such issues, the promises of technology-improved distance learning experiences will be realized and education for all students will be greatly enhanced.

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's t): -0.6740

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's t: -1.7509

Upper Cohen's t: 0.4030

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 12

Synopsis of Study: The purpose of this study was to determine if differences in students' time-of-day preferences impacted their performance on an algebra test. A measure of student's time-of-day preference (morning or afternoon) was obtained and the test was administered to members of both groups during morning and afternoon (not the same students).

Analysis 1:

Issue addressed: Do student's who have different preferences (morning or afternoon) perform differently if they take the test in the morning?

Statistical Significance Information:

Difference between groups:

$F(1,64) = 5.44, p < .05$

Relevant Results/Conclusions:

There was a significant difference between afternoon-preferenced students and morning-preferenced student taking the test in the morning. (p.298)

The results indicate clearly that the time-of-day element in learning stule may play a signficacnt part in the instructional environment. When time preference and testing environment were matched, significant differences emerged between test results—but only for the morning test (p. 298)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.2849

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.0024

Upper Cohen's f: 0.5283

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 12

Synopsis of Study: The purpose of this study was to determine if differences in students' time-of-day preferences impacted their performance on an algebra test. A measure of student's time-of-day preference (morning or afternoon) was obtained and the test was administered to members of both groups during morning and afternoon (not the same students).

Analysis 2:

Issue addressed: Do student's who have different preferences (morning or afternoon) perform differently if they take the test in the afternoon?

Statistical Significance Information:

Difference between groups:

$F(1,64) = 3.81, p < .055$

Relevant Results/Conclusions:

There was a small difference between afternoon-preferenced students and morning-preferenced student taking the test in the afternoon. (p.298)

The results indicate clearly that the time-of-day element in learning stule may play a signficacnt part in the instructional environment. When time preference and testing environment were matched, significant differences emerged between test results—but only for the morning test (p. 298)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): 0.2385

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.0

Upper Cohen's f: 0.4805

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 76

Synopsis of Study: The purpose of this study was to determine if the type of supervision pre-service teachers experienced impacted their development of clarity skills, pedagogical reasoning and actions, and attitudes toward several aspects of their field experience. Pre-service teachers were assigned either to the experimental group which engaged in peer coaching techniques or to the control group which experienced traditional mentoring experiences.

Analysis 1:

Issue addressed: Do student's who have different supervision experiences have different attitudes toward their experience upon completion?

Statistical Significance Information:

Difference between groups on overall measure:

$T(30) = .67, p > .51$

Relevant Results/Conclusions:

We did not find statistical significance for the overall rating.(p.260)

Evidence presented here indicates that peer coaching is a feasible vehicle for institutitng collaborative efforts; therefore, peer coaching warrants consideration as a potentially serviceable solution for strengthening field-based training of prospective teachers (p.261)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's d): -.7929

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's d: -.2840

Upper Cohen's d: -1.3018

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 76

Synopsis of Study: The purpose of this study was to determine if the type of supervision pre-service teachers experienced impacted their development of clarity skills, pedagogical reasoning and actions, and attitudes toward several aspects of their field experience. Pre-service teachers were assigned either to the experimental group which engaged in peer coaching techniques or to the control group which experienced traditional mentoring experiences.

Analysis 2:

Issue addressed: Do pre-service teachers who have different supervision experiences demonstrate differences in clarity skills

Statistical Significance Information:

Difference between groups on overall measure:

$f(1, 30) = 41.66, p < .001$

Relevant Results/Conclusions:

Posttreatment results showed statistically significant differences in favor of the experimental group for overall demonstration of clarity skills.(p.260)

Evidence presented here indicates that peer coaching is a feasible vehicle for institutitng collaborative efforts; therefore, peer coaching warrants consideration as a potentially serviceable solution for strengthening field-based training of prospective teachers (p.261)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): .8068

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.5213

Upper Cohen's f: 1.0874

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 78

Synopsis of Study: The purpose of this study was to determine if participation by families in a literary intervention project helped their young student's gain literacy skills. Parents and families participated in a monthly training session for five months to provide them with skills and materials to help their kindergarten age children with literacy skills. Gains on various measure were compared with gains by children in the same schools and classes that did not participate in the program.

Analysis 1:

Issue addressed: Is the family intervention program effective in helping children gain vocabulary skills?

Statistical Significance Information:

Difference between groups on overall measure across time:

$f(1, 247) = 32.08, p < .001$

Relevant Results/Conclusions:

When examining the effect of the interaction of group affiliation with time using repeated measures ANOVA we found that Project EASE participants made statistically significantly greater gains than the control group on Vocabulary..(p.532)

It appeared from the posttest measures on the CAP vocabulary subtests that those students who participated in the intervention were better able to recall more superordinate terms, which in turn have been shown to relate to the reading skills of elementary aged children. (p. 538)

Because vocabulary knowledge, story comprehension, and story sequencing are precisely the language skills that relate most strongly to literacy accomplishments (citation), the improvement on these measures strongly confirms the relevance of the intervention to improved reading outcomes.(p.539)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): .3597

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.2309

Appendix B

Article Reviewer Instructions and Cases

Upper Cohen's f: 0.4878

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 78

Synopsis of Study: The purpose of this study was to determine if participation by families in a literary intervention project helped their young student's gain literacy skills. Parents and families participated in a monthly training session for five months to provide them with skills and materials to help their kindergarten age children with literacy skills. Gains on various measure were compared with gains by children in the same schools and classes that did not participate in the program.

Analysis 2:

Issue addressed: Is the family intervention program effective in helping children gain sound awareness skills?

Statistical Significance Information:

Difference between groups on overall measure across time:

$f(1, 247) = 7.45, p < .01$

Relevant Results/Conclusions:

When examining the effect of the interaction of group affiliation with time using repeated measures ANOVA we found that Project EASE participants made statistically significantly greater gains than the control group on Sound Awareness.(p.532)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): .1733

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.0474

Upper Cohen's f: 0.2985

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 78

Synopsis of Study: The purpose of this study was to determine if participation by families in a literary intervention project helped their young student's gain literacy skills. Parents and families participated in a monthly training session for five months to provide them with skills and materials to help their kindergarten age children with literacy skills. Gains on various measure were compared with gains by children in the same schools and classes that did not participate in the program.

Analysis 3:

Issue addressed: Is the family intervention program effective in helping children gain story comprehension skills?

Statistical Significance Information:

Difference between groups on overall measure across time:

$f(1, 229) = 6.85, p < .01$

Relevant Results/Conclusions:

When examining the effect of the interaction of group affiliation with time using repeated measures ANOVA we found that Project EASE participants made statistically significantly greater gains than the control group on Story Comprehension.(p.532)

The impact of participation in Project EASE on children's language scores is striking. (p. 537)

Because vocabulary knowledge, story comprehension, and story sequencing are precisely the language skills that relate most strongly to literacy accomplishments (citation), the improvement on these measures strongly confirms the relevance of the intervention to improved reading outcomes.

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): .1874

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.0448

Upper Cohen's f: 0.3288

Appendix B

Article Reviewer Instructions and Cases

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 78

Synopsis of Study: The purpose of this study was to determine if participation by families in a literary intervention project helped their young student's gain literacy skills. Parents and families participated in a monthly training session for five months to provide them with skills and materials to help their kindergarten age children with literacy skills. Gains on various measure were compared with gains by children in the same schools and classes that did not participate in the program.

Analysis 4:

Issue addressed: Is the family intervention program effective in helping children gain language skills?

Statistical Significance Information:

Difference between groups on overall measure across time:

$f(1, 246) = 35.46, p < .001$

Relevant Results/Conclusions:

Although all the children in the sample showed statistically significant gains in all three literacy composites over time, we were able to attribute a statistically significant gain in Language skills to the Project EASE intervention. (p.532)

The impact of participation in Project EASE on children's language scores is striking. (p. 537)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's f): .3789

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's f: 0.2494

Upper Cohen's f: 0.5077

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 73

Synopsis of Study: The purpose of this study was to determine if praise impacted the amount of time college students' spent on homework. Additionally, it was investigated if praise impacted achievement. Students maintained a log of time spent on homework and were either placed into the 'praised' group (when receiving the log, the instructor momentarily reviewed and told the student 'good job', 'very good', or 'great work') or were in the 'non-praised' group…these students' were merely thanked when they turned in their log. At the end of the course, the average amount of time spent on homework for 17 randomly selected homework assignments was calculated and compared, as well as performance on an instructor-created final examination.

Analysis 1:

Issue addressed: Does praise impact the amount of time spent on homework?

Statistical Significance Information:

Difference between groups:

$t(59) = 9.788, p < .001$

Relevant Results/Conclusions:

Results revealed that students studied significantly more outside of the classroom when exposed to the verbal praise treatment than when exposed to the no verbal praise treatment. (p. 387)

Although the results of this study may not generalize to all college student populations, they demonstrate the profound impact of properly administered verbal praise on college students' motivation to engage in homework. (p. 388)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's d): 2.4881

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's d: 1.8196

Upper Cohen's d: 3.1566

Please answer item C on the review sheet

Appendix B

Article Reviewer Instructions and Cases

Study Number: 73

Synopsis of Study: The purpose of this study was to determine if praise impacted the amount of time college students' spent on homework. Additionally, it was investigated if praise impacted achievement. Students maintained a log of time spent on homework and were either placed into the 'praised' group (when receiving the log, the instructor momentarily reviewed and told the student 'good job', 'very good', or 'great work') or were in the 'non-praised' group…these students' were merely thanked when they turned in their log. At the end of the course, the average amount of time spent on homework for 17 randomly selected homework assignments was calculated and compared, as well as performance on an instructor-created final examination.

Analysis 2:

Issue addressed: Does praise on homework through the length of a course impact the performance on the end of course assessment?

Statistical Significance Information:

Difference between groups:

$t(59) = 1.929, p > 0.05$ ns

Relevant Results/Conclusions:

Although the difference was not statistically significant (on the end of course exam), the direction of the means suggested that the students exposed to verbal praise not only studied more for each lesson but also achieved more than those not exposed to verbal praise. (p. 387)

In addition, my findings suggest that students who experience verbal praise for doing homework perform somewhat better on an instructor-created, criterion-referenced final examination than those who experience no verbal praise for their homework habits. (p. 388)

Before continuing, please answer item A on the review sheet

Calculated Effect Size (Cohen's d): .4800

Before continuing, please answer item B on the review sheet

Calculated Confidence Interval (95%)

Lower Cohen's d: -.0292

Upper Cohen's d: .9891

Please answer item C on the review sheet

Appendix C

SAS Code

```
proc printto print='C:\Cohen_ci.lst';
* +-----------------------------------------------------------------+
  This program calculates confidence bands for two group effect size
  (Cohen's d) using both an interval inversion approach through the
  macro at the beginning and then using z-bands.
  This first part calculates endpoints using Steiger

  Raw values are input about midway through program for two group
  Ns, means and std deviations.  Depending on data provided, these
  inputs might need to be modified.

  Last modification:  4 Sept 2003
  +-----------------------------------------------------------------+;
* +-------------------------------------------------------------------
---+
   Input to the macro:
        data = name of data set
            effect_size = obtained sample value of Cohen d
            n1 = sample size of group one
            n2 = sample size of group two
   Output is printed table of confidence intervals
* +-------------------------------------------------------------------
---+;
%macro EFFECT_CI(data, effect_size, n1, n2);

proc iml;

start find_delta(obs_stat, n1, n2, pctl, delta_t);

  df = n1 + n2 - 2;
  * Step 1: Find value of delta that is a little too high;

  OK = 0;
  delta_t = 0; * start the loop with population effect size = 0;
  loop = 0;
  do until (OK = 1);
     nc = delta_t # sqrt(n1#n2/(n1+n2));
     cumprob = PROBT(obs_stat,df,nc);
     if cumprob<pctl then OK = 1;
     if cumprob>pctl then delta_t = delta_t + .1;
     loop = loop + 1;
     if loop > 1500 then do;
         print 'Looping too much!' loop delta_t nc obs_stat cumprob;
         OK = 1;
     end;
  END;
  * print 'Estimating High' delta_t nc obs_stat pctl cumprob ok
not_poss;

  high = delta_t;

  * Step 2: Find value of delta that is a little too low;

  OK = 0;
```

SAS Code

```
  delta_t = 0; * start the loop with population effect size = 0;
  loop = 0;
  do until (OK = 1);
     nc = delta_t # sqrt(n1#n2/(n1+n2));
     cumprob = PROBT(obs_stat,df,nc);
     if cumprob>pctl then OK = 1;
     if cumprob<pctl then delta_t = delta_t - .1;
     loop = loop + 1;
     if loop > 1500 then do;
         print 'Looping too much!' loop delta_t nc obs_stat cumprob;
         OK = 1;
     end;
  END;
  * print 'Estimating Low' delta_t nc obs_stat pctl cumprob ok not_poss;

  low = delta_t;

  * Step 3: Successively halve the interval between low and high
            to obtain final value of percentile;

      change = 1;
      loop = 0;
      small = .00000000001;
      do until (change<small);
       half = (high + low)/2;
       nc = half # sqrt(n1#n2/(n1+n2));
       cum_h = PROBT(obs_stat,df,nc);
       if cum_h < pctl then high =  half; * still too high;
       if cum_h > pctl then low  = half;  * still too low;
       change = abs(high - low);
       loop = loop + 1;
       if loop > 1500 then small = .000000001;
       if loop > 3000 then small = .01;
          * print high low change;
       Delta_t = (high + low)/2;
              * print Delta_t;
      end;
finish;

use &data;
 read all var{&effect_size} into effect_vec;
 read all var{&n1} into n1;
 read all var{&n2} into n2;
 k = nrow(effect_vec);
  file print;
  put @1 'Confidence Intervals Around Sample cohen d steiger and
fouladi' //
      @16 '99% CI' @36 '95% CI' @56 '90% CI' /
      @2 'Effect' @10 '-------------------' @30 '-------------------'
@50 '-------------------' /
      @3 'Size' @12 'Lower     Upper' @32 'Lower     Upper' @52 'Lower
Upper' /
       @1 '--------' @10 '--------- ---------' @30 '--------- ---------
' @50 '--------- ---------';
```

