LONGITUDINAL STABILITY OF EFFECT SIZES IN EDUCATIONAL RESEARCH

JOSHUA STEPHENS

Bachelor of Science in Natural Resources

Ohio State University

December 2000

Master of Science in Natural Resources

Ohio State University

June 2003

Master of Education

Cleveland State University

August 2005

submitted in partial fulfillment of requirement for the degree

DOCTORATE IN URBAN EDUCATION

at the

CLEVELAND STATE UNIVERSITY

MAY 2013

DEDICATION

I dedicate this work to my loving and supportive wife. She has never given up on me, regardless of how ridiculous my plans seemed and she deserves much of the credit for this project. She has been the one to push me to go and write even when I didn't want to. She has been the one who has agreed to parent our five children alone while I seclude myself in some coffee shop or library. And she has been the one I have kept in mind when I considered quitting. We have never quit on each other and I would never quit on this project as long as I knew that I had her support.

I would also like to dedicate this work to my children who have, often without realizing it, provided me with the energy to keep writing. As I have watched them grow and achieve their own academic goals, I have often been reminded that there is always more to learn. I hope they are proud of our accomplishment and they can see that knowledge and dedication are the keys to a happy life.

ACKNOWLEDGEMENTS

In undertaking this task, I have relied heavily on the support of many people. All of them deserve some sort of recognition. While it is impossible to accurately account for everyone who has helped me achieve this goal, I will endeavor to highlight those who I feel have been instrumental and, without whom, I would have certainly failed.

First, I would like to acknowledge the guidance and support of my chairman, Dr. Jeremy Genovese. I first met Dr. Genovese as a master's student. When I began the doctoral dissertation writing process, I realized that if I could produce a work that was acceptable to him that I would have truly pushed myself to new academic heights. We began to talk about educational fads, research methodology, and pseudo-science. It is remarkable how a nebulous conversation between almost six years ago has turned into this project. Without Dr. Genovese's assistance, I would have been mired in confusion and frustration. He was always able to help me sort out a difficulty or elucidate a poorly understood idea. His intellectual prowess remains an inspiration to me.

Second, I would like to acknowledge the members of my committee. Dr. Stead has been instrumental in assisting me in understanding the methodological elements of this work. His comments have been highly valuable. Dr. Harper helped me to remain motivated. Many times Dr. Harper has seen me in the halls of Julka asked me when a manuscript would be forthcoming and reminded me to just get this thing done. Dr. Hamlen provided advice to me on how to proceed with writing my dissertation and what to do when I was finished. This advice has stayed with me through this process and I thank her for it. And, lastly, to Dr. McNamara, I wish to extend my sincerest thanks for agreeing to serve on my committee. Her expertise and insight has been much appreciated.

Special thanks are also owed to members of my family. First, my wife has been my constant support throughout this process. She has managed a household of five children, often times alone, with grace and expertise in order to permit me the stolen hours I needed to write this document. Second, Nancy Smialek deserves special consideration. Beyond supporting me, my mother-in-law has watched my children and cooked meals for us, thus giving me the precious time needed to complete this project. Third, my parents and grandparents have inspired me and motivated me throughout this process and I would like to thank them for everything they have done. Fourth, Tom Hall, PhD, whom I respect deeply, granted me editing assistance to help me make this project what I wanted it to be.

Thanks to all of you for what you have done. While no man is an island, I am fortunate enough to be surrounded by a sea of supportive people who have helped me in innumerable ways throughout these years.

LONGITUDINAL STABILITY OF EFFECT SIZES IN EDUCATIONAL RESEARCH

JOSHUA STEPHENS

ABSTRACT

Effect sizes are the statistic generated by meta-analyses, a commonly used statistic in education research. Meta-analyses are widely used by education practitioners, administrators, and policy makers as a means to decide best classroom or school practices. It has been suggested by authors in other fields, most notably Jennions and Moller (2001), that effect sizes have declined over time due to various sources of bias. This paper examines the question of whether shifting effect sizes can be observed in educational research and attempts to explain possible causes of this observation. It uses the methodological framework used by Jennions and Moller (2001) and applies it to educational meta-analyses conducted from 1970 to the present. It finds that, contrary to the findings of Jennions and Moller (2001), that effect sizes in educational research have increased over time. Likely explanations regard systemic bias in the conduct and publication of educational research. The paper concludes with recommendations for future research to examine causal factors contributing to this phenomenon.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

Prevalence of Declining Effect Sizes in Educational Research

> An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganised. We need a sort of clearing-house for the mind: a depot where knowledge and ideas are received, sorted, summarised, digested, clarified and compared.
>
> —H. G. Wells, 1938

"Philosophers of science state that an area of knowledge becomes a science when it accumulates both: (a) a body of applicable facts and theory and (b) agreement on methodologies of research and inquiry that will produce replicable observations among observers over time" (Asher, p. 144 – 145, 1990). Educational researchers have certainly developed the first of these two points. However, it is the degree to which we have perfected the second of these two points that is the focus of this paper.

Meta-analysis is a statistical technique where many studies are analyzed together in order to determine the effect of a particular intervention or phenomenon. Gene Glass first described its use in the social sciences in 1976 as a way to make meaning out of a sometimes overwhelming array of studies and data in psychology (Asher, 1990). Glass (1976) describes it thusly:

> Meta-analysis refers to the analysis of analyses . . . the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating findings. It connotes a rigorous alternative to the causal, narrative discussions of research studies that typify our attempts to make sense of the rapidly expanding literature. (p.3)

Since then, the technique is widely used in a variety of fields, including education.

Prior to the advent of meta-analysis in the social sciences, quantitative methods of research synthesis were generated and later found to be lacking for various reasons (Rosenthal, 1978). Despite these attempts at quantitatively synthesizing research results, literature reviews tended to be done mostly qualitatively and presented the researcher's best effort at categorizing, summarizing, and describing the current state of a particular field of study. However, with the ever - increasing number of studies on particular topics, (Littell et al., 2008) the literature review becomes less and less meaningful and a better quantitative method of cumulating and synthesizing past research is needed. Asher (1990) concludes, "the traditional literature review in the first chapter of doctoral dissertations is somewhat obsolete. . . " (p.148).

Meta-analysis has emerged as one of the primary methods of synthesizing and understanding large bodies of literature. However, a troubling phenomenon has been observed by meta-analysts in ecology, medicine, and other natural sciences. The effect sizes generated via meta-analyses tend to decrease over time. Thus, what once appeared to be true and well supported by research apparently diminishes in validity over time. Often treatments, be they medical, ecological, or other, have been prescribed based on initial meta-analyses. However, it now appears that it is possible that these initial prescriptions may not have been as robust as once believed.

In education, practitioners, policy-makers, and the public continually search for solutions to various pressing issues. This leads to a focus on what interventions will be effective and what evidence supports these claims (Gibbs, 2003). Often the evidence cited comes from meta-analyses that may be subject to the declining effect size phenomenon. Thus, the purpose of this study is to determine whether diminishing effect sizes are present in educational meta-analyses and, if present, to account for this. This study utilizes a model where meta-analyzed educational phenomena are assessed to determine if effect sizes have decreased over time.

1.1 Statement of the Problem

This study attempted to address the following research question: Do effect sizes in published meta-analyses of educational research tend to diminish over time?

1.2 Definitions

The following definitions will be used throughout this paper:

Bias.

Bias is described as the "deviation of results or inferences from the truth, or processes leading to such deviation; or any trend in the collection, analysis, interpretation, publication or review of data that can lead to conclusions that are systematically different from the truth" (Song et al., 2000, p. 16).

Confirmation bias.

A fundamental issue relating to the synthesizing literature has to do with confirmation bias. Confirmation bias is the tendency of individuals to search for information that confirms their pre-determined beliefs and discount information that contradicts those beliefs (Bushman & Wells, 2001). It is rooted in issues relating to how individuals settle cognitive dissonance.

When individuals process new information, they tend to cling to their preconceived ideas, particularly when the new data are complex (Anderson & Sechler, 1986, Hilton & von Hippell, 1990). Confirmation bias can play a heavy role in narrative literature reviews. As researchers examine and synthesize a body of literature, deeply held psychological processes can sway how they process and agglomerate that literature.

Confirmation bias has its roots in the psychology of human judgment. Francis Bacon first described the phenomenon in 1621 (Bacon, 1960). It is easier for humans to see how evidence supports a position as opposed to how that evidence contradicts a position.

Confirmation bias is demonstrated by a finding that shows how peer reviewers have been are biased against manuscripts they were asked to review if the results of the manuscript contravened their particular perspectives (Mahoney, 1977).

Confirmation bias can also be a factor in individual studies as researchers may not report all outcomes or may only report selected outcomes (Dickersin, 2005). This creates the illusion of a body of literature that generally agrees with the point that the researcher is attempting to make and deludes the reader into a false sense of clarity. Even if a researcher includes studies illustrating contravening evidence to his or her pre-determined position, a qualitatively conducted literature review allows the researcher to deemphasize the findings of studies that contravene that pre-determined position. This is an even more nefarious sort of ambiguation as it provides a false sense of objectivity to a literature review that is, in fact, biased.

Database bias.

This form of bias, sometimes referred to as indexing bias, occurs when literature databases may not include all published studies on a given topic (Felson, 1992). This may

manifest as a tendency for low-circulation journals that are often not indexed in commonly used databases to publish studies with negative or non-positive results (Egger & Davey-Smith, 1998). Several studies have suggested that there is a greater possibility of missing relevant studies by searching a single electronic database (Dickersin et al., 1985; Gotzsche & Lange, 1991; Silagy, 1993). This will bias future literature searches as the databases provide systematically different results than what research has actually been conducted (Song et al., 2000).

Dissemination bias.

This is a type of bias that occurs when the dissemination profile of a study's results depends on the direction or strength of its findings (Song et al., 2000).

Dissemination profile.

This term refers to the accessibility of research results that range on a spectrum from completely inaccessible to readily accessible (Song et al., 2000).

Full publication bias.

This type of bias refers to the situation when research results may be presented at a conference but may not be fully published due to not finding significance (Song et al., 2000). At issue is the fact that abstracts presented at conferences will present only limited data while full publications will provide many more details. DeBellefeuille et al. (1992) found that abstracts with positive results were more likely to be presented at conferences and published in full than studies with negative or neutral findings. However, other studies have found that there is no association between study outcome and full publication (Scherer et al., 1994; Chalmers et al., 1990; Landry, 1996).

Grey literature bias.

Grey literature refers to reports, working papers, dissertations, and conference abstracts that often have very limited dissemination (Auger, 1998). Grey literature bias occurs when the results of these studies are systematically different from those reported in peer-reviewed journals (Song et al., 2000). In psychology and education, this trend seems to be even more pronounced. Several studies have found that theses and dissertations in these fields tend to be published more often if they show significant results and also that the average effects reported in journal articles tended to be higher than those found in theses or dissertations (Smart, 1964; Smith, 1980; Glass, McGaw, & Smith, 1981; White, 1982). Results from the field of medicine show a similar bias (Detsky et al., 1987; Devine, 1999). McAuley et al. (1999) sampled 135 meta-analyses, 38 of which included grey literature. They then found that those meta-analyses that included grey literature showed a diminished effectiveness of treatments by approximately 12%.

Hot stuff bias.

Hot stuff bias refers to the occurrence of researchers and investigators publishing results regarding topics that are popular, despite the fact that these results may be only weakly supported (Sackett, 1979).

Language bias.

This bias, sometimes referred to as the Tower of Babel bias, refers to the difficulties faced by non – English speaking researchers in getting published in the most prestigious international journals, most of which are published in English (Gregoire, Derderian, & Lorier, 1995, Vandenbrouche, 1989, Bakewell, 1992). Other research has demonstrated that English language journals tended to print statistically significant results more often than German language journals (Egger et al., 1997). Gregoire et al. (1995) found that meta-

analyses conducted in English only tended to underreport effect sizes, thus making language bias a particularly concerning issue for meta-analysts.