Appendix C

SAS Code

```
do i = 1 to k;

  obs_stat = effect_vec[i,1] # sqrt(n1[i,1]#n2[i,1]/(n1[i,1]+n2[i,1]));

  run find_delta(obs_stat, n1[i,1], n2[i,1], .005, delta005);
  run find_delta(obs_stat, n1[i,1], n2[i,1], .995, delta995);
  run find_delta(obs_stat, n1[i,1], n2[i,1], .025, delta025);
  run find_delta(obs_stat, n1[i,1], n2[i,1], .975, delta975);
  run find_delta(obs_stat, n1[i,1], n2[i,1], .05, delta05);
  run find_delta(obs_stat, n1[i,1], n2[i,1], .95, delta95);

  print_effect = effect_vec[i,1];
  file print;
  put @1 print_effect 8.4 @10 delta995 8.4 @20 delta005 8.4 @30 delta975
8.4 @40 delta025 8.4 @50 delta95 8.4 @60 delta05 8.4;
end;
quit;

%mend EFFECT_CI;


data one;
input journ $ article analysis $ n1 n2 mn1 mn2 sd1 sd2;


nsample1 = n1;
   nsample2 = n2;
   d = 0;
   vard = 0;

width_z_99  = 0;
width_z_95  = 0;
width_z_90  = 0;
lo_z_99 = 0;
hi_z_99 = 0;
lo_z_95 = 0;
hi_z_95 = 0;
lo_z_90 = 0;
hi_z_90 = 0;

    * +----------------------------------+
      Compute sample means and variances
    * +----------------------------------+;
      n1 = n1;
      n2 = n2;
      mn1 = mn1;
      mn2 = mn2;
      var1 = sd1**2;
      var2 = sd2**2;

    * +----------------------------------------+
      Compute sample value of d and its variance
    * +----------------------------------------+;
      d = (mn1- mn2)/(((((n1-1)*var1) + ((n2 -1)*var2)) / (n1 + n2 -
2))**0.5);
```

```
      vard = ((n1 + n2)/(n1 * n2)) + d**2/ (2*(n1 + n2))  ;


  *   +-------------------------------------------------+
       Compute endpoints of CI using normal distribution
    * +-------------------------------------------------+;
       lo_z_99 = d - (2.576*sqrt(vard));
       hi_z_99 = d + (2.576*sqrt(vard));
       lo_z_95 = d - (1.96*sqrt(vard));
       hi_z_95 = d + (1.96*sqrt(vard));
       lo_z_90 = d - (1.645*sqrt(vard));
       hi_z_90 = d + (1.645*sqrt(vard));


   * +-----------------------+
      Normal Z Bands
     +-----------------------+;
        width_z_99 = width_z_99 + (hi_z_99 - lo_z_99);
        width_z_95 = width_z_95 + (hi_z_95 - lo_z_95);
        width_z_90 = width_z_90 + (hi_z_90 - lo_z_90);


* +-----------------------------------------------------------------+
 just computing sample delta
* +-----------------------------------------------------------------+;


width_z_99  = 0;
width_z_95  = 0;
width_z_90  = 0;
lo_z_99 = 0;
hi_z_99 = 0;
lo_z_95 = 0;
hi_z_95 = 0;
lo_z_90 = 0;
hi_z_90 = 0;

    * +---------------------------------+
       Compute sample means and variances
    * +---------------------------------+;
       n1 = n1;
       n2 = n2;
       mn1 = mn1;
       mn2 = mn2;
       var1 = 6.93**2;
       var2 = 5.71**2;

    * +---------------------------------------+
       Compute sample value of d and its variance
    * +---------------------------------------+;
       d = (mn1- mn2)/(((((n1-1)*var1) + ((n2 -1)*var2)) / (n1 + n2 -
2))**0.5);
       vard = ((n1 + n2)/(n1 * n2)) + d**2/ (2*(n1 + n2))  ;
```

Appendix C

SAS Code

```
  *   +-------------------------------------------------+
       Compute endpoints of CI using normal distribution
    * +-------------------------------------------------+;
       lo_z_99 = d - (2.576*sqrt(vard));
       hi_z_99 = d + (2.576*sqrt(vard));
       lo_z_95 = d - (1.96*sqrt(vard));
       hi_z_95 = d + (1.96*sqrt(vard));
       lo_z_90 = d - (1.645*sqrt(vard));
       hi_z_90 = d + (1.645*sqrt(vard));


   * +----------------------+
      Normal Z Bands
     +----------------------+;
        width_z_99 = width_z_99 + (hi_z_99 - lo_z_99);
        width_z_95 = width_z_95 + (hi_z_95 - lo_z_95);
        width_z_90 = width_z_90 + (hi_z_90 - lo_z_90);
*  If journals are coded by:
    1:  Research Reading Quarterly
    2:  Journal of Educational Research
    3:  Journal of Personality and Social Psychology;

diff = mn1-mn2;

vardiff = ((n1 + n2)/(n1 * n2)) + diff**2/ (2*(n1 + n2)) ;

       crit_t99 = TINV(.995,n1+n2-2,0);
       crit_t95 = TINV(.975,n1+n2-2,0);
       crit_t90 = TINV(.95,n1+n2-2,0);

       lo_t_99 = diff - (crit_t99*sqrt(vardiff));
       hi_t_99 = diff + (crit_t99*sqrt(vardiff));
       lo_t_95 = diff - (crit_t95*sqrt(vardiff));
       hi_t_95 = diff + (crit_t95*sqrt(vardiff));
       lo_t_90 = diff - (crit_t90*sqrt(vardiff));
       hi_t_90 = diff + (crit_t90*sqrt(vardiff));
         width_t_99 = hi_t_99 - lo_t_99;
         width_t_95 = hi_t_95 - lo_t_95;
         width_t_90 = hi_t_90 - lo_t_90;


cards;
JER 1 A 4 23 3.69 3.94 .59 .33
JER 1 B 4 23 3.56 3.88 .33 .31
JER 1 C 4 23 3.65 3.88 .59 .34
JER 1 D 4 23 4.15 4.23 .34 .17
JER 1 E 4 23 3.48 3.62 .47 .44
JER 1 F 4 23 3.49 3.79 .48 .27
JER 1 G 11 13 3.69 3.79 .28 .44
JER 1 H 11 13 3.72 3.60 .29 .22
JER 1 I 11 13 3.65 3.65 .19 .43
JER 1 J 11 13 3.83 4.25 .19 .10
JER 1 K 11 13 3.58 3.42 .44 .23
JER 1 L 11 13 3.56 3.63 .13 .40
```

SAS Code

```
JER 5 A 55 15 4.79 4.21 .50 .81
JER 5 B 55 15 4.71 4.07 .54 .83
JER 5 C 55 15 4.46 4.13 .75 .83
JER 5 D 55 15 4.36 3.93 .88 1.16
JER 5 E 55 15 4.04 4.27 .87 .70
JER 5 F 55 15 3.92 3.72 .96 1.58
JER 5 G 55 15 4.02 3.29 1.18 1.27
JER 5 H 55 15 3.06 3.40 1.07 1.12
JER 5 I 55 15 2.42 3.13 1.32 1.13
JER 5 J 55 15 2.17 3.47 1.20 1.36
JER 76 D 32 32 3.75 4.50 1.18 .63
JER 76 E 32 32 3.87 4.75 1.02 .45
JER 76 F 32 32 4.31 4.88 .87 .34
JER 76 G 32 32 4.31 4.80 .70 .41
JER 76 H 32 32 4.69 4.56 .80 .73
JER 76 I 32 32 4.56 4.75 1.03 .45
JER 4 A 247 247 27.56 26.84 9.45 9.68
JER 4 B 149 149 31.10 30.76 8.16 8.47
JER 4 C 98 98 22.18 20.87 8.74 8.27
JER 4 D 94 94 22.27 21.57 9.03 9.31
JER 4 E 45 45 34.51 34.53 7.62 8.41
JER 4 F 33 33 24.94 24.52 7.36 7.05
JER 4 G 59 59 20.48 19.07 9.23 8.49
JER 4 H 35 35 25.31 25.78 7.92 9.22
JER 4 I 35 35 29.50 28.36 6.76 5.86
JER 4 J 35 35 36.75 37.07 6.32 6.55
JER 73 A 30 31 34.7 46.8 5.3 4.4
JER 73 B 30 31 83.5 86.0 5.6 4.8
JPSP 63 M 54 41 4.27 2.05 3.17 1.69
JPSP 63 N 54 41 4.89 6.02 1.84 2.62
JPSP 63 Y 146 146 3.27 -.96 .85 1.59
JPSP 52 D 38 39 11.96 4.94 8.15 7.04
JPSP 52 E 46 38 10.58 11.96 8.43 8.15
JPSP 52 F 39 38 16.49 9.78 11.04 9.00
JPSP 52 G 46 39 13.88 16.49 9.18 11.04
JPSP 52 H 39 38 3.25 4.53 1.41 1.03
JPSP 52 I 39 46 3.25 4.30 4.41 1.33
JPSP 52 J 38 46 4.53 4.30 1.03 1.33
JPSP 52 L 39 38 4.11 4.89 1.02 .72
JPSP 52 M 39 46 4.11 4.52 1.02 0.81
JPSP 52 N 38 46 4.89 4.52 .72 .81
JPSP 52 P 39 38 -.92 1.21 1.53 1.63
JPSP 52 Q 39 46 -.92 .13 1.53 1.57
JPSP 52 R 38 46 1.21 .13 1.63 1.57
JPSP 52 AA 30 29 14.83 4.70 7.70 3.43
JPSP 52 AB 30 29 14.83 7.97 7.70 4.87
JPSP 52 AC 29 29 4.70 7.97 3.43 4.87
JPSP 52 AD 29 30 19.79 7.70 7.18 6.20
JPSP 52 AE 29 29 19.79 12.28 7.18 7.43
JPSP 52 AF 30 29 7.70 12.28 6.20 7.43
JPSP 52 AH 29 30 3.24 5.20 .83 1.00
JPSP 52 AI 29 29 3.24 4.28 .83 1.29
JPSP 52 AJ 30 29 5.20 4.28 1.00 1.29
JPSP 52 AL 29 30 3.86 4.99 1.01 .55
JPSP 52 AM 29 29 3.86 4.49 1.01 .94
JPSP 52 AN 30 29 4.99 4.49 .55 .94
```

Appendix C

SAS Code

```
JPSP 52 AQ 29 30 -.37 1.23 1.42 1.50
JPSP 52 AR 29 29 -.37 .34 1.42 2.13
JPSP 52 AS 30 29 1.23 .34 1.50 2.13
JPSP 52 AZ 17 17 4.24 2.19 1.79 1.52
JPSP 52 BA 17 17 2.19 3.47 1.52 2.24
JPSP 52 BB 17 17 4.24 3.46 1.79 2.24
JPSP 52 BE 15 16 20.06 7.13 11.32 5.52
JPSP 52 BF 15 15 20.06 10.06 11.32 6.79
JPSP 52 BG 16 15 7.13 10.06 5.52 6.79
JPSP 52 BH 16 15 20.94 7.75 7.17 9.50
JPSP 52 BI 16 15 20.94 12.06 7.17 8.06
JPSP 52 BJ 15 15 7.75 12.06 9.50 8.06
JPSP 52 BL 16 15 3.44 4.62 1.59 1.63
JPSP 52 BM 16 15 3.44 4.56 1.59 1.26
JPSP 52 BN 15 15 4.62 4.56 1.63 1.26
JPSP 52 BP 16 15 3.92 5.19 .95 .82
JPSP 52 BQ 16 15 3.92 4.90 .95 .93
JPSP 52 BR 15 15 5.19 4.90 .82 .95
JPSP 52 BV 16 16 20.81 7.25 9.42 5.34
JPSP 52 BW 16 16 20.81 5.75 9.42 4.92
JPSP 52 BX 16 16 7.25 5.75 5.34 4.92
JPSP 52 BZ 16 16 16.94 7.13 9.74 6.48
JPSP 52 CA 16 16 16.94 13.31 9.74 8.67
JPSP 52 CB 16 16 7.13 13.31 6.48 8.67
JPSP 52 CE 16 16 3.84 4.86 .74 .88
JPSP 52 CF 16 16 3.84 4.41 .74 .80
JPSP 52 CG 16 16 4.86 4.41 .88 .80
JPSP 52 CK 16 16 4.38 6.81 2.16 2.23
JPSP 52 CL 16 16 6.81 4.94 2.23 1.84
JPSP 52 CM 16 16 4.38 4.94 2.16 1.84
JPSP 52 CP 16 16 4.13 2.56 1.20 1.44
JPSP 52 CQ 16 16 4.13 2.94 1.20 1.84
JPSP 52 CR 16 16 2.56 2.94 1.44 1.84
JPSP 69 D 17 17 9.24 12.35 4.04 2.62
JPSP 69 E 21 20 9.76 9.35 3.48 3.01
JPSP 69 I 22 22 5.27 4.45 2.07 2.13
JPSP 69 J 23 23 3.30 4.17 2.32 2.23
JPSP 69 K 23 22 5.89 4.51 1.01 1.23
JPSP 69 L 23 22 4.02 2.86 1.89 1.68
JPSP 58 A 64 64 4.49 2.25 .50 .74
JPSP 58 B 64 64 4.48 3.01 .50 .68
JPSP 58 C 64 64 4.12 2.32 .59 .66
JPSP 58 D 64 64 3.98 3.35 .83 .71
JPSP 58 E 64 64 3.56 2.94 .74 .87
rrq 18 A 23 26 3.6 2.6 1.3 1.4
rrq 49 A 23 26 59.6 53.7 5.95 12.4
rrq 49 H 23 26 29.8 23.2 5.8 8.2
rrq 49 I 23 26 3.6 2.6 1.3 1.4
JER 3 A 1036 1131 2.52 2.58 1.01 1.07
JER 3 B 1036 1131 2.83 2.91 .91 .91
JER 3 C 1036 1131 2.28 2.29 1.12 1.16
JER 3 D 1036 1131 3.07 3.10 .88 .89
JER 3 E 1036 1131 1.98 2.06 1.15 1.15
JER 3 F 1036 1131 2.24 2.41 1.07 1.10
JER 3 G 1036 1131 2.44 2.50 1.05 1.06
JER 3 H 1036 1131 2.21 2.15 1.16 1.21
```

```
JER 3 I 1036 1131 2.37 2.28 1.05 1.16
JER 3 J 1036 1131 1.29 1.17 1.28 1.24
JER 3 K 1036 1131 1.83 1.80 1.24 1.25
JER 3 L 1036 1131 1.46 1.45 .57 .60
JER 3 M 1036 1131 1.58 1.58 .54 .55
JER 3 N 1036 1131 1.27 1.23 .67 .67
JER 3 O 1036 1131 1.50 1.50 .56 .58
JER 3 P 1036 1131 1.11 1.12 .71 .69
JER 3 Q 1036 1131 1.26 1.31 .60 .60
JER 3 R 1036 1131 1.42 1.41 .62 .62
JER 3 S 1036 1131 1.36 1.36 .60 .60
JER 3 T 1036 1131 1.43 1.35 .61 .65
JER 3 U 1036 1131  .78  .69 .76 .74
JER 3 V 1036 1131 1.12 1.08 .67 .68
JER 3 W 1036 1131 2.04 2.10 .75 .75
JER 3 X 1036 1131 2.06 2.10 .76 .76
JER 3 Y 1036 1131 1.54 1.52 .99 1.01
JER 3 Z 1036 1131 2.46 2.43 .75 .79
JER 3 AA 1036 1131 1.72 1.74 .90 .93
JER 3 AB 1036 1131 1.59 1.56 .91 .94
JER 3 AC 1036 1131 1.97 1.95 .88 .90
JER 3 AD 1036 1131 1.98 1.95 .81 .81
JER 3 AE 1036 1131 1.98 1.86 .91 .92
JER 3 AF 1036 1131 1.11 1.02 1.10 1.02
JER 3 AG 1036 1131 1.43 1.37 1.01 .98
;
  * The following calls the macro for Interval Inversion;


      %EFFECT_CI(one, d, n1, n2);

PROC FREQ;
 TABLES JOURN * ARTICLE;
title1 'Cohen d Confidence Intervals z transformation';
proc print;
var journ article analysis n1 n2 mn1 mn2 var1 var2 vard d hi_z_99
lo_z_99 d hi_z_95 lo_z_95 d hi_z_90 lo_z_90 d ;
*proc print;
 * var d lo_z_99 hi_z_99 lo_z_95 hi_z_95 lo_z_90 hi_z_90;
*proc print;
 *  var width_z_99 width_z_95 width_z_90;


title1 'Difference of Means Confidence Intervals by t-test';
*proc print;
*var journ article n1 n2 mn1 mn2 sd1 sd2 diff;
*proc print;
*var crit_t99 crit_t95 crit_t90;
proc print;
var journ article n1 n2 mn1 mn2 sd1 sd2 diff hi_t_99 lo_t_99 diff
hi_t_95 lo_t_95 diff hi_t_90 lo_t_90;
*proc print;
*var width_t_99 width_t_95 width_t_90;
```

```
run;
```

Appendix C

SAS Code

```
* +-----------------------------------------------------------------+
  This program calculates confidence bands for the effect size
  (Cohen's f) in ANOVA analyses using both an interval inversion
  approach and z transformation.

  Raw values are input about midway through program for total N,
  number of groups. and the F value obtained in the original analysis.
  Depending on data provided, these
  inputs might need to be modified.

  Last modification:  4 Sept 2003
  +-----------------------------------------------------------------+;
* +-----------------------------------------------------------------------
---+
   Input to subroutine:
        data = name of data set
            F_obt = obtained value of F
            N = sample size
            K = number of groups
        u = degrees of freedom numerator
        v = degress of freedom denominator
   Output is printed table of confidence intervals--at least I hope
someday :-)
* +-----------------------------------------------------------------------
---+;


data one;
input journ $ article analysis $ N k F_obt;

u = k-1 ;
v = N - k;
eta2=((k-1)*F_obt)/((k-1)*F_obt + N);
f=(eta2/(1-eta2))**.5;
loweta2_90 = 0;
loweta2_95 = 0;
loweta2_99 = 0;
higheta2_90 = 0;
higheta2_95 = 0;
higheta2_99 = 0;
widtheta2_90 = 0;
widtheta2_95 = 0;
widtheta2_99 = 0;
lowf_90 = 0;
lowf_95 = 0;
lowf_99 = 0;
highf_90 = 0;
highf_95 = 0;
highf_99 = 0;
widthf_90 = 0;
widthf_95 = 0;
widthf_99 = 0;
```

Appendix C

SAS Code

```
*++++++++++++++++++++++++++++++++++++++++++++
CalculatiNg the upper aNd lower bouNds of eta2
usiNg O&F_obt 3..called loweta2_95 aNd higheta2_95.
CurreNtly calculatioNs are oNly doNe usiNg the
95th perceNtile, peNdiNg resolutioN of method
++++++++++++++++++++++++++++++++++++++++++++;

z=log((1+sqrt(eta2))/(1-sqrt(eta2)));

loweta2_95 = z - ((2*(1.96))/(sqrt(N)));
higheta2_95 =  z + ((2*(1.96))/(sqrt(N)));

    low95 = exp(loweta2_95);
    high95 = exp(higheta2_95);
   loweta2_95 = ((low95-1)/(low95+1))**2;
    higheta2_95 = ((high95-1)/(high95+1))**2;

if loweta2_95<0 theN loweta2_95=0;
if higheta2_95>1 theN higheta2_95=1;
widtheta2_95 = higheta2_95-loweta2_95;

loweta2_99 = z - ((2*(2.576))/(sqrt(N)));
higheta2_99 =  z + ((2*(2.576))/(sqrt(N)));

    low99 = exp(loweta2_99);
    high99 = exp(higheta2_99);
   loweta2_99 = ((low99-1)/(low99+1))**2;
    higheta2_99 = ((high99-1)/(high99+1))**2;

if loweta2_99<0 theN loweta2_99=0;
if higheta2_99>1 theN higheta2_99=1;
widtheta2_99 = higheta2_99-loweta2_99;


loweta2_90 = z - ((2*(1.645))/(sqrt(N)));
higheta2_90 =  z + ((2*(1.645))/(sqrt(N)));

    low90 = exp(loweta2_90);
    high90 = exp(higheta2_90);
   loweta2_90 = ((low90-1)/(low90+1))**2;
    higheta2_90 = ((high90-1)/(high90+1))**2;

if loweta2_90<0 theN loweta2_90=0;
if higheta2_90>1 theN higheta2_90=1;
widtheta2_90 = higheta2_90-loweta2_90;



* ++++++++++++++++++++++++++++++++++++++
This set of CIs (called lowf_95 aNd
highf_95) are coNstructed by calculatiNg
f for the lower eta2 aNd upper eta2 calculated
earlier ...this method is the oNe more
appropriate???
```

SAS Code

```
++++++++++++++++++++++++++++++++++++++++;
lowf_95 = (loweta2_95/(1-loweta2_95))**.5;
highf_95 = (higheta2_95/(1-higheta2_95))**.5;
widthf_95 = highf_95-lowf_95;

lowf_99 = (loweta2_99/(1-loweta2_99))**.5;
highf_99 = (higheta2_99/(1-higheta2_99))**.5;
widthf_99 = highf_99-lowf_99;

lowf_90 = (loweta2_90/(1-loweta2_90))**.5;
highf_90 = (higheta2_90/(1-higheta2_90))**.5;
widthf_90 = highf_90-lowf_90;

Smpl_eta2 = eta2;

cards;
RRQ 78 A 229 2 .04
RRQ 78 B 229 2 .71
RRQ 78 C 248 2 .19
RRQ 78 D 248 2 32.08
RRQ 78 E 247 2 .72
RRQ 78 F 195 2 6.85
RRQ 78 G 248 2 4.80
RRQ 78 H 248 2 12.86
RRQ 78 I 229 2 8.52
RRQ 78 J 229 2 .56
RRQ 78 K 248 2 2.08
RRQ 78 L 248 2 7.45
RRQ 78 M 248 2 1.42
RRQ 78 N 247 2 .89
RRQ 78 O 195 2 .09
RRQ 78 P 248 2 .06
RRQ 78 Q 248 2 .03
RRQ 78 R 229 2 .57
RRQ 78 S 229 2 1.14
RRQ 78 T 248 2 .28
RRQ 78 U 248 2 .16
RRQ 78 V 248 2 .13
RRQ 78 W 248 2 1.53
RRQ 78 X 248 2 2.63
RRQ 78 Y 247 2 .81
RRQ 78 Z 247 2 8.13
RRQ 78 AA 247 2 1.59
RRQ 78 AB 247 2 35.46
RRQ 78 AC 247 2 3.69
RRQ 78 AD 247 2 1.92
RRQ 78 AE 247 2 0.00
RRQ 78 AF 247 2 .78
RRQ 32 B 58 2 5.85
RRQ 32 C 58 2 18.05
RRQ 32 D 58 2 2.43
RRQ 32 E 116 2 8.41
RRQ 32 F 116 2 3.13
RRQ 32 G 58 2 6.88
RRQ 32 H 58 2 7.61
RRQ 32 I 58 2 13.81
```

```
RRQ 32 J 58 2 10.05
RRQ 32 K 58 2 11.48
RRQ 32 L 58 2 7.79
RRQ 32 M 58 2 68.9
RRQ 32 N 58 2 4.02
RRQ 32 O 58 2 11.56
RRQ 32 P 58 2 14.88
RRQ 32 Q 58 2 90.93
RRQ 32 R 58 2 25.15
RRQ 32 S 58 2 4.20
RRQ 32 T 58 2 5.71
RRQ 32 U 58 2 4.20
RRQ 32 V 58 2 10.74
RRQ 32 W 58 2 11.99
RRQ 32 X 58 2 33.19
RRQ 32 Y 58 2 17.19
RRQ 32 Z 58 2 4.67
RRQ 32 AA 58 2 8.05
RRQ 35 A 158 3 15.10
RRQ 35 B 158 3 26.35
RRQ 35 C 158 3 27.10
RRQ 35 D 158 3 15.37
RRQ 35 E 158 2 15.19
RRQ 35 F 60 2 4.71
RRQ 35 G 60 2 6.99
RRQ 35 H 90 3 3.10
RRQ 35 I 90 3 6.59
RRQ 35 J 90 3 8.79
RRQ 35 K 90 3 9.71
RRQ 35 L 90 3 7.18
RRQ 35 M 90 3 9.17
RRQ 35 N 91 3 9.47
RRQ 35 O 91 3 7.18
RRQ 35 P 90 3 5.10
RRQ 35 Q 91 3 9.47
RRQ 35 R 91 3 4.86
RRQ 35 S 91 3 8.64
RRQ 35 T 91 2 5.88
RRQ 35 U 46 2 5.99
RRQ 35 V 46 2 10.72
RRQ 35 W 46 2 6.32
RRQ 35 X 46 2 5.50
RRQ 35 Y 46 2 10.69
RRQ 35 Z 139 3 6.9
RRQ 35 AA 85 2 4.8
RRQ 35 AB 140 3 9.3
RRQ 35 AC 86 2 4.9
RRQ 35 AD 140 3 13.3
RRQ 35 AE 86 2 5.9
RRQ 35 AF 139 3 10.0
RRQ 35 AG 85 2 8.9
RRQ 35 AH 140 3 49.2
RRQ 35 AI 140 3 38.5
RRQ 35 AJ 140 3 38.5
RRQ 35 AK 53 2 27.0
RRQ 35 AL 53 2 58.1
```