A related phenomenon to language bias is country bias where there is a difference in the reporting of results between different countries (Song et al., 2000). One study observed that estimated efficacy of a medical treatment was greater for studies published in English language journals published outside of the United States than it was for studies published inside the United States (Ottenbacher & Defabio, 1985). Other studies have found that results of acupuncture efficacy studies published in certain Asian countries varied significantly from the results published in Western European and North American countries (Vickers et al., 1998).

Meta-analysis.

This is the quantitative attempt to synthesize research using effect sizes as a means of comparison.

Multiple publication bias.

Multiple publication bias is a multiplier-effect bias where studies that support previously supported studies or those studies that find significant results tend to generate multiple publications (Song et al., 2000). This has been described as either overt or covert (Tramer et al., 1997) and may be very difficult to detect across various publication venues and in different times. Overt duplication is when results are re-analyzed and re-published in a different format and the original research is appropriately cited. Cover duplication is when the same results are simply republished in a different place or time without proper citation to the original publication.

Though the practice of publishing the same results in multiple journals has been criticized for wasting resources (Angell & Relman, 1989), it has also been argued that publishing the same data in different ways may assist in the dissemination of important results (Song et al., 2000). Studies with significant results were more likely to generate multiple publications and more likely to be published in more widely circulated journals (Easterbrook et al., 1991).

Narrative review (or narrative literature review).

This refers to an attempt to identify literature on a particular topic to generate a base upon which to conduct further research (Davies, 2000).

Outcome reporting bias.

This sort of bias refers to when a study measures multiple outcomes but reports only those that are significant. This is a particularly vexing issue for meta-analyses as those results may be biased by including only those outcomes that are significant while ignoring or not reporting those that are non-significant (Song et al., 2000).

Positive results bias.

Positive results bias refers to the tendency of authors to submit and for editors to accept positive research results over null results (Song et al., 2000). This phenomenon was first observed by Sterling in 1959 in a study that demonstrated that 97% of studies published in four major psychological journals were statistically significant, thus supporting the conclusion that studies with non-significant results were underreported (Sterling, 1959). Since that time, a number of studies have demonstrated that studies that show statistical significance tend to be published at a greater rate than those that do not show significance (e.g. Begg, 1994; Dickersin, 2005; Scherer, Langenberg, & von Elm, 2007; Torgerson, 2006;

Littell et al., 2008; Smart, 1964; Bozarth & Roberts, 1972; Greenwald, 1975; Hubbard & Armstrong, 1997; Davidson, 1986; Moscati et al., 1994; Moher et al., 1994; Mulward & Gotzsche, 1996; Csada et al., 1996). In 1995, Sterling and his team concluded that in practices leading to publication bias have changed little in more than three decades (Sterling et al., 1995).

Moreover, this trend seems to be more pronounced in psychology than in other disciplines (Song et al., 2000). Greenwald (1975) found that the probability of psychological researchers submitting their research results for publication if they were significant was 0.49 while the probability of submitting non-significant results was only 0.06. Coursol and Wagner (1986) found that among members of the American Psychological Association, 66% of studies with significant findings were published while only 22% of studies with non-significant findings were.

Power.

Power is the probability that a population difference that is actually different from zero will be detected by a particular test (Carlton & Strawderman, 1996).

Place of publication bias.

This form of bias is based on the certain popular journals tending to be more likely to publish particular studies due to editorial policy or reader preference (Ben-Shlomo & Davey-Smith, 1994). Simes (1987) found that journals with lower circulation tended more to publish studies with negative results than journals with higher circulation.

Publication bias.

This type of bias occurs when the publication of research results depends on their nature and direction (Dickersin, 1990).

Research synthesis.

Research synthesis is a term describing a family of methods, of which meta-analysis is one, for summarizing, integrating, and cumulating results of a group of different studies on a particular research topic or question (Davies, 2000).

Time-lag bias.

This sort of bias (sometimes referred to as the pipeline effect) occurs when studies that demonstrate significance are published earlier than those that demonstrate non-significance (Rosenthal, 1988; Jadad & Rennie, 1998). In examples from medicine, it was found that the time between the approval and inception of a study and its publication was significantly different between those studies finding significance and those that did not. (Stern & Simes, 1997; Ioannidis, 1998; Misakian & Bero, 1998; Rothwell & Robertson, 1997; Song & Gilbody, 1998).

Retrieval bias.

This form of bias occurs when there is a difference between quantitatively produced estimates based on all studies as opposed to an estimate based on only retrieved studies. This is often due to the fact that unpublished studies are not retrieved (Song et al., 2000).

Media attention bias.

This occurs when research results that receive media coverage generate new studies and new citations (Song et al., 2000). Mass media tends to over-emphasize sensational stories and this may lead researchers to study these phenomenon in order to appeal to a broader audience. A related observance is that mass media outlets tend to over-report positive findings from academic journals that may then impact how future studies are conducted and reported (Koren & Klein, 1991).

1.3 Significance of the Study

There are several reasons why this research is pertinent. First, the primary concern of this study is with the application of increased rigor into educational literature. It is of the utmost importance that educators and educational policy-makers utilize practices and policies based on the strongest empirical evidence. Since public school funding is a limited resource, it is important for that funding to be spent wisely and not on ineffective innovations. This research is particularly important in education because of the politicized nature of public education funding. Since public funding finances education, it is important that curricular and programming innovations that are implemented in public schools be well-reasoned and researched in order to avoid wasted funding and energy. Meta-analysis promises to provide a statistic that can simplify a complex set of information into one figure. However, if the statistic does not provide meaningful information, then the educational innovations based on that statistic will result in misdirected funds and energy.

Second, this research will fulfill a need for stricter accountability for education researchers who may be misapplying meta-analysis to their research questions that could be better answered by another statistic. The field of education research is wrought with low-quality research. This finding raises serious questions regarding the validity of meta-analysis. If a meta-analysis combines many studies that are of low quality, then the output of the analysis is questionable. Thus, this study will assist education researchers in evaluating whether meta-analysis or another form of research synthesis is best.

Third, this research will assist educators to assess the value of meta-analyzed studies to their practice. It has been suggested in the field of social work, a field many find similar to education, that practitioners often lack the capabilities to locate, evaluate, and utilize

academic research (Shlonsky et al., 2011). As such, it becomes imperative that techniques of research synthesis can be utilized and understood by practitioners but these techniques must help practitioners to reliably and accurately understand complex fields of research.

Fourth, researchers in other fields may be able to use this research in order to determine whether meta-analysis is appropriate for their purposes. Horder (2001) writes that "'science' must be acknowledged as being a historical edifice: it not only consists of the latest results, but, more accurately, it is composed of the sum total of a massive accumulation of earlier-acquired data, interpretation and assumptions" (p. 124). This paper is an attempt to add to this "accumulation" and work towards greater degrees of rigor in educational research.

1.4 Limitations of the Study

This study is limited in its utility in several ways. First, as with any meta-analysis, there may be an inadequate selection of studies to analyze. This is potentially a very serious problem as it could seriously bias the results. Second, there are many software packages that can be used to conduct meta-analysis. Since it is beyond the scope of this paper to evaluate each of the merits of these various programs, a potential limitation is that METAWIN, the software package selected for this study, has flaws that could affect final results.

Chapter II

Review of the Literature

With the rising importance of meta-analysis and its emphasis on effect size, it becomes imperative that the research question addressed by this study is more fully examined.

2.1 Educational Fads

Short-lived educational policy or pedagogical trends are much-bemoaned facets of American education. These tend to be attempts to fix pressing and entrenched issues in the public schools through a new approach that is often not well supported by research. These attempts are undertaken with the best of intentions and may be an attempt to correct problems created by earlier fads (Chaddock, 1998). The primary issue with education fads is that they waste limited financial and academic resources, thus undermining the success of students.

Fads tend to be things that appear original but so commonsensical that individuals are struck by their apparent truth (Birnbaum, 2000). “The case is put so simply, forcefully, and

fashionably that any other view sounds untenable, or even politically incorrect. The clarity of the message can lull the listener into uncritical acceptance. Since everybody is saying these sorts of things, surely they must be right" (Hilmer & Donaldson, 1996, p. 6).

Characterizing fads raises some interesting points. Fads are products in which some person or entity has an interest selling. Those people or entities have vested interests in promoting their particular product (Birnbaum, 2000). A narrative is typically used to justify a fad's utility (Roe, 1994). Fads are given the air of certainty and comfort that will assist individuals or organizations in dealing with various difficulties using rhetoric in order to convince an audience of its validity. They are generally derived from knowledge and simplified so as to be communicated to a broad audience (Tornatzky & Fleischer, 1990). Fads are introduced into an organization and ultimately they do not infiltrate throughout the entire system and are ultimately rejected. This complicates the task of preemptively identifying which innovations will remain and which will fade away. Fads have also been characterized as memes as they are ideas that act parasitically in an organization even if they do no good to the host. However, even if the fad does no good it may produce a placebo effect, as an organization will improve coincidentally with the implementation of the fad. This can cause policymakers or administrators to falsely attribute success to the fad (Birnbaum, 2000).

"[T]he history of education is blotched by both faddish ideas and methods that don't work and by persistent failure to institutionalize ideas and methods that *do* work" (Kozloff, 1992). Kozloff describes a broad range of educational innovations that he sees as having had a pernicious impact on education, such as: whole language, invented spelling, inquiry learning, discovery learning, learning styles, multiple intelligences, brain-based teaching,

constructivist math, portfolio assessment, authentic assessment, journaling, self-esteem raising, learning centers, sustained silent reading, developmentally appropriate practices, balanced literacy, and student-centered education. He holds that these fads are products of both romantic modernism and progressivism and that these forces have combined in modern, public education to push out what has been proven to work in favor of newer ideas promulgated by ideologically or financially motivated schools of education and others who stand to profit when school districts adopt their ideas (Kozloff, 2002). While his *ad hominem* criticisms are impossible to prove, his point does bear merit that there is much to gain by interested parties when schools adopt new curricular models or other policy interventions.

Rigby (1998) points out that fads have the following negative impacts: imbalances in strategic resources, internal divisiveness, unrealistic expectations, and loss of employee responsibility. However, Birnbaum (2000) indicates that fads may also have the following positive effects: recognizing the importance of data, emphasizing alternative values, producing variety, diversifying interaction, and promoting activity. Thus while fads are generally viewed as negative, it is possible to find some value in them.

2.2 Single-Studies

The single study is a well-established form of basic research in all scientific disciplines. It provides information on one particular phenomenon and attempts to answer a particular question. While the single study is the building block of all scientific endeavors, it is not well-suited to all purposes. Single studies have been criticized as being inferior to research syntheses on a number of counts. The criticism of single studies that bears the most relevance to this work involves generalizability. Cook et al., (1992) states that single studies "are limited in the gereralisability of the knowledge they produce about concepts,

populations, settings, and times" (p.3). This lack of generalizability is particularly problematic in fields such as education where the diversity of research subjects is at times staggering. Research syntheses, such as meta-analysis, on the other hand, can lead to levels of generalizability not possible with single studies (Cooper & Hedges, 1994).

There are a host of other concerns regarding single studies. For instance, one particularly poor method of conducting single studies is the one group pre- and post-designs without a control group. Lipsey and Wilson (1993) found that this design, combined with publication bias, tends to inflate mean effect sizes. While it is beyond the scope of this research to explicate every shortcoming of single studies, it is important to note that there are many potential shortcomings that can alter the findings of narrative literature reviews or meta-analyses.

2.3 Significance Testing

Tests of statistical significance or non-significance do not necessarily suffice to fully describe a phenomenon (Glass, McGaw, & Smith, 1981). This simple evaluation technique that is commonly taught to undergraduate and graduate statistics students as the primary way to evaluate differences between samples is not sufficient in the evaluation of a null hypothesis in a single study as it lacks the power needed to accurately answer research questions. The weaknesses of statistical significance testing are compounded when synthesizing and cumulating a body of literature.