SAS Code

```
RRQ 35 AM 53 2 53.8
RRQ 35 AN 53 2 29.2
RRQ 35 AO 53 2 10.9
RRQ 35 AP 53 2 64.5
RRQ 35 AQ 53 2 36.9
RRQ 35 AR 53 2 50.0
RRQ 35 AS 53 2 6.3
RRQ 47 A 88 2 284.09
RRQ 47 B 88 3 3.61
RRQ 47 C 88 2 3.89
RRQ 47 D 88 3 57.02
RRQ 47 E 88 3 10.26
RRQ 47 F 88 2 14.10
RRQ 47 G 88 2 428.82
RRQ 47 H 88 3 3.73
RRQ 47 I 88 3 6.22
RRQ 47 J 88 3 32.43
RRQ 47 K 88 3 32.43
RRQ 47 L 88 2 374.57
RRQ 47 M 88 2 32.11
RRQ 47 N 88 3 6.51
RRQ 47 O 88 3 5.47
RRQ 47 P 88 2 329.66
RRQ 47 Q 88 3 6.23
RRQ 47 R 88 2 136.73
RRQ 47 S 88 2 7.60
RRQ 47 T 88 3 9.23
RRQ 47 U 88 2 178.00
RRQ 47 V 88 2 700.61
RRQ 47 W 88 3 9.14
RRQ 47 X 88 2 8.42
RRQ 47 Y 88 3 3.90
RRQ 47 Z 88 3 21.24
RRQ 47 AA 88 2 620.89
RRQ 47 AB 88 2 20.61
RRQ 47 AC 88 3 11.64
RRQ 47 AD 88 3 6.14
RRQ 47 AE 88 2 6.97
RRQ 47 AF 88 3 9.99
RRQ 47 AG 88 3 27.87
RRQ 47 AH 88 2 45.16
RRQ 47 AI 88 2 33.65
RRQ 47 AJ 88 3 21.85
RRQ 47 AK 88 2 7.63
RRQ 47 AL 88 3 8.06
RRQ 48 A 118 3 31.6
RRQ 42 A 151 4 124.81
RRQ 42 B 151 12 1.73
RRQ 42 C 151 4 2.90
RRQ 46 A 91 3 4.57
RRQ 46 B 91 3 113.5
RRQ 46 C 91 3 89.29
RRQ 46 D 91 3 73.99
RRQ 46 E 91 3 113.26
RRQ 46 F 91 3 62.09
RRQ 79 A 83 2 4.72
```

SAS Code

```
RRQ 79 B 83 2 16.72
RRQ 79 C 83 2 6.27
RRQ 79 D 83 2 5.09
RRQ 79 E 83 2 53.66
RRQ 79 F 83 2 16.42
RRQ 79 G 83 2 21.78
RRQ 79 H 83 2 8.55
RRQ 79 I 83 2 52.98
RRQ 79 J 83 2 9.83
RRQ 79 K 83 2 48.03
RRQ 79 L 83 3 17.68
RRQ 79 M 83 3 26.29
RRQ 79 N 83 3 74.26
RRQ 79 O 83 3 92.84
RRQ 79 P 83 2 78.81
RRQ 79 Q 83 2 11.23
RRQ 79 R 83 2 182.44
RRQ 79 S 83 3 97.11
RRQ 79 T 83 2 8.91
RRQ 79 U 83 2 3.40
RRQ 79 W 83 2 29.3
RRQ 79 Z 83 3 15.58
RRQ 77 A 71 3 .02
RRQ 77 B 71 3 1.80
RRQ 77 C 71 3 68.84
RRQ 77 D 71 3 46.72
RRQ 77 E 71 3 16.38
RRQ 77 F 71 3 9.72
RRQ 77 G 71 3  202.44
RRQ 77 H 71 3  8.48
RRQ 77 I 71 3 14.05
RRQ 77 J 71 3 29.95
RRQ 77 L 71 3 13.32
JER 76 A 64 2 23.71
JER 76 B 64 2 49.77
JER 76 C 64 2 41.66
JER 12 A 67 2 .16
JER 12 B 67 2 8.95
JER 12 C 67 2 9.23
JER 12 D 67 2 13.81
JER 12 E 67 2 0
JER 12 F 67 2 5.44
JER 12 G 67 2 3.81
JER 12 H 74 2 9.90
JER 12 I 74 2 6.25
JER 12 J 74 2 8.25
JER 12 K 74 2 18.32
JER 12 L 74 2 0.04
JER 12 M 74 2 10.27
JER 12 N 74 2 .12
JER 31 A 47 3 .98
JER 31 B 145 9 6.42
JER 31 C 145 9 1.04
JER 31 D 47 3 7.34
JER 74 A 92 4 4.64
JER 74 B 92 4 4.65
```

SAS Code

```
JER 74 C 92 4 8.16
JER 74 D 92 4 4.63
JER 74 E 92 4 5.32
JER 74 F 92 4 7.47
JER 74 G 92 4 6.33
JER 74 H 92 4 7.95
JER 74 I 92 4 10.79
JER 74 J 92 4 10.25
JER 74 K 92 4 6.32
JER 74 L 92 4 6.22
JER 74 M 92 4 7.23
JER 74 N 92 4 7.59
JER 74 O 92 4 12.55
JER 74 P 92 4 7.66
JER 74 Q 92 4 6.74
JER 74 R 92 4 4.82
JER 74 S 92 4 15.40
JER 74 T 92 4 6.29
JER 74 U 92 4 7.88
JER 74 V 92 4 5.17
JER 74 W 92 4 4.71
JER 74 X 92 4 5.92
JER 74 Y 92 4 5.23
JPSP 56 A 124 2 10.82
JPSP 56 B 124 2 3.97
JPSP 56 C 124 2 7.01
JPSP 56 D 74 4 3.98
JPSP 56 E 38 2 0.00
JPSP 56 F 34 2 8.17
JPSP 56 G 69 4 4.55
JPSP 56 H 33 2 1.05
JPSP 56 I 36 2 4.75
JPSP 63 A 116 2 27.75
JPSP 63 B 116 2 144.98
JPSP 63 C 112 2 1.06
JPSP 63 D 112 2 5.38
JPSP 63 E 163 2 31.32
JPSP 63 F 163 2 53.18
JPSP 63 G 159 2 7.23
JPSP 63 H 159 2 3.94
JPSP 63 I 159 2 5.22
JPSP 63 J 95 2 70.42
JPSP 63 K 95 2 1.87
JPSP 63 L 95 2 12.78
JPSP 63 O 95 2 6.15
JPSP 63 P 93 2 3.19
JPSP 63 Q 93 2 5.36
JPSP 63 R 146 4 29.19
JPSP 63 S 146 4 15.25
JPSP 63 T 144 4 16.55
JPSP 63 U 145 4 105.5
JPSP 63 V 145 4 11.29
JPSP 63 W 145 4 .91
JPSP 63 X 145 4 .07
JPSP 63 Z 140 2 24.47
JPSP 63 AA 142 4 2.96
```

```
JPSP 63 AF 140 2 26.21
JPSP 63 AG 142 4 3.48
JPSP 63 AH 142 4 2.21
JPSP 53 A 63 2 7.46
JPSP 53 B 63 2 13.97
JPSP 53 C 63 2 10.33
JPSP 53 D 63 2 3.47
JPSP 53 E 63 2 3.07
JPSP 53 F 62 2 .86
JPSP 53 G 73 2 6.16
JPSP 53 H 73 2 9.4
JPSP 53 I 73 2 4.97
JPSP 53 J 73 2 5.37
JPSP 53 L 72 2 27.95
JPSP 53 M 72 2 2.15
JPSP 53 N 128 2 17.61
JPSP 53 O 128 2 5.31
JPSP 53 P 128 2 21.17
JPSP 53 Q 128 2 1.79
JPSP 53 R 128 2 5.26
JPSP 53 S 128 2 21.11
JPSP 53 T 128 2 1.80
JPSP 53 U 128 2 6.39
JPSP 53 V 128 2 11.75
JPSP 53 W 128 2 17.78
JPSP 53 Y 128 2 13.14
JPSP 53 Z 128 2 18.97
JPSP 53 AB 128 2 9.73
JPSP 53 AC 128 2 24.18
JPSP 53 AD 128 2 7.11
JPSP 53 AE 128 2 8.29
JPSP 53 AF 128 2 5.95
JPSP 53 AG 128 2 1.71
JPSP 53 AH 128 2 6.15
JPSP 53 AI 128 2 2.26
JPSP 68 A 40 3 5.44
JPSP 68 B 40 3 3.46
JPSP 68 C 40 3 4.29
JPSP 68 D 40 3 4.32
JPSP 68 E 62 3 3.21
JPSP 68 F 36 3 4.91
JPSP 68 G 36 3 5.18
JPSP 68 H 65 3 1.73
JPSP 68 J 47 3 5.43
JPSP 68 K 35 3 .39
JPSP 68 L 79 3 3.65
JPSP 68 M 79 3 3.2
JPSP 68 N 82 3 3.33
JPSP 68 O 82 3 3.16
JPSP 68 P 82 3 .65
JPSP 68 Q 47 3 2.13
JPSP 68 R 47 3 .84
JPSP 68 S 47 3 .83
JPSP 68 T 47 3 2.91
JPSP 68 U 47 3 5.91
JPSP 52 A 123 3 15.90
```

SAS Code

```
JPSP 52 B 123 3 8.64
JPSP 52 C 123 3 4.63
JPSP 52 F 123 3 11.14
JPSP 52 K 123 3 7.73
JPSP 52 O 123 3 17.62
JPSP 52 S 123 3 .79
JPSP 52 T 123 3 .18
JPSP 52 U 51 3 1.06
JPSP 52 V 51 3 .17
JPSP 52 W 51 3 1.01
JPSP 52 X 51 3 2.04
JPSP 52 Y 88 3 27.54
JPSP 52 Z 88 3 22.68
JPSP 52 AG 88 3 25.62
JPSP 52 AK 88 3 12.95
JPSP 52 AO 88 3 6.58
JPSP 52 AP 88 3 .28
JPSP 52 AT 88 3 2.70
JPSP 52 AU 51 3 1.03
JPSP 52 AV 51 3 .68
JPSP 52 AW 51 3 .07
JPSP 52 AX 51 3 1.41
JPSP 52 AY 51 3 4.98
JPSP 52 BC 46 3 10.79
JPSP 52 BD 46 3 10.50
JPSP 52 BK 46 3 3.16
JPSP 52 BO 46 3 8.64
JPSP 52 BS 46 3 2.13
JPSP 52 BT 46 3 .59
JPSP 52 BU 48 3 23.38
JPSP 52 BY 48 3 5.57
JPSP 52 CC 48 3 .35
JPSP 52 CD 48 3 6.38
JPSP 52 CH 48 3 6.18
JPSP 52 CI 48 3 .32
JPSP 52 CJ 48 3 .54
JPSP 52 CN 48 3 2.49
JPSP 69 A 76 2 2.50
JPSP 69 B 76 2 2.39
JPSP 69 C 76 2 5.23
JPSP 69 F 76 2 1.66
JPSP 69 G 76 2 .59
JPSP 69 H 76 2 .04
JPSP 69 M 87 2 1.20
JPSP 69 N 87 2 .60
JPSP 69 O 87 2 11.01
JPSP 69 U 87 2 2.2
JPSP 69 V 87 2 .01
JPSP 58 F 254 2 63.66
JPSP 58 G 255 3 2.69
JPSP 58 H 255 3 2.88
JPSP 58 I 254 2 22.92
JPSP 55 A 54 2 35.3
JPSP 55 B 54 2 69.94
JPSP 55 C 54 2 4.15
JPSP 55 D 54 2 5.84
```

```
JPSP 55 E 54 2 0.62
JPSP 55 F 54 2 8.3
JPSP 55 G 88 2 22.25
JPSP 55 H 88 2 146.73
JPSP 55 I 86 2 11.39
JPSP 55 J 86 2 4.68
JPSP 55 K 86 2 3.99
JPSP 55 L 86 2 4.86
JPSP 55 M 86 2 4.05
JPSP 55 N 86 2 .51
JPSP 55 O 86 2 11.39
JPSP 60 A 77 2 4.75
JPSP 60 B 77 2 4.29
JPSP 60 C 371 2 8.07
JPSP 60 D 371 2 69.89
JPSP 60 E 350 2 18.61
JPSP 60 F 350 2 13.04
;
title1 'Eta2 and Cohen f confidence intervals using z transformation';
/*PROC FREQ;
  tables journ * article;
 proc priNt ;
  var journ article eta2 f N k;
 proc print;
 var loweta2_99 higheta2_99 loweta2_95 higheta2_95 loweta2_90
higheta2_90 widtheta2_99 widtheta2_95 widtheta2_90;
 proc print;
  var lowf_99 highf_99 lowf_95 highf_95 lowf_90 highf_90 widthf_90
widthf_95 widthf_99;
  ruN;*/
PROC FREQ;
 TABLES JOURN * ARTICLE;
  proc print;
  var journ article analysis N k higheta2_99 loweta2_99 eta2 higheta2_95
loweta2_95 eta2 higheta2_90 loweta2_90 eta2;
  proc print;
  var journ article analysis N k highf_99 lowf_99 f highf_95 lowf_95 f
highf_90 lowf_90 f;

proc iml;

* +-----------------------------------------------------------------------
---+
   Subroutine eta2_PCTL
   Calculates percentiles from the sampling distribution of r-square
   using the inversion method of Steiger and Fouladi (1997).
   Inputs are
      SMPL_eta2 = obtained sample value of eta-square
            k = number of regressor variables
            N = sample size
         PCTL = desired percentile from the sampling distribution

   Output is
     LASTetap2 = the population r-square that provides SMPL_eta2 at the
                 pctl percentile
*
```