As the number of studies increases, paradoxically, the power of the t-test diminishes. "When statistical significance is used as the criterion and *more* studies are available for review (i.e. as *K* increases), then it is *less* likely that there will be detection of a true population difference" (Carlton and Strawderman, 1996, p. 69). Testing for statistical

significance, as the number of relevant studies in a field increases, makes detecting differences between control and treatment groups less likely and makes reproducibility of results increasingly difficult (Hedges & Olkin, 1980, 1985). Aggregating tests of statistical significance can even lead to contradictory conclusions as primary studies examined individually may seem to indicate one finding but their aggregated significance scores contraindicate that finding (Glass, McGaw, & Smith, 1981). Thus, the use of statistical significance to estimate robustness in large bodies of studies is an inherently weak way to determine relationships between variables. This has led researchers to look for ways to determine these relationships with greater validity.

2.4 Meta-Analysis

Meta-analysis is an attempt to use a common measure, generally effect sizes, to generate a statistically more powerful answer to a research question based on a body of literature. Effect sizes tell researchers the magnitude of relationships between variables. In a meta-analysis, effect sizes are calculated for each study, weighted by sample size and study quality, and then averaged to produce an overall effect size (Littell et al., 2008). While typical data analysis uses multiple observations of a phenomenon as data points, meta-analysis uses multiple studies as data points (Wolf, 1986, Littell et al., 2008). Meta-analyses reanalyze data from original studies to generate effect sizes and then analyzes these effect sizes to examine trends (Littell et al., 2008). As researchers comb through the available literature, code it, and account for differences between studies, some researchers believe that their literature reviews become stronger than those done in a qualitative or narrative fashion (Asher, 1990). This technique can minimize sampling error and bias by synthesizing a large and complex body of research in a robust and methodical manner. Many researchers believe

that this "very powerful" and "relatively simple" statistical technique holds great promise (Asher, 1990). The overall effect size, generated through meta-analysis, is considered by some to be a more robust way to answer research questions in a variety of fields, including education.

Brief history of meta-analysis.

Meta-analysis began in the 1930s as agricultural researchers attempted to combine studies to draw more meaningful data from the large pool of information published in their field (Wolf, 1986). Four decades later, Gene Glass applied some of the same techniques to psychological data (Asher, 1990). Glass is credited with coining the term in a 1976 speech and it was quickly adopted throughout the social sciences. Researchers felt a pressing need for methods to organize and evaluate the exploding number of research reports published after WWII (Chalmers, Hedges, & Cooper, 2002; Wachter, 1998; Glass, McGaw, and Smith, 1981). Since that point, there have been a large number of meta-analyses conducted on a broad array of topics (Wolf, 1986).

Purpose of meta-analysis.

Meta-analysis can serve two primary functions: theory building and assisting in the formation of best practices.

Asher (1990) states, "meta-analysis results should be the primary basis of theory building" (p. 148). Meta-analysis is appropriate for a broad range of statistical applications including: "synthesizing research on correlations, epidemiological data (incidence and prevalence rates), accuracy of diagnostic tests, prognostic accuracy (etiological and risk factors), and treatment effects" (Littell et al., 2008, p. 5). One of the primary strengths of meta-analysis is that it illuminates type II errors by assisting researchers in showing

examples of where conclusions may seem more significant than they actually are. Meta-analysis's focus on effect sizes assists researchers in moving beyond simply testing for significance, a statistical test that is prone to a set of particular weaknesses. Meta-analysis's focus on the standardization of data imposes a set of criteria on a particular topic to make that topic easier to grasp and understand. In fields such as education, where the research enterprise "is a rough-hewn, variegated undertaking of huge proportions" (Glass, McGaw, and Smith, 1981, p. 12) this theory building power of meta-analysis is all the more important.

Meta-analysis can also be used as a tool to discern best practices in a variety of fields. Meta-analysis is now the primary statistical tool used by the biomedical sciences as it is can reliably be used to discern treatment effects across a large number of studies (Littell et al., 2008). The field of education, similar to the fields of medicine, social work, psychology, and others, has moved towards evidence-based practice. Evidence-based practice in education means "integrating individual teaching and learning expertise with the best available evidence from systematic research on educational interventions and practice" (Davies, 2000, p. 11). The integrative property of meta-analysis makes it a potentially powerful tool for deciding upon which educational interventions and policies are most effective. With recent emphasis being placed on evidence-based practices in education, (Sackett et al., 1991) meta-analysis has taken on a greater prominence as educators, administrators, and school boards have sought to provide quantitative evidence for various proposals and policy prescriptions. Indeed, the potential for research syntheses to inform policy and practice is great and not fully explored (Chalmers, Hedges, & Cooper, 2002). Some researchers (see Feinstein, 1995) point out, however, that such pooled and aggregated data, while appropriate for policy makers, is not useful to actual practitioners as practitioners tend to not have the statistical

expertise to utilize the results of meta-analyses and most meta-analyses are not interpreted for practitioners in a way they could use.

The use of meta-analytic techniques in education is quite widespread and has become widely accepted, markedly increasing since the 1990s (Littell, et al., 2008). Kulik and Kulik (1989) reviewed 150 meta-analyses of various educational interventions. Lipsey and Wilson (1993) reviewed 302 meta-analyses in the social sciences, two-thirds of which dealt with some educational phenomenon. Both of these works indicate that most educational interventions have a moderate effect. Neither large effects nor negative effects are frequently observed in education. Given the gap between the knowledge and research needs of education practitioners and educational researchers, meta-analysis becomes increasingly attractive (Hargreaves, 1997; Hillage et al., 1998; Tooley & Darby, 1998) and it appears that it will continue to grow in popularity.

Conducting meta-analysis.

Crafting well-done meta-analyses is of the utmost importance. Many issues play into doing so and they will be overviewed in this section. They include: issues of classification; inter-coder reliability;

Glass, McGaw, & Smith (1981) discusses one of the key issues in the crafting of meta-analysis, namely that of classification of variables. When researchers are collecting research to be considered in the meta-analysis, the classification of variables can become a highly complex process. In Glass and Smith (1977), the researchers reviewed over 400 psychotherapy single-case studies in order to determine the efficacy of psychotherapy. They had to code each study to determine the particular type of psychotherapy used. This coding process became more complicated than they initially thought, as many of the primary studies

did not follow the same coding scheme for types of psychotherapy used by the meta-analysts. To overcome this difficulty, Glass and Smith employed a team of 25 clinicians and researchers to independently review the studies in order to bolster the validity of the conclusions drawn from the meta-analysis. These reviewers agreed to group types of therapies into broader categories, thus making the classification system simpler. This simplification, however, comes at the loss of some data specificity. This case illustrates one of the complexities to which meta-analysts must attend.

Reliability generally refers to consistency of measurement and, in a meta-analysis, specifically refers to inter-coder reliability (Glass, McGaw, & Smith, 1981). One of the key phases in the meta-analytic process is to have independent raters code different variables in each of the studies to determine either eligibility for inclusion or other features of the studies that may be pertinent to the meta-analysis. Issues of reliability emerge when different coders reach differing conclusions, thus complicating the work of meta-analysis. To diminish this problem, the following solutions are recommended: providing coders with explicit instructions, specifying defining characteristics of variables to be studied, rigorously attempting to be as thorough as possible, using multiple judges for each study, correction of flagrant issues that arise between judges. However, there are limits to the *a priori* preparation that can be imposed on coders. Despite these limitations, it is critical that coders be prepared as thoroughly as possible and that reliability testing procedures are enacted.

Advantages of meta-analysis.

While studies conducted in the natural sciences allow for definitive conclusions to be drawn based on clear, empirical data, studies in the social sciences are often too complicated and too fraught with unexpected and unaccounted variables for this to be the case (Wolf,

1986; Cook et al., 1992). Social science research, and education research in particular, is often disorganized and contradictory and, as such, it is necessary for the more systematic approach offered by meta-analysis to provide a less biased assessment of evidence (Littell, 2005). Studies published on the same research topic will reach opposing results and be reported in different journals (Wolf, 1986), thus allowing for the possibility of different readers reaching differing conclusions on the same topic and both readers being able to cite research as their source. As such, a single study of a topic is rarely sufficient to make conclusions. This necessitates a research synthesis of some sort.

Researchers have traditionally relied on qualitative literature reviews for the task of organizing and synthesizing the current state of the literature. However, the traditional literature review has serious flaws that can limit the accuracy of the conclusions drawn. These flaws were discussed in a prior section. So researchers have now attempted to move to meta-analytic techniques to draw quantitative conclusions that may be more valid (Davies, 2000). The advantage of meta-analysis over the traditional literature review is the focus of this section.

Meta-analyses give better parameter estimates of treatment effects than traditional literature reviews or vote counting. This increased validity stems from the fact that by combining results of multiple studies, meta-analyses increase the statistical power that can be used to detect significant effects (Littell et al., 2008). Just as it is true that single case studies provide less robust information than examining multiple cases, the meta-analysis, with its reach extended to many studies, provides better estimates of the parameter in question (Littell et al., 2008).

Traditional literature reviews cannot accurately account for moderating effects, such as treatment, participant, or study design characteristics that influence the parameter in question (Littell, et al. 2008). As opposed to a single study, rigorously conducted research syntheses of multiple studies can produce more robust estimates of treatment effects and can be highly useful for estimating program impacts (Littell, 2008). One of the strengths of meta-analysis is that it is capable of utilizing between-study variations to describe moderators of treatment effects to assess moderating effects that may not have been possible to assess in a single study (Littell et al. 2008). Meta-analysis can do this in a rigorous, scientific way that other researchers can follow and understand and attempt to replicate. Inconsistent findings between studies can indicate moderating effects not obvious under the traditional literature review.

While many people believe that meta-analyses require large number of studies, Littell et al. (2008) states that meta-analysis can be used with a minimum of only two studies and can accurately be conducted with studies that have both small and large sample sizes.

Criticisms of meta analysis.

Many authors have levied criticisms at meta-analysis. Some early attempts at meta-analysis were challenged as having only the veneer of rigor and validity. Eysenck (1978) called the process "mega-silliness." Shapiro (1994) called it "smeta-analysis." Feinstein (1995) called meta-analysis "statistical alchemy for the 21st century." Meta-analysis is limited by the quality of the research question, the quality and completeness of the literature upon which the meta-analysis will be based, and data searches used to find that literature (Davies, 2000). Meta-analysis, as is the case with most research techniques, can be abused and misapplied, both intentionally or unintentionally (Littell et al., 2008). Hedges (1986)

writes “Meta-analysis has become the latest fashion in some circles. As you might expect, there is a great deal of garbage being published . . . Many of the meta-analyses are ill-conceived, poorly executed, and minimally interpreted” (as quoted in Asher, 1990). Just like any other statistical or analytical tool, meta-analysis can be misused and misapplied to construct invalid and misleading results. In one particularly disturbing example, Jorgensen, Hilden, & Gotzsche (2006) found that meta-analyses funded by pharmaceutical companies tended to have results biased towards the success of the drugs being tested. While this may be an extreme case, it indicates the need for the rigorous application of scientific review to any published work. High-quality meta-analyses require specific research questions, well considered populations to study, and clearly defined outcomes to be assessed (Davies, 2000).

Dissimilar data.

One criticism deals with the notion that logical conclusions cannot be drawn by comparing and aggregating dissimilar data. This is sometimes referred to as the “apples and oranges” problem (Glass et al., 1981; Slavin, 1984; Wolf, 1986). This is fundamentally a problem relating to the inadequate conceptualization of the problem (Littell et al., 2008). In a large pool of studies, there will be such a myriad of population variables, differences in research methods, data analysis methods, and data interpretation paradigms and techniques that will confound any attempt to meaningfully combine these studies. This criticism is particularly relevant when a researcher is primarily interested in only one of the variables being studied, as the statistical interference of the other variables will decrease the validity of the results. When the researcher is interested in studying many variables and interaction effects, then this problem is much minimized since meta-analysis is such a powerful tool for dealing with multiple variables and interaction effects (Littell et al., 2008). Furthermore,

through careful coding and incorporation of key variables into the analysis, this problem can be much minimized. Careful and deliberate coding of different variables between studies will allow for correlations to be seen between studies that will further illuminate the body of literature (Glass et al., 1981).

Inconsistent study quality.