Appendix C

SAS Code

```
+-------------------------------------------------------------------------
-+;

  start eta2_PCTL(Smpl_eta2,k,N,pctl,lastetap2,OOPS);

  *print 'Values within eta2_PCTL Subroutine';

  eta2_tilde = Smpl_eta2/(1 - Smpl_eta2);

  * Step 1: Find value of etap-squared that is a little too high;
  OOPS = 0;
  OK = 0;
  etap2 = 0;
  loop = 0;
  flag = 0;
  flag1 = 0;
  flag2 = 0;
  do until (OK = 1);
   etap_tild = etap2/(1-etap2);
   gamma2 = 1 /(1-etap2);
   phi_1 = (N-1)*(gamma2 - 1) + k;
   phi_2 = (N-1)*(gamma2##2 - 1) + k;
   phi_3 = (N-1)*(gamma2##3 - 1) + k;
   G = (phi_2 - SQRT(phi_2##2 - (phi_1#phi_3))) / phi_1;
   v = (phi_2 - 2#etap_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
   nc = (etap_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
   obt_stat = (eta2_tilde#(N-k-1))/(v#G);

   * +-----------------------------------------+
      Be sure the computation is possible in SAS
   * +-----------------------------------------+;
      little = FINV(.0000001,v,n-k-1,nc);
      big = FINV(.99999,v,n-k-1,nc);
      not_poss = 1;
      if (obt_stat>little & obt_stat<big) then do;
         cumprob = PROBF(obt_stat,v,n-k-1,nc);
         not_poss = 0;
      end;
      *print 'Step 1A:' little big cumprob nc;

  IF (not_poss = 1 & etap2 < .99) then do;
     etap2 = etap2 + .01;
     cumprob = 1;
  end;
  IF (not_poss = 1 & etap2 > .98) then do;
     flag = 1;
     OK = 1;
     cumprob = 1;
  end;
  IF not_poss = 0 then do;
   if cumprob<pctl then OK = 1;
   if cumprob>pctl then etap2 = etap2 + 0.01;
   if etap2 > 0.99 then do;
      OK = 1;
```

```
      etap2 = .99;
      flag = 1;
   end;
    loop = loop + 1;
    if loop > 1500 then do;
*      print 'Looping too much!' loop eta2_tilde k v pctl cumprob ok nc
etap2;
      OK = 1;
    end;
  END;
   * print 'Estimating High' eta2_tilde k v pctl cumprob ok nc etap2
not_poss;
  end;
  high = etap2;
  if flag = 1 then do;
     high = 1.00;
     flag1 = 1;
  end;
    * print 'End of High Loop:' high;
    *  print high;

  * Step 2: Find value of etap-squared that is a little too low;
  OK = 0;
  etap2 = .99;
  flag = 0;
  do until (OK = 1);
   etap_tild = etap2/(1-etap2);
   gamma2 = 1 /(1-etap2);
   phi_1 = (N-1)*(gamma2 - 1) + k;
   phi_2 = (N-1)*(gamma2##2 - 1) + k;
   phi_3 = (N-1)*(gamma2##3 - 1) + k;
   G = (phi_2 - SQRT(phi_2##2 - (phi_1#phi_3))) / phi_1;
   v = (phi_2 - 2#etap_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
   nc = (etap_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
   obt_stat = (eta2_tilde#(N-k-1))/(v#G);

   * +----------------------------------------+
      Be sure the computation is possible in SAS
   * +----------------------------------------+;
      little = FINV(.0000001,v,n-k-1,nc);
      big = FINV(.99999,v,n-k-1,nc);
      not_poss = 1;
      if (obt_stat>little & obt_stat<big) then do;
         cumprob = PROBF(obt_stat,v,n-k-1,nc);
         not_poss = 0;
      end;
      * print 'Step 1B:' little big cumprob nc;
      * print 'Step1B:' little big not_poss obt_stat etap2;
  IF (not_poss = 1 & etap2 > .01) then do;
     * print 'Prog is in this one!';
     etap2 = etap2 - .01;
     cumprob = 1;
  end;
  IF (not_poss = 1 & etap2 < .02) then do;
     * print 'Program is here!';
```

```
     flag = 1;
     OK = 1;
     cumprob = 1;
  end;
  IF not_poss = 0 then do;
   if cumprob>pctl then OK = 1;
   if cumprob<pctl then etap2 = etap2 - .01;
   if etap2 < 0.01 then do;
      OK = 1;
      etap2 = .01;
      flag = 1;
   end;
  END;
  *  print 'Estimating Low' eta2_tilde k v pctl cumprob ok nc etap2
not_poss
flag;
  end;
  low = etap2;
  if flag = 1 then do;
     low = 0;
     flag2 = 1;
  end;
   *  print low;

  * Step 2: Successively halve the interval between low and high
            to obtain final value of percentile;

IF (flag1 = 0 | flag2 = 0) then do;
  change = 1;
  loop = 0;
  small = .00000000001;
  do until (change<small);
   half = (high + low)/2;
     etap_tild = half/(1-half);
     gamma2 = 1 /(1-half);
     phi_1 = (N-1)*(gamma2 - 1) + k;
     phi_2 = (N-1)*(gamma2##2 - 1) + k;
     phi_3 = (N-1)*(gamma2##3 - 1) + k;
     G = (phi_2 - SQRT(phi_2##2 - (phi_1#phi_3))) / phi_1;
     v = (phi_2 - 2#etap_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
     nc = (etap_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
     obt_stat = (eta2_tilde#(N-k-1))/(v#G);

   * +-----------------------------------------+
      Be sure the computation is possible in SAS
   * +-----------------------------------------+;
      little = FINV(.0000001,v,n-k-1,nc);
      big = FINV(.99999,v,n-k-1,nc);
      not_poss = 1;
      if (obt_stat>little & obt_stat<big) then do;
         cum_h = PROBF(obt_stat,v,n-k-1,nc);
         not_poss = 0;
      end;
      * print 'Step 2:' little big cumprob nc;
```

```
   *print 'not possible = ' not_poss;
if not_poss = 1 then do;
    change = 0;
    OOPS = 1;
    lastetap2 = 0;
end;
if not_poss = 0 then do;
   if cum_h < pctl then do;  * still too high;
      high =  half;
   end;
   if cum_h > pctl then do;  * still too low;
      low = half;
   end;
   change = abs(high - low);
   loop = loop + 1;
   if loop > 1500 then small = .000000001;
   if loop > 3000 then small = .01;
    * print high low change;
  end;
  lastetap2 = (high + low)/2;
    * print lastetap2;
end;
end;

IF (flag1 = 1 & flag2 = 1) then do;
  lastetap2 = 0;
  OOPS = 1;
end;
finish;

use  one;

 read all var{Smpl_eta2} into Smpl_eta2;
 read all var{u} into U;
 read all var{v} into V;
 read all var{N} into N;
 read all var{k} into K;
 k_total = nrow(Smpl_eta2);
  file print;
  put @1 'Confidence Intervals Around Sample eta2' //
      @16 '99% CI' @36 '95% CI' @56 '90% CI' /
      @2 'eta2' @10 '------------------' @30 '------------------' @50
'------------------' /
      @3 '' @12 'Lower     Upper' @32 'Lower     Upper' @52 'Lower
Upper' /
       @1 '--------' @10 '--------- ---------' @30 '--------- ---------
' @50 '--------- ---------';

do i = 1 to k_total;

  run eta2_PCTL(Smpl_eta2[i,1],k[i,1],N[i,1],0.005,eta2_005,oops005);
  run eta2_PCTL(Smpl_eta2[i,1],k[i,1],N[i,1],0.995,eta2_995,oops995);
  run eta2_PCTL(Smpl_eta2[i,1],k[i,1],N[i,1],0.025,eta2_025,oops025);
  run eta2_PCTL(Smpl_eta2[i,1],k[i,1],N[i,1],0.975,eta2_975,oops975);
  run eta2_PCTL(Smpl_eta2[i,1],k[i,1],N[i,1],0.05,eta2_05,oops05);
```

```
  run eta2_PCTL(Smpl_eta2[i,1],k[i,1],N[i,1],0.95,eta2_95,oops95);


  print_eta2 = Smpl_eta2[i,1];
  file print;
  put @1 print_eta2 8.4 @10 eta2_995 8.4 @20 eta2_005 8.4 @30 eta2_975
8.4 @40 eta2_025 8.4 @50 eta2_95 8.4 @60 eta2_05;

end;
*proc iml;

start find_NC(F_obt, u, v, ncc, pctl, f);
  OK=0;
  nc=0;
  target = pctl;
  loop = 0;
  do until (OK = 1);
     cumprob = PROBF(F_obt, u, v, nc);
     if cumprob<target then OK = 1;
     if cumprob>target then nc = nc + 3.0;
     loop = loop + 1;
  end;

  low = nc;
  high = 0;

      change = 1;
      loop = 0;
      small = .00000000001;
      do until (change<small);
       half = (high + low)/2;
       cum_h = PROBF(F_obt, u, v, half);
       if cum_h < pctl then do;
               low = half;
         end;
         if cum_h > pctl then do;
              high = half;
         end;
       change = abs(high - low);
       loop = loop + 1;
       if loop > 1500 then small = .000000001;
       if loop > 3000 then small = .01;
          * print high low change;
       ncc = (high + low)/2;
         f = ((ncc/(u + v + 1)))**.5;
 * print ncc;
end;

finish;

use  one;

 read all var{F_obt} into F_obt;
 read all var{u} into U;
 read all var{v} into V;
```

SAS Code

```
 read all var{f} into effect_vec;
 k_total = nrow(effect_vec);
  file print;
  put // @1 'Confidence Intervals Around Sample Cohen f' //
      @16 '99% CI' @36 '95% CI' @56 '90% CI' /
      @2 'Effect' @10 '--------------------' @30 '--------------------'
@50 '--------------------' /
      @3 'Size' @12 'Lower     Upper' @32 'Lower     Upper' @52 'Lower
Upper' /
       @1 '--------' @10 '--------- ---------' @30 '--------- ---------
' @50 '--------- ---------';

do i = 1 to k_total;


  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_005, .005, f005);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_995, .995, f995);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_025, .025, f025);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_975, .975, f975);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_05, .05, f05);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_95, .95, f95);

  print_effect = effect_vec[i,1];
  file print;
  put @1 print_effect 8.4 @10 f995 8.4 @20 f005 8.4 @30 f975 8.4 @40
f025 8.4 @50 f95 8.4 @60 f05 8.4;
end;
quit;
```

Appendix C

SAS Code

```
 data one;

    iNput journ $ article analysis $ N k r2;
* ++++++++++++++++++++++++++++++++++++++++++++
   This calculates coNfideNce iNtervals for the effect
      size for regressioN aNalyses (f2) usiNg
      a log traNsformatioN (O&F 3--Fisher Z)  Two
      approaches were used (testiNg, as we discussed)
      Very differeNt results.
   Last edit:  Aug 23
  ++++++++++++++++++++++++++++++++++++++++++++;

fsquare = 0;
lowr2_90 = 0;
lowr2_95 = 0;
lowr2_99 = 0;
highr2_90 = 0;
highr2_95 = 0;
highr2_99 = 0;
widthr2_90 = 0;
widthr2_95 = 0;
widthr2_99 = 0;
lowf2_90 = 0;
lowf2_95 = 0;
lowf2_99 = 0;
highf2_90 = 0;
highf2_95 = 0;
highf2_99 = 0;
widthf2_90 = 0;
widthf2_95 = 0;
widthf2_99 = 0;


u = k ;
v = N - k - 1;
F_obt = (r2/u)/((1- r2)/v);
F2 = r2/(1-r2); * I computed f2 here;
Smpl_R2 = r2;



*++++++++++++++++++++++++++++++++++++++++++++
CalculatiNg the upper aNd lower bouNds of eta2
usiNg O&F 3..called loweta2_95 aNd higheta2_95.
CurreNtly calculatioNs are oNly doNe usiNg the
95th perceNtile, peNdiNg resolutioN of method
++++++++++++++++++++++++++++++++++++++++++++;

z=log((1+sqrt(r2))/(1-sqrt(r2)));

lowr2_95 = z - ((2*(1.96))/(sqrt(N)));
highr2_95 =  z + ((2*(1.96))/(sqrt(N)));

    low95 = exp(lowr2_95);
```

```
    high95 = exp(highr2_95);
   lowr2_95 = ((low95-1)/(low95+1))**2;
    highr2_95 = ((high95-1)/(high95+1))**2;

if2 lowr2_95<0 theN lowr2_95=0;
if2 highr2_95>1 theN highr2_95=1;
widthr2_95 = highr2_95-lowr2_95;

lowr2_99 = z - ((2*(2.576))/(sqrt(N)));
highr2_99 =  z + ((2*(2.576))/(sqrt(N)));

    low99 = exp(lowr2_99);
    high99 = exp(highr2_99);
   lowr2_99 = ((low99-1)/(low99+1))**2;
    highr2_99 = ((high99-1)/(high99+1))**2;

if2 lowr2_99<0 theN lowr2_99=0;
if2 highr2_99>1 theN highr2_99=1;
widthr2_99 = highr2_99-lowr2_99;


lowr2_90 = z - ((2*(1.645))/(sqrt(N)));
highr2_90 =  z + ((2*(1.645))/(sqrt(N)));

    low90 = exp(lowr2_90);
    high90 = exp(highr2_90);
   lowr2_90 = ((low90-1)/(low90+1))**2;
    highr2_90 = ((high90-1)/(high90+1))**2;

if2 lowr2_90<0 theN lowr2_90=0;
if2 highr2_90>1 theN highr2_90=1;
widthr2_90 = highr2_90-lowr2_90;



* ++++++++++++++++++++++++++++++++++++++++
This set of2 CIs (called lowf2_95 aNd
highf2_95) are coNstructed by calculatiNg
f2 f2or the lower r2 aNd upper r2 calculated
earlier ...this method is the oNe more
appropriate???
++++++++++++++++++++++++++++++++++++++++;
lowf2_95 = lowr2_95/(1-lowr2_95);
highf2_95 = highr2_95/(1-highr2_95);
widthf2_95 = highf2_95-lowf2_95;

lowf2_99 = lowr2_99/(1-lowr2_99);
highf2_99 = highr2_99/(1-highr2_99);
widthf2_99 = highf2_99-lowf2_99;

lowf2_90 = lowr2_90/(1-lowr2_90);
highf2_90 = highr2_90/(1-highr2_90);
widthf2_90 = highf2_90-lowf2_90;

cards;
```

```
JER 7 A 48 4 .80
JER 7 B 48 3 .75
JER 7 C 48 4 .74
JER 7 R 39 1 .61
JER 7 S 39 1 .60
JER 7 T 39 3 .71
RRQ 51 A 89 1 .13
RRQ 51 B 89 1 .28
RRQ 51 C 89 1 .17
RRQ 51 D 89 1 .35
RRQ 51 E 89 1 .34
RRQ 51 F 89 1 .30
RRQ 51 G 89 1 .31
RRQ 51 H 89 1 .39
RRQ 51 I 89 1 .88
RRQ 51 J 89 1 .47
RRQ 51 K 89 1 .52
RRQ 51 L 89 1 .08
RRQ 51 M 89 1 .08
RRQ 51 N 89 1 .45
RRQ 51 O 89 1 .49
RRQ 51 P 89 1 .51
RRQ 51 Q 89 1 .44
RRQ 51 R 89 1 .88
RRQ 51 S 47 1 .27
RRQ 51 T 47 1 .41
RRQ 51 U 47 1 .25
RRQ 51 V 47 1 .10
RRQ 51 W 89 1 .53
RRQ 51 X 89 1 .55
RRQ 51 Y 89 1 .17
RRQ 51 Z 89 1 .13
RRQ 51 AA 47 1 .06
RRQ 51 AB 47 1 .06
RRQ 51 AC 47 1 .29
RRQ 51 AD 47 1 .16
RRQ 51 AE 89 1 .37
RRQ 51 AF 89 1 .38
RRQ 51 AG 89 1 .18
RRQ 51 AH 89 1 .17
RRQ 51 AI 89 1 .26
RRQ 51 AJ 89 1 .13
RRQ 51 AK 89 1 .33
RRQ 51 AL 89 1 .19
RRQ 78 AG 195 1 .29
RRQ 78 AH 195 1 .10
RRQ 78 AI 195 1 .08
RRQ 78 AJ 195 3 .42
RRQ 78 AK 195 4 .49
RRQ 78 AL 195 5 .54
RRQ 78 AM 149 1 .35
RRQ 78 AN 149 1 .17
RRQ 78 AO 149 1 .04
RRQ 78 AP 149 1 .22
RRQ 78 AQ 149 3 .45
RRQ 78 AR 149 4 .48
```

SAS Code

```
RRQ 32 A 60 7 .55
JER 2 A 2307 5 .26
JER 2 B 2307 5 .18
JER 2 C 644 5 .23
JER 2 D 644 5 .23
JER 2 E 2307 4 .70
JER 2 F 2307 4 .56
JER 2 G 644 5 .74
JER 2 H 644 5 .62
JER 2 I 2307 9 .62
JER 2 J 2307 9 .69
JER 2 K 2307 9 .70
JER 2 L 2307 9 .71
JER 2 M 2307 9 .71
JER 2 N 2307 9 .71
JER 2 O 2307 9 .71
JER 2 P 2307 9 .71
JER 2 Q 2307 9 .71
JER 2 R 2307 5 .52
JER 2 S 2307 5 .54
JER 2 T 2307 5 .56
JER 2 U 2307 5 .57
JER 2 V 2307 5 .58
JER 4 x 3856 5 .26
JER 4 x 3856 5 .18
JER 4 x 3856 5 .23
JER 4 x 3856 5 .23
JER 4 x 3856 5 .70
JER 4 x 3856 5 .56
JER 4 x 3856 5 .74
JER 4 x 3856 5 .62
JPSP 57 A 638 3 .75
JPSP 57 B 621 3 .67
JPSP 57 C 649 3 .77
JPSP 57 D 642 3 .73
JPSP 57 E 599 3 .70
JPSP 57 F 630 3 .81
JPSP 57 G 630 3 .78
JPSP 57 H 640 3 .74
JPSP 57 I 624 3 .76
JPSP 57 J 650 3 .71
JPSP 57 K 643 3 .64
JPSP 57 L 655 3 .69
JPSP 57 M 633 3 .69
JPSP 57 N 631 3 .74
JPSP 57 O 640 2 .24
JPSP 57 P 624 2 .22
JPSP 57 Q 650 2 .42
JPSP 57 R 650 2 .42
JPSP 57 S 658 2 .26
JPSP 57 T 633 2 .31
JPSP 57 U 632 2 .35
;
```

Appendix C

SAS Code

```
*  the following card set is absent the large N with large R2 (middle 4)
and will run complete, even with large R2 when there is small N;

*cards;
*JER 4 3856 5 .26
JER 4 3856 5 .18
JER 4 3856 5 .23
JER 4 3856 5 .23
JER 7 48 4 .80
JER 7 48 3 .75
JER 7 48 4 .74
JER 7 39 1 .61
JER 7 39 1 .60
JER 7 39 3 .71

;

proc freq;
tables journ article;
title1 'R2 and F2 Confidence Intervals using Z transformation';
 /*proc priNt ;
  var r2 f2 N k;
 proc print;
 var lowr2_99 highr2_99 lowr2_95 highr2_95 lowr2_90 highr2_90 widthr2_99
widthr2_95 widthr2_90;
 proc print;
  var lowf2_99 highf2_99 lowf2_95 highf2_95 lowf2_90 highf2_90
widthf2_90 widthf2_95 widthf2_99;*/
proc print;
 var journ article analysis N k highr2_99 lowr2_99 r2 highr2_95 lowr2_95
r2 highr2_90 lowr2_90 r2;
proc print;
 var journ article analysis N k highf2_99 lowf2_99 f2 highf2_95 lowf2_95
f2 highf2_90 lowf2_90 f2;

run;
proc iml;

* +-------------------------------------------------------------------
---+
   Subroutine R2_PCTL
   Calculates percentiles from the sampling distribution of r-square
   using the inversion method of Steiger and Fouladi (1997).
   Inputs are
      SMPL_R2 = obtained sample value of r-square
            k = number of regressor variables
            N = sample size
         PCTL = desired percentile from the sampling distribution

   Output is
     LASTRHO2 = the population r-square that provides SMPL_R2 at the
                pctl percentile
*
+-------------------------------------------------------------------
-+;
start R2_PCTL(Smpl_R2,k,N,pctl,lastrho2,OOPS);
```

Appendix C

SAS Code

```
*print 'Values within R2_PCTL Subroutine';

R2_tilde = Smpl_R2/(1 - Smpl_R2);

* Step 1: Find value of rho-squared that is a little too high;
OOPS = 0;
OK = 0;
rho2 = 0;
loop = 0;
flag = 0;
flag1 = 0;
flag2 = 0;
do until (OK = 1);
 rho_tild = rho2/(1-rho2);
 gamma2 = 1 /(1-rho2);
 phi_1 = (N-1)*(gamma2 - 1) + k;
 phi_2 = (N-1)*(gamma2##2 - 1) + k;
 phi_3 = (N-1)*(gamma2##3 - 1) + k;
 G = (phi_2 - SQRT(phi_2##2 - (phi_1#phi_3))) / phi_1;
 v = (phi_2 - 2#rho_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
 nc = (rho_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
 obt_stat = (R2_tilde#(N-k-1))/(v#G);

 * +-----------------------------------------+
    Be sure the computation is possible in SAS
 * +-----------------------------------------+;
    little = FINV(.0000001,v,n-k-1,nc);
    big = FINV(.99999,v,n-k-1,nc);
    not_poss = 1;
    if (obt_stat>little & obt_stat<big) then do;
       cumprob = PROBF(obt_stat,v,n-k-1,nc);
       not_poss = 0;
    end;
    *print 'Step 1A:' little big cumprob nc;

IF (not_poss = 1 & rho2 < .99) then do;
   rho2 = rho2 + .01;
   cumprob = 1;
end;
IF (not_poss = 1 & rho2 > .98) then do;
   flag = 1;
   OK = 1;
   cumprob = 1;
end;
IF not_poss = 0 then do;
 if cumprob<pctl then OK = 1;
 if cumprob>pctl then rho2 = rho2 + 0.01;
 if rho2 > 0.99 then do;
    OK = 1;
    rho2 = .99;
    flag = 1;
 end;
  loop = loop + 1;
```

```
    if loop > 1500 then do;
*      print 'Looping too much!' loop R2_tilde k v pctl cumprob ok nc
rho2;
      OK = 1;
    end;
  END;
   * print 'Estimating High' R2_tilde k v pctl cumprob ok nc rho2
not_poss;
  end;
  high = rho2;
  if flag = 1 then do;
     high = 1.00;
     flag1 = 1;
  end;
    * print 'End of High Loop:' high;
    *  print high;

  * Step 2: Find value of rho-squared that is a little too low;
  OK = 0;
  rho2 = .99;
  flag = 0;
  do until (OK = 1);
   rho_tild = rho2/(1-rho2);
   gamma2 = 1 /(1-rho2);
   phi_1 = (N-1)*(gamma2 - 1) + k;
   phi_2 = (N-1)*(gamma2##2 - 1) + k;
   phi_3 = (N-1)*(gamma2##3 - 1) + k;
   G = (phi_2 - SQRT(phi_2##2 - (phi_1#phi_3))) / phi_1;
   v = (phi_2 - 2#rho_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
   nc = (rho_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
   obt_stat = (R2_tilde#(N-k-1))/(v#G);

   * +-----------------------------------------+
      Be sure the computation is possible in SAS
   * +-----------------------------------------+;
      little = FINV(.0000001,v,n-k-1,nc);
      big = FINV(.99999,v,n-k-1,nc);
      not_poss = 1;
      if (obt_stat>little & obt_stat<big) then do;
         cumprob = PROBF(obt_stat,v,n-k-1,nc);
         not_poss = 0;
      end;
      * print 'Step 1B:' little big cumprob nc;
      * print 'Step1B:' little big not_poss obt_stat rho2;
  IF (not_poss = 1 & rho2 > .01) then do;
     * print 'Prog is in this one!';
     rho2 = rho2 - .01;
     cumprob = 1;
  end;
  IF (not_poss = 1 & rho2 < .02) then do;
     * print 'Program is here!';
     flag = 1;
     OK = 1;
     cumprob = 1;
  end;
```

Appendix C

SAS Code

```
  IF not_poss = 0 then do;
   if cumprob>pctl then OK = 1;
   if cumprob<pctl then rho2 = rho2 - .01;
   if rho2 < 0.01 then do;
      OK = 1;
      rho2 = .01;
      flag = 1;
   end;
  END;
  *  print 'Estimating Low' R2_tilde k v pctl cumprob ok nc rho2
not_poss
flag;
  end;
  low = rho2;
  if flag = 1 then do;
     low = 0;
     flag2 = 1;
  end;
   *  print low;

  * Step 2: Successively halve the interval between low and high
            to obtain final value of percentile;

IF (flag1 = 0 | flag2 = 0) then do;
  change = 1;
  loop = 0;
  small = .00000000001;
  do until (change<small);
   half = (high + low)/2;
     rho_tild = half/(1-half);
     gamma2 = 1 /(1-half);
     phi_1 = (N-1)*(gamma2 - 1) + k;
     phi_2 = (N-1)*(gamma2##2 - 1) + k;
     phi_3 = (N-1)*(gamma2##3 - 1) + k;
     G = (phi_2 - SQRT(phi_2##2 - (phi_1#phi_3))) / phi_1;
     v = (phi_2 - 2#rho_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
     nc = (rho_tild#sqrt(gamma2)#sqrt((N-1)#(N-k-1)))/G##2;
     obt_stat = (R2_tilde#(N-k-1))/(v#G);

   * +----------------------------------------+
      Be sure the computation is possible in SAS
   * +----------------------------------------+;
      little = FINV(.0000001,v,n-k-1,nc);
      big = FINV(.99999,v,n-k-1,nc);
      not_poss = 1;
      if (obt_stat>little & obt_stat<big) then do;
         cum_h = PROBF(obt_stat,v,n-k-1,nc);
         not_poss = 0;
      end;
      * print 'Step 2:' little big cumprob nc;
   *print 'not possible = ' not_poss;
if not_poss = 1 then do;
    change = 0;
    OOPS = 1;
```

```
    lastrho2 = 0;
end;
if not_poss = 0 then do;
   if cum_h < pctl then do;  * still too high;
      high =  half;
   end;
   if cum_h > pctl then do;  * still too low;
      low = half;
   end;
   change = abs(high - low);
   loop = loop + 1;
   if loop > 1500 then small = .000000001;
   if loop > 3000 then small = .01;
    * print high low change;
  end;
  lastrho2 = (high + low)/2;
    * print lastrho2;
end;
end;

IF (flag1 = 1 & flag2 = 1) then do;
  lastrho2 = 0;
  OOPS = 1;
end;
finish;

use  one;

 read all var{Smpl_R2} into Smpl_R2;
 read all var{u} into U;
 read all var{v} into V;
 read all var{N} into N;
 read all var{k} into K;
 k_total = nrow(Smpl_r2);
  file print;
  put @1 'Confidence Intervals Around Sample R2' //
      @16 '99% CI' @36 '95% CI' @56 '90% CI' /
      @2 'R2' @10 '-------------------' @30 '-------------------' @50 '-
------------------' /
      @3 '' @12 'Lower     Upper' @32 'Lower     Upper' @52 'Lower
Upper' /
       @1 '--------' @10 '--------- ---------' @30 '--------- ---------
' @50 '--------- ---------';

do i = 1 to k_total;

  run R2_PCTL(Smpl_R2[i,1],k[i,1],N[i,1],0.005,r2_005,oops005);
  run R2_PCTL(Smpl_R2[i,1],k[i,1],N[i,1],0.995,r2_995,oops995);
  run R2_PCTL(Smpl_R2[i,1],k[i,1],N[i,1],0.025,r2_025,oops025);
  run R2_PCTL(Smpl_R2[i,1],k[i,1],N[i,1],0.975,r2_975,oops975);
  run R2_PCTL(Smpl_R2[i,1],k[i,1],N[i,1],0.05,r2_05,oops05);
  run R2_PCTL(Smpl_R2[i,1],k[i,1],N[i,1],0.95,r2_95,oops95);


  print_r2 = Smpl_r2[i,1];
```

```
  file print;
  put @1 print_r2 8.4 @10 r2_995 8.4 @20 r2_005 8.4 @30 r2_975 8.4 @40
r2_025 8.4 @50 r2_95 8.4 @60 r2_05;

end;



proc iml;

start find_NC(F_obt, u, v, ncc, pctl, f2); * I added f2 to the arguments
here;
  OK=0;
  nc=0;
  target = pctl;
  loop = 0; * I initialized loop here;
  do until (OK = 1);
     cumprob = PROBF(F_obt, u, v, nc);
     if cumprob<target then OK = 1;
     if cumprob>target then nc = nc + 3.0;
     loop = loop + 1;
  end;

  low = nc;
  high = 0;

      change = 1;
      loop = 0;
      small = .00000000001;
      do until (change<small);
       half = (high + low)/2;
       cum_h = PROBF(F_obt, u, v, half);
       if cum_h < pctl then do;
                low = half;
         end;
         if cum_h > pctl then do;
               high = half;
         end;
       change = abs(high - low);
       loop = loop + 1;
       if loop > 1500 then small = .000000001;
       if loop > 3000 then small = .01;
          * print high low change;
       ncc = (high + low)/2;
         f2 = (ncc/(u + v + 1));
 *print ncc;
end;

finish;

use  one;

 read all var{F_obt} into F_obt; * I changed this vector to F_obt;
 read all var{u} into U; * I changed this vector to U;
 read all var{v} into V; * I changed this vector to V;
```