Another criticism contends that the results of meta-analysis cannot be interpreted since both well-designed and poorly-designed studies are included together (Eysenck, 1978). This is sometimes referred to as the "garbage in, garbage out" criticism (Littell et al., 2008). In a field such as education where thousands of studies are published every year and many of them may be of low quality, this problem certainly looms large. This criticism states that when many studies are combined together, their varying degrees of quality will diminish the validity of the results obtained. Moreover, when many poorly designed or implemented studies are included in one meta-analysis the results are highly questionable.

However, this problem can also be solved by a coding mechanism. Meta-analysts pre-determine a rubric for judging the quality of studies and then assign studies a weight based on this rubric. In particular, the potential mediating effects of substantive and methodological characteristics of studies should be included. Weighting studies based on their quality is how meta-analysts answer this criticism (Glass et al., 1981). However, some researchers take issue with the quality of the coding performed. Feinstein (1995) dismisses such coding as it generally only give credit for whether or not researchers of a particular study explain their method. Feinstein believes that this obscures the fact that it is quite possible that even if a method is explained, it may or may not have been conducted properly. So, for example, a study to be included in a meta-analysis would receive credit on a scoring rubric for

explaining how a procedure was conducted, even if the procedure was bizarre or blatantly inappropriate for the study.

Biased towards positive effects.

Another criticism of meta-analysis is that published research findings are often biased in favor of those studies finding significant results and that this bias is transferred into the meta-analysis (Davies, 2000). Since most meta-analyses are based solely on published studies and these studies tend to report primarily statistically significant findings, then the meta-analyses based on these studies will suffer from a higher Type I error rate (Kraemer & Andrews, 1982; Littell et al., 2008).

This is true even in studies where non-significant findings are reported (Littell et al., 2008, Chan et al., 2004; Williamson & Gamble, 2005; Williamson et al., 2006). This translates into a greater tendency of meta-analyses to tabulate effect sizes that fail to confirm the null hypothesis (Littell et al., 2008). Thus, since meta-analysis takes into account a broad number of published studies, this line of criticism contends that meta-analyses will necessarily lead to a higher rate of Type I errors (Glass et al., 1981).

This criticism has two possible solutions. First, meta-analysts can conduct careful reviews of unpublished papers by using particular Internet search engines to examine theses and dissertations, papers presented at conferences, and other sources of unpublished work. This more careful review will round out the literature search and attenuate the Type I error problem. Second, there is a statistical technique whereby one can estimate the number of non-significant studies that would have to be conducted in order to overturn the results of the meta-analysis. Using this technique, the researcher conducts the meta-analysis with the understanding that he or she may be committing a Type I error and then estimates how many

studies would have to be published in order to nullify the findings of the meta-analysis. If this number is impractically large, then the meta-analysis can be judged to be sound (Glass et al., 1981). Carson, Schriesheim, and Kinicki (1990) refer to this as the fail safe N.

Methodological criticisms.

Another criticism of meta-analysis deals with the quality of the analysis itself. Jadad et al. (2000) found that in a review of 50 meta-analyses and systematic reviews that most published reviews in peer-reviewed journals had methodological flaws that impaired their utility. In a complicated data analysis technique, such as meta-analysis, errors are prone to develop unless researchers carefully and methodically move through the data analysis process. It is important to have independent raters assess the scales used to determine eligibility criteria for studies to be included in a meta-analysis and to use these scales in a rigorous manner (Glass, McGaw, & Smith, 1981).

Andrews, Guitar, and Howie (1980) is a much - criticized study dealing with stuttering therapy. These researchers set their inclusion requirement for including studies into the analysis to only those studies with three or more participants, thus overlooking single subject designs, thus invalidating their results (Cordes, 1998; Ingham, 1984). Ingham and Bothe (2002) go on to further describe how selection of studies may invalidate results, such as in the case of Thomas and Howell's (2001) meta-analysis of stuttering therapy techniques. Ingham and Bothe conclude that not only was that work not a meta-analysis, but also it was severely compromised by decisions regarding which studies to include. These examples demonstrate the need for those engaged in meta-analysis to carefully select studies for inclusion so as to maximize the validity and power of their results.

Another methodological criticism is that meta-analyses will utilize multiple iterations of one study by using different trials published in one study several times or by using subgroups from a single study that are presented in another paper (Davies, 2000). This can muddy the meaning behind the results as that one study is then given undue precedence. However, this can be dealt with statistically through a coding procedure that acknowledges this fact (Glass et al., 1981).

Another methodological criticism is that many published meta-analyses are conducted using outdated techniques and do not satisfactorily account for known sources of bias (Littell et al., 2008). One source of bias relates to the failure of primary studies to account for intervention attrition where study participants discontinue participation in a study. This issue is rarely reported in primary studies and is even less likely to be accounted for in a meta-analysis (Davies, 2000).

The quality of statistical reporting in primary studies is highly variable and difficult to account for in a meta-analysis since primary researchers often fail to report each particular step in their research process (Wolf, 1986; Davies, 2000). Cook et al. (1992) discuss that there are several methods of dealing with issues of inadequate statistical reporting by meta-analysts including: use of external sources to establish validity and reliability of instruments used in primary studies, contacting the primary investigator(s) for clarification, and reporting deficiencies of the data from primary sources. If original studies that are to be included in a meta-analysis were not based on random assignment of study participants to experimental and control groups then causal inferences cannot be made by the meta-analyst (Cooper & Hedges, 1994). This means that conclusions drawn from secondary research may be almost as limited as those made by primary research.

Another criticism of meta-analysis is that it relies overly much on effect sizes as its central measure. Effect size is the measure of the magnitude of the relationship between two variables. This is a term that is understood by researchers but is often misunderstood by practitioners. Thus, the results of meta-analysis may be misinterpreted and misapplied due to inadequate understanding of the measure. This issue, however, is easily overcome by translating the effect size measure into a metric that is more meaningful to practitioners (Littell et al., 2008). Beyond this, however, is the broader criticism that meta-analyses tend to be reported in a staggering array of scales and units with little thought to standardization. This lack of standardization makes interpreting a meta-analysis open to significant bias as practitioners or researchers can be confused or misled by the various statistical measures employed (Feinstein, 1995).

Littell et al. (2008) makes another criticism of meta-analysis in that some newer techniques of meta-analysis have yet to be validated. This lack of validation means that some meta-analytical studies are fundamentally irrelevant as their techniques are improper.

One issue in conducting meta-analysis deals with scales used to measure various variables. While some studies utilize the same scale and can be readily combined in one meta-analysis, many times a meta-analysis will incorporate studies utilizing multiple scales (Glass, McGaw, & Smith, 1981). In order to combine studies using various scales the standardized mean difference is used which is a tabulation of the difference in means between the treatment groups divided by the pooled standard deviation of the measurements. This transformation ensures that all results are measured on the same scale, thus minimizing the so-called “apples and oranges” criticism of meta-analysis. This violation of the fundamental principle of comparing homogenous units is one of the most serious criticisms

of meta-analysis. Statistically, "6-month-old children, small dogs, large cats, and huge fish can be regarded as a homogeneous group" (Feinstein, 1995, p. 76) even though inclusion of all of these into one aggregated analysis be nothing less than absurd. While meta-analysts claim that pooling large amounts of data from heterogeneous groups improves generalizability, critics maintain that doing so produces "imprecision, confusion, and perhaps delusion (Feinstein, 1995, p. 76).

Gotzsche et al. (2007) caution that the tabulation of standard mean differences entails many difficult calculations and is fraught with potential errors. In their analysis of 27 medical meta-analyses using this technique, they found errors in 10 of them (37%). Thus, it is imperative that meta-analyses that utilize this technique do so with extreme caution in order to avoid arriving at spurious conclusions.

External validity.

A key criticism of meta-analysis for education is that meta-analyses tend to have low external validity (Littell et al., 2008; Davies, 2000). Littell et al. (2008) make this criticism as applied to a social work context and it is safe to say that this criticism holds as well for education. While most studies included in a meta-analysis are conducted in tightly controlled university settings, practitioners in the field will use the results of these studies and the university and field settings may be markedly different. In particular, educational research has been criticized for its atomized nature that makes policy prescriptions difficult to make based on research (Davies, 2000). Slavin (1986) discusses the need for meta-analysis to study variables that are applicable in the field. Conducting meta-analyses on variables that are only of interest to academics is of much less use. Since meta-analyses group many studies together, this validity problem is amplified. A key issue in translating meta-analyses to

education is that effect sizes or standard deviations do not readily translate into the scales used by educators. External validity could be improved if more studies conducted in field settings using scales familiar to educators were conducted and included in a meta-analysis.

Temporally limited utility.

Another criticism of meta-analysis stems from the fact that meta-analyses are conducted at one particular point in time and thus reflects only the current state of the literature. When new studies on a particular topic are released, then the meta-analysis on those studies is immediately less valid. Thus, there is a need to constantly update meta-analytical research to keep it valid (Littell et al., 2008).

Quantitative focus.

Some researchers also take fault with meta-analysis's focus on only quantitative data. While it is strictly true that meta-analysis cannot incorporate non-numeric data, qualitative data can be used to inform the construction of a meta-analysis and to provide contextual information that may be used to help to interpret the results of the meta-analysis (Littell et al., 2008). However, Davies (2000) indicates that synthesizing high-quality qualitative has garnered increasing attention from researchers in education and other social sciences using tools such as meta-ethnography. The use of qualitative studies can form the basis of contextualizing the results of a meta-analysis and may allow more educational users to access meta-analytic findings (Davies, 2000).

Since one of the key reasons for researchers moving to the more rigorous, quantitative literature review as opposed to the less rigorous, qualitative literature review is to avoid bias, conducing meta-analyses carefully and methodically is critical.

Academic or philosophical criticisms.

Another criticism of meta-analysis relates to the academic value of the technique. Many academics deride secondary analyses as derivative or parasitic. It has often been deemed as unworthy of publication and not listed on researchers' curriculum vitae. As the technique became increasingly established and respected throughout the 1990s, this criticism has been much diminished (Chalmers, Hedges, & Cooper, 2002). Glass (1976) defends the practice thusly:

> The man who adds his bit of fact to the total of knowledge has a useful and necessary function. But who would deny that a role by far the greater is played by the original thinker and critic who discerns the broader outlines of the plan, who synthesizes from existing knowledge through detection of the false and illumination of the true relationships of things a theory, a conceptual model, or a hypothesis capable of test. (p. 417)

Despite this eloquent defense, the practice of meta-analysis, and research synthesis more broadly, remains somewhat of a lesser-respected practice. Feinstein (1995) contends, "meta-analytic results of observational research are often acts of politics, not science" (p. 76).

Finally, the most overarching criticism of meta-analysis is that it provides a semblance of objectivity when in fact this is impossible. When meta-analysts use such rigorous techniques in their attempt to account for every variable, it necessarily leads those reading a meta-analysis to believe that the meta-analysis is unquestionably the correct interpretation of the phenomenon being studied (Wolf, 1986). Feinstein (1995) emphasizes the idea that meta-analysis fails to meet the requirements for scientific inquiry as it mixes too many facts together in a manner that is fundamentally flawed. This criticism can only be answered by the continued use of the scientific method to analyze and interpret complex social phenomena.

2.5 Alternatives to Meta-Analysis

The two alternatives to meta-analysis for synthesizing literature are narrative summaries and vote counting, both of which have been described as "haphazard" and "opportunistic" (Petticrew and Roberts, 2006; Davies, 2000) but remain the standard for research synthesis in most fields. A narrative summary, otherwise known as the traditional literature review, describes primary studies and attempts to come to conclusions about which direction the evidence seems to indicate about the research question at hand. This is done through a mostly qualitative, value-laden process that is generally invisible to those reading the literature review. A more quantitative approach to literature synthesis is vote counting, first described by Light and Smith (1971). This is a simple procedure where researchers tally whether studies indicated statistical significance or non-significance in order to impose some degree of quantitative rigor onto the results. The weaknesses of both methods will be discussed below.

Traditional literature review.

The traditional literature review has several disadvantages that a meta-analytic approach can ameliorate. Traditional literature reviews, according to Pillemer (1984) have been characterized as

> subjective, relying on idiosyncratic judgments about such key issues as which studies to include and how to draw overall conclusions. Studies are considered one at a time, with strengths and weaknesses selectively identified and casually discussed. Since the process is informal, it is not surprising that different reviewers often draw very different conclusions from the same set of studies. (p. 28)

Narrative reviews often make no attempt to generalize findings, but rather identify elements of the various studies in a body of literature in order to base future research. They are generally selective as they do not involve a systematic, rigorous, and exhaustive search of all

of the literature (Davies, 2000). These weaknesses are so abundant in the social sciences, that one team of researchers has commented "If a review purports to be an authoritative summary of what 'the evidence' says, then the reader is entitled to demand that this is a comprehensive, objective, and reliable overview, and not a partial review of a convenience sample of the author's favorite studies" (Petticrew & Roberts, 2006, p. 6). Carefully conducted systematic reviews utilizing meta-analyses offer a transparency not present in traditional, narrative summaries of research findings that supports a scientific method of collecting and reporting results (Littell et al., 2008). These weaknesses of the traditional literature review can be seriously detrimental to the quality of the conclusions drawn from the literature and meta-analysis has the potential to minimize those weaknesses (Littell et al., 2008). There is frequently no attempt to standardize techniques between researchers and the process is often left to be a matter of private judgment, style, and creativity. While it is by no means advisable to enforce a set of rules for research synthesis and stymie the creative forces that generate primary research, it is desirable to promote the greatest degree of clarity, explicitness, and openness possible so that the scientific method can be fruitfully engaged in the social sciences (Glass, McGaw, and Smith, 1981).

An example that illuminates the need for more quantitatively based research syntheses comes from Cooper and Rosenthal (1980). These researchers had a group of individuals who were at least graduate students or better assess a group of seven studies relating to sex differences and persistence to a task. They divided up the group into a treatment group that received basic training in quantitative research synthesis and another that did not receive such training. The group that did not receive the statistical training incorrectly came to the conclusion that there was no relationship between sex and persistence

nearly 75% of the time while only 31% of the treatment group made the same mistake. It should be pointed out that this was an attempt to integrate only seven studies. This is compounded by the fact that most literature bases are so complex and large that a narrative literature review that accurately synthesizes treatment effects is beyond the mental capacity of most researchers (Bushman & Wells, 2001). Narrative reviews tend to be opportunistic as researchers review only the studies that are readily available (referred to as the file drawer problem) (Rosenthal, 1979; Wolf, 1986; Davies, 2000).

Traditional literature reviews have been criticized on the following points: bias related to previously held ideas; bias towards positive effects; inadequate explanation of differences across studies; inadequate attention paid to study quality; allegiance effects. Each of these is discussed more fully below.

Previously held ideas.

Previously held ideas tend to distort individuals' data processing through a variety of means such as: behavioral confirmation, biased attribution and recall, and biased assimilation (Anderson & Lindsay, 1998). This is due to the manner in which humans utilize heuristics to determine what knowledge is valuable. Tversky & Kahneman (1973) discuss the availability heuristic which is a frequently utilized mechanism that leads individuals to judge the prevalence of an event based not on the actual frequencies of the event but also by other variables relating to their memory such as vividness, recency, and familiarity. This is not related to preconceptions but is rather rooted in properties of the events being studied (Bushman & Wells, 2001). This can be a significant source of bias in traditional literature reviews.

Additionally, the bulk of traditional literature reviews do not take into account differences across studies as explanations of variance in research findings.

Inadequate exploration of issues relating to study design.

Most traditional literature reviews take differences in the findings between studies rather than differences in study design as their focus. However, much of the variance between studies may have more to do with study design than with actual variance between populations being studied. The traditional literature review does not examine the differences between study designs and, hence, loses the capacity to truly illuminate the true state of the literature. While the literature may seem to indicate one particular finding, this finding may be more due to differences in study design rather than differences in populations of interest (Wolf, 1986).

Bushman and Wells (2001) hold that conclusions regarding statistical significance and effect size produced via traditional literature reviews should be less trusted than those reached via meta-analysis as traditional literature reviews can be readily biased by variables not related to the research findings. Modern research syntheses in education should include quantitative attempts to summarize the current state of the field (Glass, McGaw, and Smith, 1981). Without these quantitative measures, it becomes very difficult for a researcher to reach reasonable conclusions regarding the state of a body of literature.

Allegiance effect.

Another criticism of the traditional literature review is the allegiance effect where researchers who have a vested interest in a particular finding bias the results (Luborsky et al., 1999).

Vote counting.

Vote counting is a technique to summarize the literature by simply counting the number of studies that provide answers to the research question at hand using those studies' tests of statistical significance with no regard to effect sizes (Littell et al., 2008). The category that has the highest count is taken to be the modal finding and is generally believed to be the most effective (Davies, 2000). It is a more sophisticated form of research synthesis though it is weak due to the fact that it tends to conflate study significance and study quality and ignores the importance of effect sizes, study quality, sample size, and moderating effects (Carlton & Strawderman, 1996; Davies, 2000; Glass, McGaw, & Smith, 1981; Light & Smith, 1971). As such, vote counting is rarely utilized in modern academic research.

2.6 Effect Size

Effect size is a concept developed from Cohen's (1988) work on power analysis. The originators of meta-analysis took Cohen's work and created a new statistical measure, called effect size, that could be used to describe the standardized difference in population means (Carlton & Strawderman, 1996).

Effect size is a far superior method of describing relationships between variables. "[T]here are many sound reasons for completely abandoning such reliance (on statistical significance) in favor of direct estimates of effect-size" (Carlton & Strawderman, 1996, p. 72)

Diminishing effect size.

Ecologists have discovered several examples of diminishing effect sizes (Alatalo et al., 1997; Gontard-Danek & Moller, 1999; Simmons et al., 1999; Poulin, 2000). As of yet, a generally agreed-upon interpretation of why effect sizes apparently diminish over time has not emerged. The following are possible explanations. Alatalo et al. (1997) attribute

diminishing effect sizes to changing belief systems. Palmer (2000) attributes the phenomenon to fads. Tregenza and Wedell (1997) attribute it to biased study design. Alatalo et al. (1997) have suggested that submitting findings for publication that support previously held ideas makes it easier to get published. Simmons et al. (1999) suggest that it is easier to publish confirmatory findings during early stages of research in a particular field but it becomes more difficult as a more narrowly critique of that field develops. This may be particularly emphasized in the social sciences where it takes longer to publish non-significant results (Stern & Simes, 1997).

Potential explanations for phenomenon.

Researchers who study the phenomenon of diminishing effect sizes cite two primary potential causes: dissemination bias and citation bias. This section discusses these potential explanatory factors in greater detail.

Dissemination bias.

Dissemination bias is a broad term encompassing all elements of the dissemination process of a research report which includes bias related to date of publication, language, multiple publication bias, selective reference citation, database index bias, media attributed bias, selective publication bias, familiarity of techniques, and the cost of research reports (Song et al., 2000; Rothstein, Sutton & Bornstein, 2005). This term is an overarching term that takes in many different sorts of biases that are related to the publication and dissemination process (Song et al., 2000). "Dissemination bias occurs when the dissemination profile of a study's results depends on the direction or strength of its findings" (Song et al., 2000, p. 17). Dissemination bias refers to the notion that a given literature review does not represent a random sampling of all studies in a given field. It is a type of

non-random sampling error similar to that found when conducting primary research (Song et al., 2000).

Both indirect and direct evidence support the existence of dissemination bias (Sohn, 1996). Examples of indirect evidence include disproportionately high percentage of positive findings in journal or larger effect sizes in small studies relative to large studies. Small studies are more vulnerable to dissemination biases as the results of these studies will be more widely spread around the true results owing to greater random error (Begg & Berlin, 1988). Direct evidence includes such things as admissions by investigators and publishers and comparison of results from published and unpublished studies (Song et al., 2000). Rotton et al. (1995) found that the most significant reason given by authors for not submitting their work for publication was the failure to find statistical significance.

The strongest evidence supporting the existence of dissemination bias comes from comparisons between published and unpublished studies (Song et al., 2000). Simes (1986) performed meta-analyses on both published and unpublished studies of a cancer treatment regimen and discovered that the published findings found that the treatment was effective but when the published and unpublished studies were analyzed together, the treatment effect was not found.

There are many specific types of dissemination bias: positive results bias, hot stuff bias, time-lag bias, grey literature bias, full publication bias, place of publication bias, outcome reporting bias, multiple publication bias, language bias, citation bias, database bias, retrieval bias, media attention bias. These forms of bias are prevalent in many disciplines and may account for observed decline in effect sizes.

Citation bias.

This is a related set of biases known as citation bias, reference bias, or one-sided reference bias that refers to the chance of a study being cited by others depending on the results of that study (Sackett, 1979, Song et al., 2000). The most common form of this is when authors of a published study tend to cite studies that support their position. This effect echoes into future literature reviews as researchers search works cited for guidance in formulating new research questions or informing old ones (Song et al., 2000).

2.7 Systematic Review

Meta-analysis falls under the broader category of systematic review, a term used in the medical and behavioral sciences to connote a more rigorously applied literature review technique. Systematic review is a technique by which a body of literature can be synthesized in a meaningful, rigorous manner. The purpose of a systematic review is to assist practitioners of various fields to understand often esoteric, academic peer-reviewed studies and utilize the best possible practices, often referred to as evidence-based or evidence-informed practice (Shlonsky, 2011). A meta-analysis is considered a critical piece of a systematic review (Littell, 2008). Littell (2008) outlines the steps in this process, which are as follows.

First, a detailed plan for the meta-analysis is developed where the objectives and methods of the procedure are outlined. All steps in the process are rigorously recorded so that reviewers can evaluate the process. Then, the researcher specifies the criteria that will be set to include or exclude studies from the meta-analysis. Possibilities include: study designs, populations, interventions, comparisons, and outcome measures. All reasons for including or excluding particular studies are recorded so that the greatest degree of transparency is achieved, thus avoiding one of the potential pitfalls of meta-analysis (Ingham, 2002). Then

the researchers utilize a systematic approach to attempt to find all potentially useful studies. Both published and unpublished studies are searched to the greatest degree possible. There is an attempt to locate what has been termed "grey literature" which are hard to find studies, many of which are unpublished or were presented at minor conferences. This minimizes the so-called "file drawer problem" of researchers quietly filing away research that did not find significant results (Hopewell, McDonald, Clarke, & Egger, 2006; Petticrew & Roberts, 2006; Rosenthal, 1979; Rosenthal, 1994; Rothstein et al., 2004; Rothstein et al., 2005). It is imperative to search unpublished sources (including research in progress) so that the problems relating to publication bias can be avoided (Davies, 2000). After this process is completed, all decisions should be made by two independent raters who work together to form a documented consensus on what studies to include. Data is then extracted from the reports to be used and notes are kept so that reviewers' decisions are transparent (Higgins & Green, 2005). The quality of studies are systematically reviewed, with particular emphasis placed on those elements of the methods that directly relate to the validity of a study's conclusions. Generating overall "study-quality" scores is a less-useful way to do this (Shadish & Myers, 2004). Wortman (1994) recommends using Campbell's threats-to-validity approach. Study findings are then represented as effect sizes and results are synthesized. Results are then reported. The standard reporting rubric has become Moher's Quality of Reporting of Meta-analyses (QUORUM) system that includes a checklist of items that should be reported and a diagram to aid authors in describing how studies were identified, screened, and selected (Moher et al., 1999)

The Cochrane and Campbell Collaborations.

Two organizations have developed recently to conduct and organize research syntheses. The Cochrane Collaboration focuses on organizing and disseminating research syntheses in the health sciences while the Campbell collaboration focuses on the social sciences (Chalmers, Hedges, & Cooper, 2002).

The Cochrane Collaboration has focused attention on evaluating sources and extent of bias in randomized clinical trials in the health sciences. This group has constructed an instrument called the Cochrane Collaboration Risk of Bias Tool that functions as a measure of methodological quality. This tool allows practitioners to determine if a the recommendations posited by a group of randomized clinical trials is likely to be biased. While some researchers have questioned the rigor of this tool, it serves as an indication that many fields of study are currently searching for ways in which to synthesize research (Armijo-Olivo, 2012).