Appendix C

SAS Code

```
 read all var{F2} into effect_vec; * I added this statement to create
effect_vec;
 k_total = nrow(effect_vec);
  file print;
  put @1 'Confidence Intervals Around Sample f2' //
      @16 '99% CI' @36 '95% CI' @56 '90% CI' /
      @2 'Effect' @10 '--------------------' @30 '--------------------'
@50 '--------------------' /
      @3 'Size' @12 'Lower     Upper' @32 'Lower     Upper' @52 'Lower
Upper' /
       @1 '--------' @10 '--------- ---------' @30 '--------- ---------
' @50 '--------- ---------';

do i = 1 to k_total;

  *obs_stat = effect_vec[i,1] # sqrt(n1[i,1]#n2[i,1]/(n1[i,1]+n2[i,1]));
  *obt_F = (r2[i,1]/u[i,1])/((1- r2[i,1])/v[i,1]);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_005, .005, f2005);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_995, .995, f2995);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_025, .025, f2025);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_975, .975, f2975);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_05, .05, f205);
  run find_NC(F_obt[i,1], u[i,1], v[i,1], nc_95, .95, f295);

  print_effect = effect_vec[i,1]; *Don't think this belongs as is
(relative to
   previous computation, but didn't want to lose the thought;
  file print;
  put @1 print_effect 8.4 @10 f2995 8.4 @20 f2005 8.4 @30 f2975 8.4 @40
f2025 8.4 @50 f295 8.4 @60 f205 8.4;
end;
quit;
```

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

The statistical significance notation used reflects how the original author reported it (format, amount of information included, etc. Also, the wording in the Findings/Results column is/are exact quotes. Any information added or deleted for the purposes of clarification are in parenthesis and italicized.