The Campbell Collaboration has focused its efforts on evaluating systematic reviews in the social sciences. It is a sister organization to the Cochrane Collaboration and focuses its efforts on three areas of social science practice and policy: social welfare, crime and justice, and education. The Campbell Collaboration utilizes standards of methodological rigor and it is considered as a leader in research synthesis and meta-analysis. The editorial board of the Campbell Collaboration conducts extensive and rigorous systematic reviews and meta-analyses in the various fields and then publishes findings on its website, www.campbellcollaboration.org, at no cost (Shlonsky et al., 2011).

The existence of these two organizations points to the perceived need to accurately and reliably synthesize research. Moreover, these organizations have established the strict

and rigorous standards for research synthesis that should serve as guidelines to all researchers utilizing meta-analysis.

Chapter III

Methods

The review of the literature reveals that meta-analysis is widely used in educational research but that there are significant concerns regarding its utility. In particular, the phenomenon of diminishing effect sizes may distort empirically observed reality and lead to the implementation of poorly supported interventions. This study analyzes whether the phenomenon of diminishing effect sizes is observable via analysis of meta-analyses over time in education research.

3.1 Research Consents

No research consents were necessary for this project as no new data was gathered.

3.2 Literature Review Procedure

This study was undertaken using a process similar to that used by Jennions and Moller (2001). This process is outlined in detail in the following sections. There are no significant differences between the Jennions and Moller (2001) process and the one used in

this study. It is hoped that this process, when applied to the social science data analyzed for this project, will elucidate the research question.

This study was conducted in 6 phases: initial study selection, primary literature consideration, literature consolidation, secondary literature consideration, tertiary literature consideration, and data analysis. Each of these steps is further described below.

Initial study selection.

First, a set of meta-analyses was selected from the EBSCOHost databases. Meta-analyses will be selected from the years 1970 to 2011. This date range is seen as having the best chance of including the greatest number of meta-analyses possible. Preliminary literature searches bear out this conclusion since database searches yield no meta-analyses prior to 1970. Studies will be included if they: specifically use one or more meta-analyses based on effect sizes, provide a comprehensive list of studies used to generate effect sizes. Then, two sets of analyses will be conducted, one evaluating the studies included in selected meta-analyses and the other evaluating the meta-analyses themselves. The process for conducting this selection is outlined below.

Primary literature consideration.

Research databases were selected based on an analysis of databases available at Cleveland State University. This potential source of bias is dealt with in the Discussion section. A database was selected if it was likely it indexed journals of interest to the study. Appendix A details which research databases were utilized for the study.

Then, each database was searched to identify articles that would possibly be included in subsequent phases. Articles were assessed based on title and publication characteristics. Appendix B contains the study selection worksheet for the primary literature consideration.

Studies that met the criteria described in Appendix B were then entered into a spreadsheet for further consideration. This process generated 464 articles that continued on to secondary consideration.

Literature consolidation.

All articles identified by the primary literature search were entered into RefWorks, a citation management service. Duplicates were then eliminated. This left 387 studies that would continue to secondary consideration.

All articles were located electronically, photocopied, or requested via inter-library loan. Portable Document Files (PDFs) of all articles were generated. A small set of articles, citations for which are provided in Appendix C, were unable to be found by either extensive database searching or through the assistance of research librarians. Also, due to financial constraints of this project, dissertations were excluded from analysis.

Secondary literature consideration.

The abstracts of all 387 articles were considered. For articles to continue onto tertiary consideration, they had to meet the requirements described in Appendix B. This closer survey of literature reduced the total number of articles to 112.

Tertiary literature consideration.

The 112 articles that remained after primary and secondary consideration were further narrowed down using study selection criteria outlined in Appendix D. This final winnowing entailed reading each article and determining whether that article provided certain key data elements. In particular, a study had to include sample sizes, publication years, and effect sizes for all studies included in the meta-analysis. Furthermore, studies had to include the year of publication, number of studies included in the meta-analysis, at least one reported

effect size, and the standard deviation for that effect size at the meta-analysis level. Studies that were selected through this process were coded into a spreadsheet using coding forms found in Appendices F and G.

Studies that did not meet these requirements were excluded, as they did not provide an adequate amount of data to further analyze. This resulted in a final set of 23 studies that were analyzed. Descriptive statistics of these studies are shown in Table 1.

Table 1

Descriptive Statistics of Included Studies

N (Meta-analyses)	N (Effect Sizes)	Year of Publication Range	Mean Year of Publication	Mean Number of Reported Effect Sizes Per Meta-Analysis
23	60	1984 – 2010	2002.3	42.7

3.3 Statistical Procedure

The final group of studies that were included in this study was then analyzed using a process outlined by Jennions and Moller (2001). This process involves the use of Spearman's ρ (rho) analyses on four sets of data on two levels.

Analytical levels.

This study was undertaken on two analytical levels. The first set of analyses deals with the effect sizes reported in the studies comprising the meta-analyses identified for inclusion in the study. This will hereafter be known as the "study level" of analysis. The second set of analyses were conducted on the meta-analyses themselves. This is hereafter known as the "meta-analysis level."

Relationships of interest.

On both the study level and the meta-analysis level, four relationships were analyzed: (i) the relationship between effect size and year of publication; (ii) the relationship between effect size and sample size; (iii) the relationship between standardized effect size and sample size; and (iv) the relationship between effect size and year of publication, after weighting for variation in sampling effort. The first three relationships were conducted using a Spearman's ρ (rho) test and were performed in SPSS.

The fourth relationship, between effect size and year of publication, after weighting for variation in sampling effort, was conducted using MetaWin 2.0. This relationship was estimated by creating a random-effects continuous model meta-analysis with year of publication as the independent variable and the inverse of sampling variance as the weighting factor. Random-effects meta-analysis was selected over a fixed – effects model as fixed – effects models become problematic when some studies have very large sample sizes. These studies then dominate the analysis and the results from the studies with smaller sample sizes then are largely ignored (Helfenstein, 2002).

MetaWin 2.0 was used to obtain a one-tailed ρ – value for year of publication generated by a randomization method with 999 replicates. A one-tailed ρ – value was chosen because the Jennions and Moller (2001) study used a one-tailed test since they postulated that a declining effect size was more likely. This study employed a one – tailed test because it was initially believed that similar research and publication biases would be in effect in both the natural and social sciences.

The effect size generated by the meta-analysis was converted to a Spearman's ρ (rho) value so that all results are reported in a uniform manner. The formula to do this is as follows:

$$\rho = \sqrt{\frac{d^2}{d^2 + 4}}$$

All Spearman's ρ – values were then converted to standard normal deviates (Z-scores) by using the formula:

$$\rho = \frac{\sqrt{Z^2}}{n}.$$

This is done so that all results are normalized thus diminishing the effects of outliers and to allow results to be evaluated using equivalent metrics.

Chapter IV

Results

This study attempted to address the following research question: Do effect sizes in published meta-analyses of educational research tend to diminish over time? The remainder of this chapter explains the results discovered through the process outlined above. Results of these analyses are in Table 2.

Table 2

Relationships (ρ) between effect size, standardized effect size, year of publication, and sample size

Method of calculation	Year versus effect Size	*n* versus effect size	*n* versus standard effect	Year versus effect after weighing for sampling variance
Weighted meta-analysis of datasets	0.105*	-0.073**	-0.073**	0.440*
Weighted meta-analysis of original meta-analyses	0.317**	-0.148	-0.148	0.333*

* Significant at the <0.001 level
**Significant at the <0.01 level

Table 3 presents the summary results from the meta-analysis used to weight for sample size when determining the relationship between publication year and effect size. Table 4 provides the descriptive statistics of the meta-analytic model. The effect size (g) generated by the meta – analysis was 0.4756. This effect size was then converted to a Spearman's ρ .

Note that Cochrane's Q is the weighted sum of squared differences between individual study effects and the pooled effect across studies. Cochrane's Q has been criticized for having low power when the number of studies is small (Gavaghan, Moore, McQay, 2000). Thus, Cochrane's Q has been converted to I^2. I^2 describes the percentage of variation across studies due to heterogeneity rather than chance (Higgins & Thompson, 2002). The formula for calculating I^2 is:

$$I^2 \equiv 100\% \times \frac{Q - df}{Q}$$

Table 3

Meta – analysis summary

Predictor	Value	Standard error	Probability (normalized)	Probability (randomized)
Intercept	-43.6598	16.6122	0.00858	0.046
Slope	0.0220	0.0083	0.00789	0.047

Table 4

Meta – analysis model

Model	df	Q	Probability (Chi-square)	I^2
Regression	1	7.0587	0.00789	
Residual	57	65.4534	0.20684	12.91%

The I^2 results indicate that the meta-analysis was well conceptualized and conducted, with 12.91% of the variation due to heterogeneity. Higgins, Thompson, Deeks, and Altman (2003) suggest that this is a low level of heterogeneity, thus favoring the conclusion that the meta-analysis conducted was conducted sufficiently to answer the research question.

Beginning at the study level, these results indicate that there is a statistically significant positive relationship between year of publication and effect size ($\rho = 0.105$, $p < 0.001$, $n = 1167$). That is, the more recently the study was conducted, the greater the effect size tends to be. However, there was also a significant relationship between sample size and both effect size and standardized effect size so the relationship was re-assessed after

accounting for sampling variance. Still, however, a statistically significant, positive relationship was observed (ρ = 0.440, $p < 0.001$, $n = 1167$).

Figure 1 shows a scatterplot of the relationship between publication year and effect size at the study level. Note that effect sizes are reported as Hedge's – *g*, a commonly used effect size measure. This figure shows a prominent cluster of effect sizes between the years 1980 to 2005 centered around -0.5 to 1.75.

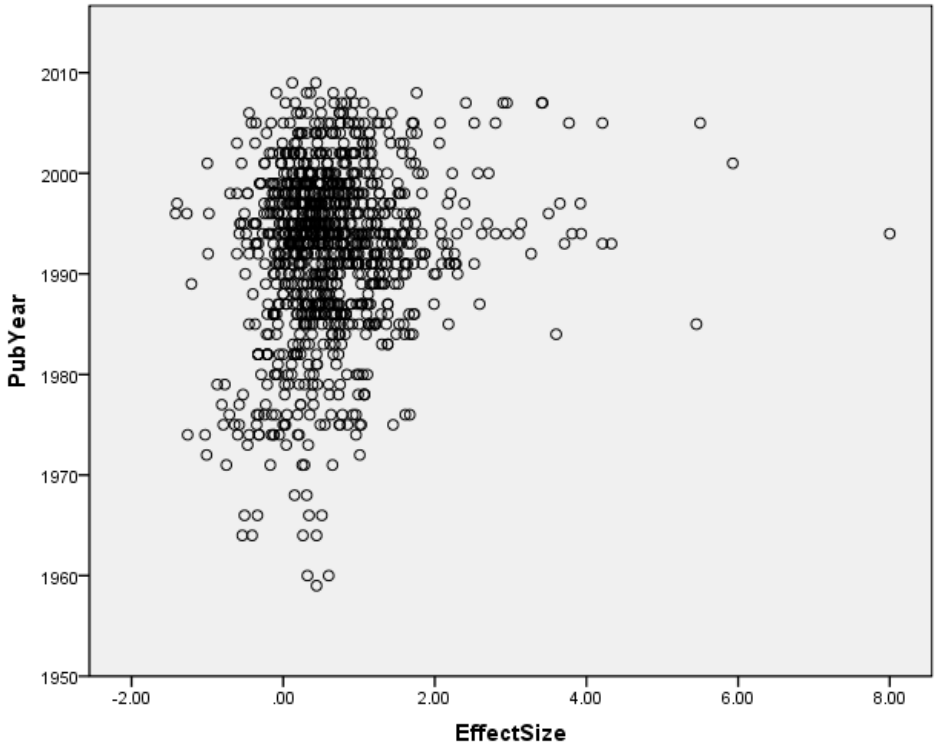

Figure 1: Publication year compared to effect size (g) at the study level

Figure 2 shows a scatterplot of the relationship between sample size and effect size at the study level. This figure shows that most sample sizes were smaller than 200 and effect sizes were centered around -0.5 to 1.75.

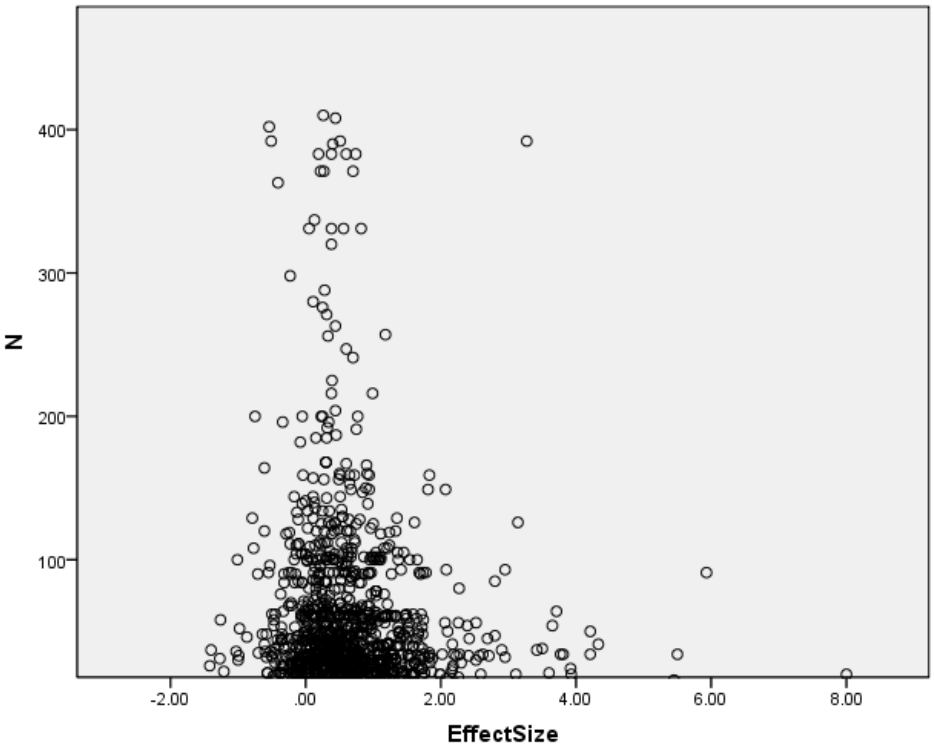

Figure 2: Sample size compared to effect size (g) at the study level

Figure 3 shows a scatterplot of the relationship between sample size and standardized effect sizes at the study level. Effect sizes were standardized using a *Z* transformation process. This shows that most effect sizes center around 0 and few studies have sample sizes larger than 200.

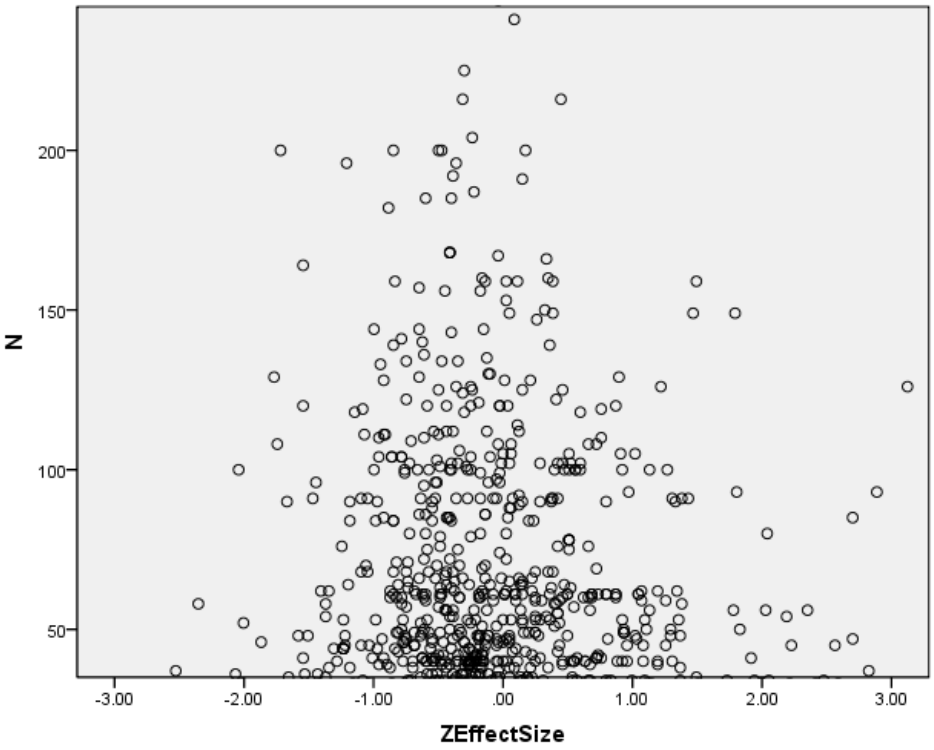

Figure 3: Sample size compared to standardized effect size (z transformed) at the study level

Figure 4 shows a scatterplot of the relationship between publication year and effect size at the study level after accounting for sample size using a random-effects continuous model meta-analysis with year of publication as the independent variable and the inverse of sampling variance as the weighting factor. This figure shows that most effect sizes fall between 0.2 and 1.8 and the year of publication tends to be 1985 and 1997.

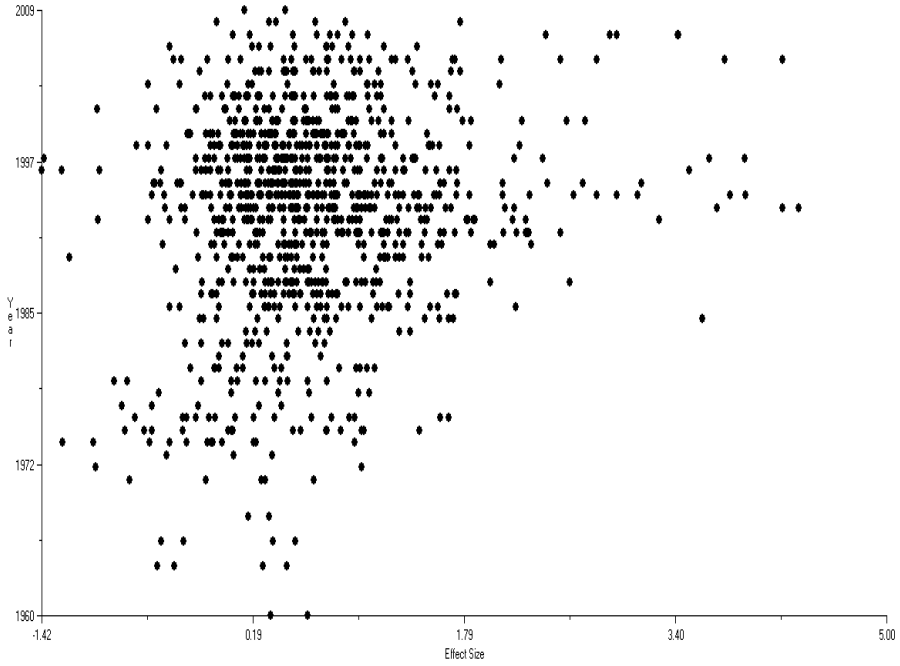

Figure 4: Year of publication compared to effect size (g) after weighting for sample size at the study level

A similar observation is found at the meta-analysis level. These results indicate that there is a statistically significant positive relationship between year of publication and effect size (ρ = 0.317, $p < 0.009$, $n = 60$). However, there was not a significant relationship between sample size and both effect size and standardized effect size. Still, however, a statistically significant, positive relationship was observed (ρ = 0.333, $p < 0.001$, $n = 60$) after accounting for sampling variance.

Figure 6 shows a scatterplot of the relationship between publication year and effect size at the meta - analysis level.

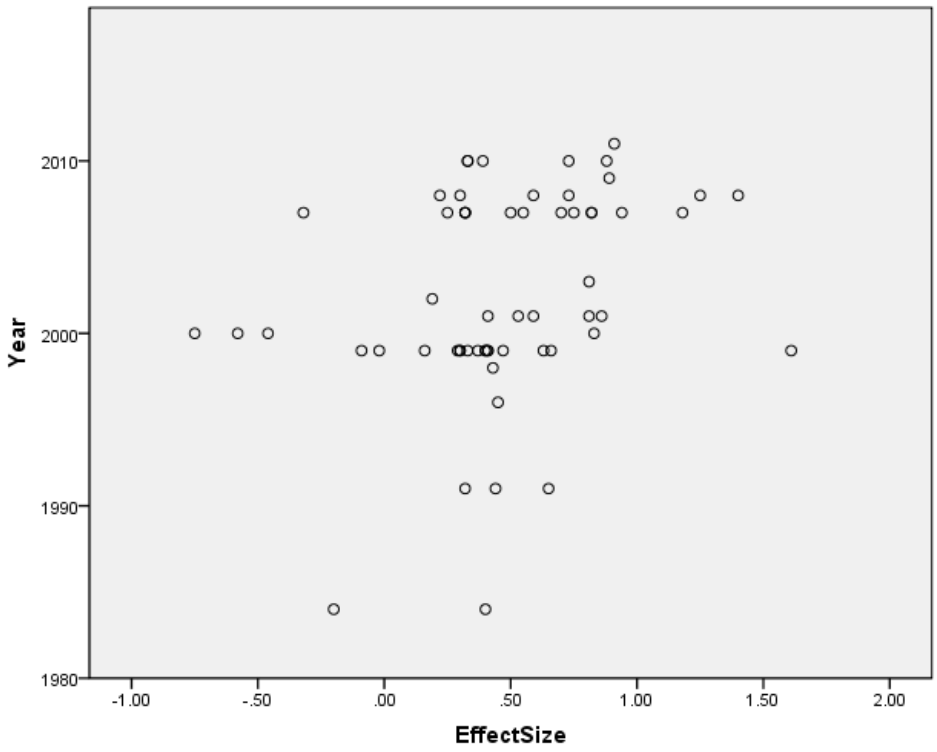

Figure 5: Year of publication compared to effect size (g) at the meta-analysis level

Figure 6 shows a scatterplot of the relationship between sample size and effect size at the meta - analysis level. This figure shows that most effect sizes cluster around 0.5 and most sample sizes were below 40.

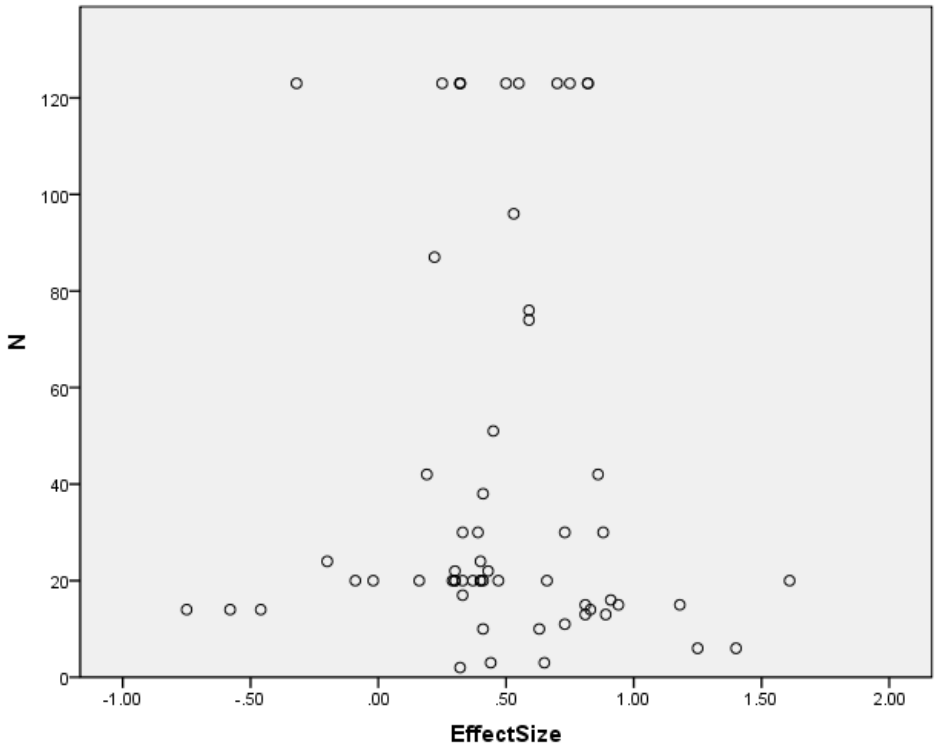

Figure 6: Sample size compared to effect size (g) at the meta-analysis level

Figure 7 shows a scatterplot of the relationship between sample size and standardized effect sizes at the meta-analysis level. Effect sizes were standardized using a *Z* transformation process. This figure shows that effect sizes are clustered around 0 and sample sizes are generally below 30.