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
2/1	Do variables that can be controlled by school systems (e.g., average class size, teacher experience, pupil-teacher ratio, teach salary, and expenditure per pupil) predict academic achievement	R^2: .26, p<.001 (reading) R^2: .18, p<.001 (math)	According to the model F statistics, both multiple regressions *(reading and math)* were statistically significant in accounting for variance in third-grade reading and mathematics scores. However, the model R^2 for the two models was relatively small, with R^2 values of .26 and .18 respectively.	Cohen f^2: .3514 (reading) Cohen f^2: .2195 (math)	(2) Slight Change Needed	.2979< f^2<.3514 (reading) .1796< f^2<.2642	(3) Much Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
2/2	Do variables that cannot be controlled by school systems (e.g., percentage White, low income, attendance, mobility) predict academic achievement	R^2: .70, p<.001 (reading) R^2: .56, p<.001 (math)	In contrast to the low model R^2 values obtained for the can control regression models, the R^2 values obtained for the cannot control regression models were considerably higher. We therefore concluded that the cannot control models accounted more accurately for variance in Grade 3 achievement scores than did the can control models.	Cohen f^2: 2.3333 (reading) Cohen f^2: .1.2727 (math)	(1) No Change Needed	2.1149 <f^2< 2.5706 (reading) 1.1397 <f^2<1.4176	(1) No Change Needed
4/1	Does the use of video as an accommodation on a math test to avoid impact of reading ability on performance on a math test for all students?	t value not reported. p = .08	Students taking the video version of the test scored slightly higher than those taking the standard version, although that difference was not statistically significant. As our results indicate, accommodations are unnecessary for the majority of students.	Cohen d: .0753	(1) No Change Needed	-.1012 <d<.25170	(2) Slight Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
4/2	Does the use of video as an accommodation on a math test to avoid impact of reading ability on performance on a math test for students with low math ability?	*t* value not reported, p=.05	Of the subgroups examined, only the low mathematics group showed a preference that reached significance.	Cohen *d*: .1537	(2) Slight Change Needed	-.1265 <*d*<.4344	(3) Much Change Needed
7/1	Do student's with different time preferences (morning or afternoon) perform differently if they take a test in the morning?	$F(1,64) = 5.44$, $p<.05$	There was a significant difference between afternoon-preferenced students and morning-preferenced studs taking the test in the morning. The results indicate clearly that the time-of-day element in learning style may play a significant part in the instructional environment. When time preference and testing environment were matched, significant differences emerged between test results—but only for the morning test.	Cohen *f*: .2849	(3) Much Change Needed	.0024 <*f*<.5283	(4) Complete revision needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
7/2	Do student's with different time preferences (morning or afternoon) perform differently if they take a test in the afternoon?	$F(1,64) = 3.81$, $p<.055$	here was a small difference between afternoon-preferenced students and morning-preferenced studs taking the test in the afternoon. The results indicate clearly that the time-of-day element in learning style may play a significant part in the instructional environment. When time preference and testing environment were matched, significant differences emerged between test results—but only for the morning test.	Cohen *f*: .2385	(2) Slight Change Needed	.0000 <*f*<.4805	(4) Complete revision needed
11/1	Do children differ in their explicit and implicit comprehension abilities?	$F(1,155) = 15.19$, $p<.001$	The explicit comprehension subscore was significantly higher than the implicit comprehension subscore.	Cohen *f*: .3101	(1) No Change Needed	.1499 <*f*<.4778	(3) Much Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
11/2	Do children differ in their overall ability to comprehend narrative based on their grade level (K-2)	$F(2,88) = 7.18, p<.001$ $F(2,87) = 9.17, p<.001$ $F(2,88) = 9.47, p<.001$	Older children received significantly more points than younger children on total prompted comprehension for all three task versions. Older students were significantly more likely to provide retellings with appropriate sequencing of events.	Cohen *f*: .3994 Cohen *f*: .4514 Cohen *f*: .44562	(2) Slight Change Needed	.1839 <*f*<.6321 .2328 <*f*<.6894 .2385 <*f*<.6933	(2) Slight Change Needed
11/4	Does the ability of children to retell a story differ among students in grades K-2nd?	$F(1,84)=5.9, p<.05$	Retelling (and prompted comprehension scores) improved significantly, indicating that the NC task differentiates between children who can recall main narrative elements from children who have weakness with this narrative comprehension skill.	Cohen *f*: .3236	(2) Slight Change Needed	.1058 <*f*<.5561	(3) Much Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
12/1	Does negotiation of mean (allowing students to discuss meanings of words prior to taking individual assessments) impact performance?	F=124.81, df = *3*, p<.001	An analysis of variance with repeated measures showed a statistically significant main effect for condition.	Cohen *f*: 1.5747	(1) No Change Needed	1.2960 <*f*<1.8936	(1) No Change Needed
12/2	Does the level of language ability impact effectiveness of using negotiation of meaning for students measured on comprehension?	F = 1.73, df = 9, p = .079	The interaction of condition by level of language proficiency was not significant.	Cohen *f:* 0.3350	(3) Much Change Needed	.1896 <*f*<.5395	(4) Total Revision Needed
15/1	Is there a difference in the way students learning foreign language use different types of clues, specifically contextual clues or, for learning Japanese, kanji measures or, integrating the two methods.	F(2,116)= 31.6, p<.0001	A one-way analysis of variance indicates a statistically significant effect of condition on students' choice of integrated answers.	Cohen *f*: .7218	(2) Slight Change Needed	.5190 <*f*<.9686	(2) Slight Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
17/1	Is a child's receptive vocabulary at ages 4 and 7 different, depending on ethnic background?	*R2* = .47, *p*<.01 (Age 4) *R2* = .52, *p*<.01 (Age 7) Two main effects:	Children's receptive vocabulary at ages 4 and 7 also differs strongly between groups.	Cohen f^2: .89 Cohen f^2: 1.08	(1) No Change Needed	.4552< f^2<.8868 .5796< f^2<1.8644	(2) Slight Change Needed
18/1	Do children who receive different types of praise (ability, effort, or none) differ in what they attribute their performance (effort or intelligence) to on performance measures?	Effect of 'low effort' on performance': *F*(2,120) = 8.64, *p*<.001 Effect of 'low intelligence' on performance': *F*(2,120) = 4.63, *p*<.05	Children differed in their endorsements of low effort and low ability as causes of their failure. Overall, the findings (of the study) support our hypothesis that children who are praised for intelligence when they succeed are the ones least likely to attribute their performance to low effort, a factor over which they have some amount of control	Cohen *f*: .3748 Cohen *f*: .2744	(3) Much Change Needed	.1750 <*f*<.5482 -.0621 <*f*<.4423	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
18/2	Do children who receive different types of praise (ability, effort, or none) differ in how they rate their enjoyment of tasks?	$F(2,129) = 7.73, p<.005$ Ability vs. Effort $t(81) = -3.81$ $p<.001$ Ability vs. Control $t(83) = -2.03$, $p<.05$ Control vs. Effort $t(82)=2.16$, $p<.05$	Children praised for intelligence enjoyed the tasks less than did children praised for effort; again, children in the control condition fell in between the other two groups. Children praised for intelligence were significantly less likely to enjoy the problems than were children in the effort and control conditions. Further, children in the control condition were less likely to enjoy the problems than those praised for effort. Indictment of ability also led children praised for intelligence to display more negative responses in terms of lower levels of task enjoyment than their counterparts, who received commendations for effort.	Cohen *f*: .3545 Cohen *d*: -.8816 Cohen *d*: -.4495 Cohen *d*: -.4801	(3) Much Change Needed	.1358 <*f*<.5269 -1.3495 <*d*<-.4136 -.8814 <*d*<-.0175 -.9158 <*d*<-.0043	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
18/3	Do children who receive different types of praise (ability, effort, or none) differ in regarding their future expectations of their performance?	$F(2,48) = 1.01$, *ns*	No significant differences were noted for children's expectations; children in the intelligence, effort and control conditions displayed equivalent expectations.	Cohen *f*: .1990	(3) Much Change Needed	.0000 <*f*<.4419	(4) Complete Revision Needed
18/4	Do children who receive different types of praise (ability, effort, or none) differ in how harshly they judge their performance??	$F(2,48) = 2.04$, *ns*	No significant differences were noted for children's judgement of their performance; children in the intelligence, effort and control conditions displayed equivalent expectations. These results indicate that effort, praise and intelligence do not lead children to judge their performance differently.	Cohen *f:* .2828	(4) Complete Revision Needed	.0000 <*f*<.5366	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
18/5	Do children who receive different types of praise (ability, effort, or none) differ regarding persistence?	$F(2,45) = 3.16$, $p=.05$ Ability vs. Effort $t(30) = -2.09$ $p<.05$ Ability vs. Control $t(30) = -2.22$, $p<.05$ Control vs. Effort $t(30)=-.12$, *ns*	Children praised for intelligence were less likely to want to persist on the problems after setbacks than were children praised for effort; children in the control condition closely resembled those in the effort conditions. Follow-up t-tests revealed significant differences between the intelligence condition and the effort and control conditions but no difference between the effort and control conditions.. Indictment of ability also led children praised for intelligence to display more negative responses in terms of lower levels of task persistence than their counterparts, who received commendations for effort.	Cohen *f*: .3707 Cohen *d*: -.7332 Cohen *d*: -.7777 Cohen *d*: -.0412	(2) Slight Change Needed	.0000 <*f*<.6462 -1.4609 <*d*<-.0055 -1.0582 <*d*<-.0472 -.6746 <*d*<-.7570	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
19/1	Is there a difference in confidence between individuals classified as high or low in dogmatism?	$F(1,61)$, $p<.01$	Individuals high in dogmatism were much more confident in their judgments than individuals low in dogmatism.	Cohen *f*: .2905	(2) Slight Change Needed	.0500 <*f*<.5236	(3) Much Change Needed
19/2	Are there differences in the types of reasons provided for outcomes that support an individuals opinion (pro decisions) as compared to the reasons that oppose an individual's opinion (con decisions resulting from how dogmatic an individual is?	$F(1,61)$, $p<.01$	Individuals high in dogmatism produced more pro reasons than individuals low in dogmatism. Also they produce fewer con reasons than individuals low in dogmatism. The results show that individuals high in dogmatism are more likely to generate cognitions supporting their newly created beliefs and are less likely to generate cognitions contradicting them.	Cohen *f*: .4049	(1) No change needed	.1462 <*f*<.6605	(2) Slight Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
20/1	Does gender stereotyping impact prediction of self performance by women, regardless of their ability as evidenced by previous performance?	$F(1,52) = 4.15$, $p < .05$	On ratings of estimated performance on a stereotypical task, the effects of initial confidence were completely undermined.	Cohen *f*: .3920	(3) Much Change Needed	.1162 <*f*<.6060	(4) Complete Revision Needed
21/1	Would pre-assessment belief about whether a test outcome predicts weakness or excellence impact actual performance by women on a math test?	$F(1,122) = 3.97$, $p<.05$	Women who believed that the test would indicate whether they were especially weak in math performed less well than did women who believed the test would indicate whether they were exceptionally strong.	Cohen *f*: .1789	(2) Slight Change Needed	.00198 <*f*<.3144	(4) Complete Revision Needed
21/2	Would pre-assessment belief about whether a test outcome predicts weakness or excellence impact actual performance by women on a math test?	$F(1,122) = 7.01$, $p<.01$	Men performed less well when they believed the test might indicate whether they were exceptionally strong.	Cohen *f:* .2378	(1) No Change Needed	.05960 <*f*<.4233	(2) Slight Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
24/1	Does belief that selection for leadership role is based on merit or gender-bias impact performance?	$F(1,75) = 4.75$, $p<.04$	As predicted, participants in the gender-only condition performed worse than participants in the control and gender + merit conditions. The data (from this study) were conceptually consistent with prior research in demonstrating that the belief that one has been selected for a task on the basis of gender alone.	Cohen *f:* .2484	(2) Slight Change Needed	.0225 <*f*<.4867	(3) Much Change is Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
24/2	Does membership in a 'stigamatized' race/ethnicity (African American or Latino) as compared to those in a 'non-stigmatized' race/ethnicity, impact the degree to which one suspects preferential treatment for admission into college.	$F(1,369) = 69.89$, $p<.001$	When we compared stigmatized and nonstigmatized students in the degree to which they suspected that their race or ethnicity might have helped them gain admission to college, we also found a significant difference, as expected. Stigmatized students suspected that their admission to the University of Texas at Austin had been influenced by their race or ethnicity to a greater extent than did nonstigamtized students.	$F = .4043$	(1) No Change Needed	$.3252 < f < .5470$	(2) Slight Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
24/3	Does membership in a 'stigmatized' race/ethnicity (African American or Latino) as compared to those in a 'non-stigmatized' race/ethnicity, impact the degree to which students possess academic self-confidence	$F(1,348) = 18.61$, $p<.001$	Stigmatized and nonstigmatized participants differed in academic self-confidence	$f = .2306$	(2) Slight Change Needed	$.1241 < f < .3396$	(3) Much Change Needed
24/4	Does membership in a 'stigmatized' race/ethnicity (African American or Latino) as compared to those in a 'non-stigmatized' race/ethnicity, impact the degree to which students are certain about the degree of their own self-confidence.	$F(1,348) = 18.61$, $p<.001$	Related, stigmatized students in our sample were significantly lower than nonstigmatized students in the certainty of their self-confidence ratings.	$f = .1930$	(2) Slight Change Needed	$.0872 < f < .3010$	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
26/1	Does the anticipation or prediction of future loneliness impact perserverence on tasks?	$F(2,37), 3.46$ $p<.05$	This analysis again showed significant variation among the three conditions. Participants in the future alone condition attempted the fewest problems. Again, the deficit was specific to feedback about social exclusion, insofar as participants in the misfortune control condition attempted as many problems (if not more) than the people in the future belonging condition. The decline in performance reflected both a higher rate of errors and reduced number of problems attempted. A diagnostic forecast of future social exclusion caused a significant drop in intelligent performance.	Cohen *f*: .4159	(3) Much Change Needed	.0000 <*f*<.7149	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
26/2	Does the anticipation or prediction of future loneliness impact cognitive abilities?	$F(2,37)$, $p<.01$	Hearing that one was likely to be alone later in life affected performance on a timed cognitive test. A diagnostic forecast of future social exclusion caused a significant drop in intelligent performance.	Cohen *f:* .5215	(2) Slight Change Needed	.1372 <*f*<.8318	(4) Complete Revision
27/1	Does culture, European-American or Asian-American) impact performance on a problem solving exam?	$F(1,74) = 2.50$, *ns*	The test revealed that there was no main effects of culture on the number of answers reported correctly.	Cohen *f:* .1837	(2) Slight Change Needed	.0445 <*f*<.4164	(3) Much Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
27/2	Does performance by Eastern Asian Americans differ when they work under 'think aloud' conditions or 'silent' conditions?	$t(32) = 2.67$, $p<.05$	East Asian American participants' performance was worse when they had to think aloud than when they were not thinking aloud. The results support the hypothesis that talking would interfere with East Asian American participants' performance.	Cohen *d*: .9134	(1) No Change Needed	.2069 <*d*<1.1/6199	(2) Slight Change Needed
27/3	Does performance by European Americans differ when they work under 'think aloud' conditions or 'silent' conditions?	$t(39) = .40$, *ns*	European American participants' performance, however, did not differ whether they were thinking aloud or not. The results support the hypothesis that talking would not interfere with European American participants cognitive performance.	Cohen *d*: .1258	(2) Slight Change Needed	-.4872 <*d*<.7388	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
28/1	Does praise on homework impact the amount of time spent on homework?	$t(59) = 9.788$, $p<.001$	Results revealed that students studied significantly more outside of the classroom when exposed to the verbal praise treatment than when exposed to the no verbal praise treatment. Although the results of this study may not generalize to all college student populations, they demonstrate the profound impact of properly administered verbal praise on college students' motivation to engage in homework.	Cohen *d*: 2.4881	(1) No Change Needed	1.8196 <*d*<3.1566	(1) No Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
28/2	Does praise on homework given throughout the course impact the performance on the end of course assessment?	$t(59) = 1.929$, $p>.05$, ns	Although the difference was not statistically significant (on the end of course exam), the direction of the means suggested that the students exposed to verbal praise not only studied more for each lesson but alos achieved more than those not exposed to verbal praise. In addition, my findings suggest that students who experience verbal praise for doing homework perform somewhat better on an instructor-created, criterion referenced final examination than those who experience no verbal praise for their homework habits.	Cohen *d*: .4800	(3) Much Change Needed	-.0292 <*d*<.9891	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	
			We did not find statistical significance for the overall rating.				
30/1	Do pre-service teacher's who have different supervision experiences have different attitudes toward their experience upon completion?	$t(30) = .67$, $p > .51$	Evidence presented here indicates that peer coaching is a feasible vehicle for instituting collaborative efforts; therefore, peer coaching warrants consideration as a potentially serviceable solution for strengthening field-based training of prospective teachers.	Cohen's *d*: -.7929	(3) Much Change Needed	-1.3018 < *d* < .2840	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
			Post treatment results showed statistically significant differences in favor of the experimental group for overall demonstration of clarity skills.				
30/2	Do pre-service teacher's who have different supervision experiences demonstrate differences in clarity skills?	$t(30) = 41.66$, $p<.001$	Evidence presented here indicates that peer coaching is a feasible vehicle for instituting collaborative efforts; therefore, peer coaching warrants consideration as a potentially serviceable solution for strengthening field-based training of prospective teachers.	Cohen's *d*: .8068	(2) Slight Change Needed	.5213 <*d*< 1.0874	(3) Much Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
32/1	Is a family intervention program effective in helping children gain vocabulary skills?	$F(1,247) = 32.08$, $p<.001$	When examining the effect of the interaction of group affiliation with time using repeated measures ANOVA we found that project EASE participants made statistically significantly greater gains than the control group on Vocabulary. It appeared from the posttest measures on the CAP vocabulary subtests that those students who participated in the intervention were better able to recall more superordinate terms which in turn have been shown to relate to the reading skills of elementary aged children. Because vocabulary knowledge, story comprehension, and story sequencing are precisely the language skills that relate most strongly to literacy accomplishments, the improvement on these measures strong confirms the relevance of the intervention to improved reading outcomes.	Cohen *f*: .3597	(3) Much Change Needed	.2309<*f*<.4878	(3) Much Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
32/2	Is a family intervention program effective in helping children gain sound awareness skills?	$F(1,247) = 7.45$ $p<.01$	When examining the effect of the interaction of group affiliation with time using repeated measures ANOVA we found that project EASE participants made statistically significantly greater gains than the control group on Sound Awareness	Cohen *f*: .1733	(3) Much Change Needed	.2309 <*f*<.4878	(3) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
			When examining the effect of the interaction of group affiliation with time using repeated measures ANOVA we found that project EASE participants made statistically significantly greater gains than the control group on Story Comprehension.				
32/3	Is a family intervention program effective in helping children gain story comprehension skills?	$F(1,227) = 6.85$, $p<.01$	The impact of participation in Project EASE on children's language scores is striking. Because vocabulary knowledge, story comprehension, and story sequencing are precisely the language skills that relate most strongly to literacy accomplishments, the improvement on these measures strong confirms the relevance of the intervention to improved reading outcomes.	Cohen *f*: .1874	(4) Complete Revision Needed	.0448 <*f*<.3288	(4) Complete Revision Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
32/4	Is a family intervention program effective in helping children gain language skills?	$F(1,246) = 35.46$, $p<.001$	Although all the children in the sample showed statistically significant gains in all three literacy composites over time, we were able to attribute a statistically significant gain in Language Skills to the Project EASE intervention.. The impact of participation in Project EASE on children's language scores is striking.	Cohen *f*: .3789	(3) Much Revision Needed	.2494 <*f*<.5077	(4) Complete Revision Needed
33/1	Does level of participation in a tutoring program impact student achievement in overall reading level?	$F(1,76) = 4.72$, $p = .03$	There was a statistically significant treatment effect. Overall, high level treatment children outperformed low-level treatment children in instructional reading level.	Cohen *f*: .2385	(2) Slight Change Needed	.0211 <*f*<.4669	(3) Much Change Needed

Appendix D

Summary of Analyses and Associated Statistics with Decision Made About Results

Study / Analysis Number	Summary of Issue to be Addressed by Analysis	Statistical Significance Reported	Findings/Results	Effect Size	Decision	CIs for Effect Size	Decision
33/2	Does level of participation in a tutoring program impact student achievement in reading words in isolation?	$F(1,71) = 5.09$, $p = .03$	There was a treatment effect for reading words in isolation. On average, for reading words in isolation, those who received longer treatment had higher word reading abilities overall.	Cohen *f*: .2476	(1) No Change Needed	.0300 <*f*<.4767	(3) Much Change Needed

Appendix E

IRB Exemption

UNIVERSITY OF
SOUTH FLORIDA

EXEMPTION CERTIFICATION

MEMO: Melinda R. Hess/ Jeffrey D. Kromrey, Ph.D.
College of Education, Department of Measurement and Research
EDU162

FROM: Institutional Review Board, PGS/amr

SUBJECT: Exemption Certification for Protocol No. 101759

DATE: September 15, 2003

On September 11, 2003, it was determined that your project entitled, **"Effect Sizes, Significance Tests, and Confidence Intervals: Assessing the Influence and Impact of Research Reporting Protocol and Practice,"** meets federal criteria to qualify as an exempt study.

Because the study has been certified as exempt, you will not be required to complete continuation or final review reports. However, it is your responsibility to notify the IRB prior to making any changes to the study. Please note that changes made to an exempt protocol may disqualify it from exempt status and may require an expedited or full review.

All research, regardless of the type of IRB review received, must be conducted in a manner that is consistent with the ethical principles of your profession and the federal guidelines for the protection of human subjects. As principal investigator, it is your responsibility to ensure subjects' rights and welfare are protected during the execution of this study.

The Division of Research Compliance will hold your exemption application for five years. At least 90 days before the end of the fifth year, you will be notified that your file will be closed. If your project is still ongoing, you will need to contact the Division of Research Compliance upon receipt of that letter and follow the instructions for completing a new exemption application. It is, therefore, important that you keep your address current with the Division of Research Compliance. If you have any questions, please contact the Division of Research Compliance at 813-974-5638.

Version date 8/19/03

OFFICE OF RESEARCH, DIVISION OF RESEARCH COMPLIANCE
INSTITUTIONAL REVIEW BOARDS, MPA NO. 1284-01/M1284-02XM
University of South Florida • 12901 Bruce B. Downs Blvd., MDC 035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-5618

About the Author

Melinda Hess received her Bachelor of Science degree in Electrical Engineering at the USF in 1986 while on an Air Force Reserved Officer Training Corp Scholarship. Commissioned in May 1986, she honorably served 11 years in the Air Force, serving in locations worldwide. She earned her Master's Degree in Management in 1990 from Webster University.

Melinda entered into education in 1997 teaching mathematics and beginning her doctoral studies. Additional experiences include consulting for a local school district and co-teaching graduate level educational research courses. She has presented her research at professional and technical conferences at state, regional and national levels and has been nominated for the Florida Educational Research Association's Distinguished Paper award three times, winning the award once. Additionally, she has interned with the Educational Testing Service, edited the *Florida Journal of Educational Research*, and served as President of the USF Graduate Research Association for Professional Enhancement.