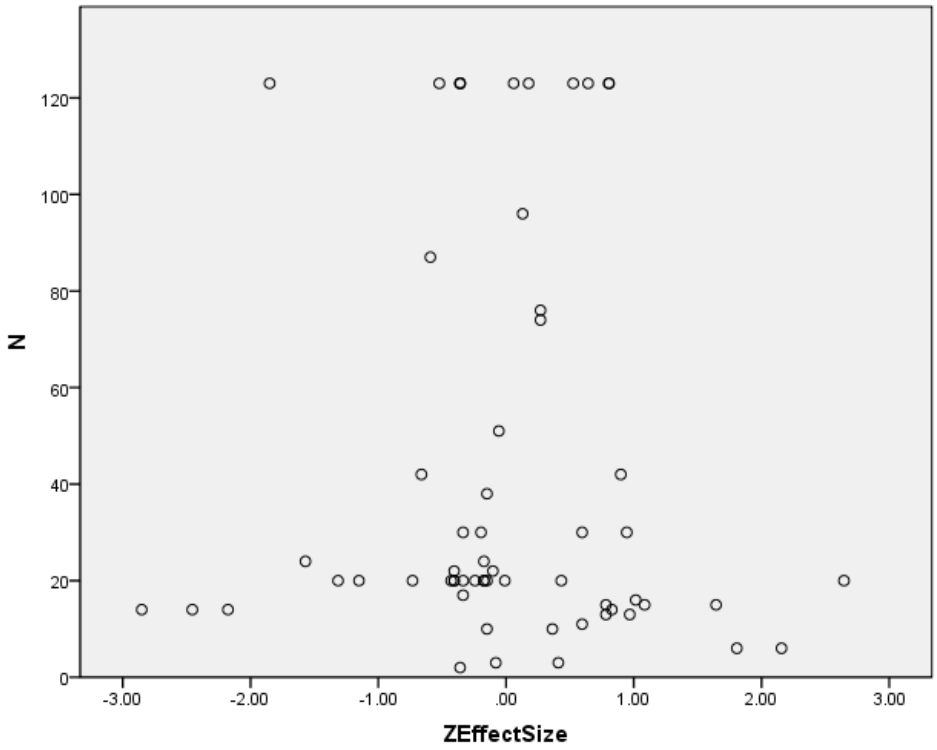

Figure 7: Sample size compared to standardized effect size (*Z* - transformed) at the meta-analysis level

Figure 8 shows a scatterplot of the relationship between publication year and effect size at the meta - analysis level after accounting for sample size using a random-effects continuous model meta-analysis with year of publication as the independent variable and the inverse of sampling variance as the weighting factor. This figure shows how most effect sizes were between 0 and 1.3 and most years of publication tended to be between 1991 and 2005.

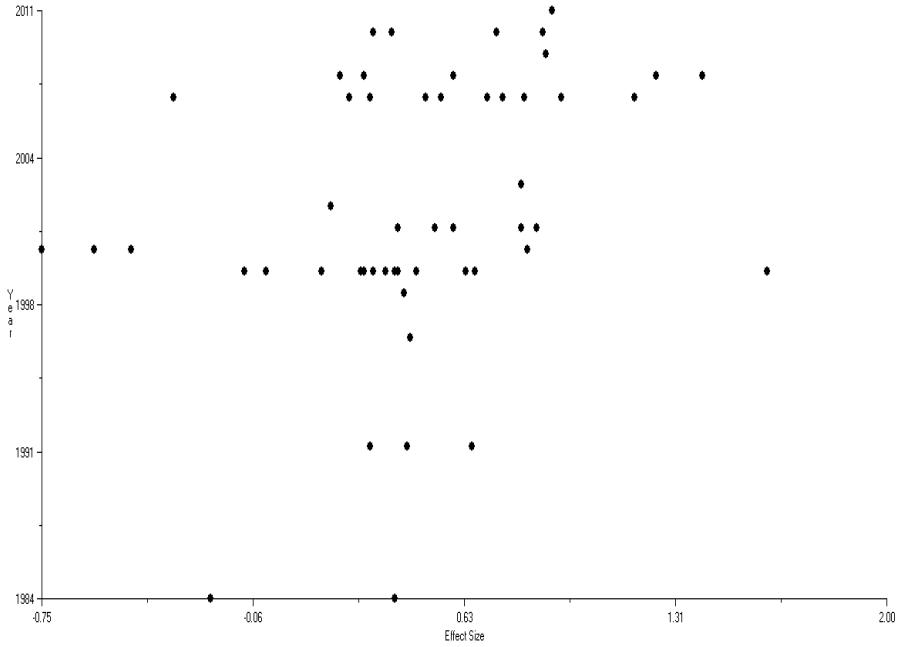

Figure 8: Year of publication compared to effect size after weighting for sample size at the meta-analysis level

It is notable that effect sizes increase at both the study and meta-analysis levels. Data were parsed out to show mean effect sizes by decade to allow for simpler understanding of how effect sizes have increased over time. Table 3 shows this descriptive information

Table 5

Mean effect sizes by decade

	N	Mean effect size (g)	Range	Standard Deviation
1970s and 1980s	2	0.100	-0.20 – 0.4	0.424
1990s	21	0.424	-0.09 – 1.61	0.329
2000s	31	0.509	-0.75 – 1.40	0.506
2010s	6	0.595	0.33 – 0.91	0.276

Chapter V

Discussion

This study has found that education meta-analyses do not follow the pattern seen in the natural sciences. Effect sizes did not decline. Rather, effect sizes tended to increase over time. This finding bears some consideration. If no statistically significant relationships had been observed between effect sizes and year of publication, then it could be assumed that meta-analysis provides a longitudinally stable measure and a strong argument could have been made for wider use of this analytical technique. However, as measured effect sizes tend to increase over the time period 1970 – 2012, one must conclude one of two things. Either there is some persistent set of biases that are impacting the conduct or publication of educational research or effect sizes are, in fact, increasing over time as the field of education develops into a more complex and sophisticated science and leaves behind educational practices that were ineffective. These two explanations will be the central thrust of this chapter and the chapter will conclude with recommendations for addressing the phenomenon

of increasing effect sizes and limitations of the present study that should temper interpretation of its results.

5.1 Persistent Bias in Educational Research

One explanation for the observed phenomenon of longitudinally increasing effect sizes is publication bias. Chapter two of this study discussed several varieties of publication bias that may impact results from any study that arise through its publication or non-publication. Given the findings of this study, it seems reasonable to conclude that it is possible that some of these forms of bias may be more active than others. In particular, the following forms of publication bias are possible explanations for the findings of this study: positive results bias; hot stuff bias; grey literature bias; and confirmation bias.

Positive results bias.

Positive results bias refers to the tendency of authors to submit and for editors to publish positive or significant research results while ignoring non-significant results (Song et al., 2000). This seems to be a very likely cause of increasing effect sizes. Since researchers generally will find statistically significant results when they are searching for literature to use to conduct meta-analyses, they will find ever - increasing effect sizes across time. Then, as other researchers use published meta-analyses to generate effect sizes for other research, this effect becomes multiplied as researchers duplicate biases from past research.

Hot stuff bias.

Another form of bias that could account for the phenomenon of increasing effect sizes is hot stuff bias. This refers to the phenomenon of journal publishers tending to publish topics that are timely or popular but which may only have relatively weak results (Sackett, 1979). This seems to be a particularly likely form of publication bias in education where fads and

trends dominate pedagogical practice. Often these trends are pushed by textbook publishers looking to profit from a product or politicians who make educational policy with little understanding of educational systems and processes.

Hot stuff bias may account for increasing effect sizes through publishers choosing articles to publish based on what they believe will promote their journal's readership. Publishers would choose articles that may be methodologically unsound to publish and then these articles are indexed in electronic indexes and used to conduct meta-analyses, thereby creating the appearance of increasing effect sizes over time. When that particular trend ends, no researcher bothers to fully repudiate it or no journal chooses to publish these repudiations so it appears that these effect sizes are significant and increasing over time.

Grey literature bias.

Grey literature refers to things such as conference presentations, dissertations, working papers, and other pieces of literature that are difficult to obtain as they are not electronically indexed in any systematic manner (Auger, 1998). Grey literature bias refers to the notion that these pieces of literature tend to show non-significant or statistically weaker results and that excluding these from meta-analyses produces an artificially high effect size (Song et al., 2000). McAuley et al. (1999) sampled 135 meta-analyses, 38 of which included grey literature, found that those meta-analyses that included grey literature showed a diminished effect size of approximately 12%.

Grey literature bias would appear to be a significant problem in the field of educational research where many universities have large numbers of master's and doctoral students who are producing volumes of research that is never published. While it is difficult to quantify specifically how much research is conducted and never included in any sort of

meta-analysis, it is safe to assume it must be a large amount every year. When one includes classroom research done by practicing teachers, the amount of grey literature skyrockets. While not all of this research would meet methodological criteria for publication or for inclusion in properly conducted meta-analyses, some certainly would. The exclusion of this grey literature could be a significant factor in the observed phenomenon of increasing effect sizes. If established researchers get their statistically significant findings published while student researchers or others who find non-significance do not, then effect sizes would tend to increase over time as no one individual or organization reputes earlier findings.

Confirmation bias.

Confirmation bias refers to the psychological phenomenon whereby humans tend to subconsciously look for ideas and information that confirms their earlier beliefs. This information tends to be more readily assimilated and utilized than does information that contradicts what an individual believes (Bushman & Wells, 2001).

Confirmation bias seems a likely cause of increasing effect sizes. As researchers look for studies to help them build the case for their study, they will naturally begin by searching for studies that confirm what they already believe. As they find increasing numbers of these studies, it seems that the results of the study are a foregone conclusion. This may lead researchers to discount or ignore studies that may disagree with what they believe is true about a research question. In a meta-analysis, this may take the form of a researcher applying more stringent selection criteria to studies that don't confirm his or her hypothesis, leading to effect sizes that increase across time.

A synthesis of biases.

It should be noted that all of the above forms of bias that were identified as the most likely explanation of the phenomenon of increasing effect sizes are probably related to one another and would be difficult to parse out and account for individually. That is, a researcher may begin a study on a popular topic (hot stuff bias) by unconsciously looking for studies that confirm a hypothesis (confirmation bias), not bothering to delve too deeply into grey literature as it would be very time consuming and frustrating (grey literature bias), and base a meta-analysis on published studies that show statistically significant effects (positive results bias.) This study may be published in a reputable journal where it is electronically indexed and other researchers pick up one or more pieces and conduct their own research (hot stuff bias again) based on the previously found positive effects (positive results bias) and publish their studies. This phenomenon, across time, is one potential accounting for the phenomenon of increasing effect sizes.

Other biases less likely to explain phenomenon.

As noted above, there are many other sources of publication bias. However, it is less likely that these sources of bias would be significant factors to explain the phenomenon of increasing effect sizes. Those sources of bias less likely to account for the observed phenomenon are: time – lag bias; full publication bias; place of publication bias; outcome reporting bias; multiple publication bias; language bias; database bias; retrieval bias; and media attention bias.

Time-lag bias would seem to support the phenomenon of longitudinally diminishing effect sizes, as observed by Jennions & Moller (2001). In the natural sciences there may be more of an importance placed on refuting the work of other scholars than is seen in educational research. This would account for diminishing effect sizes over time in the natural

sciences while educational research has the opposite phenomenon as there is less emphasis placed on repeating earlier studies.

Full publication bias and outcome reporting bias are a set of related biases where only partial results of studies are reported. While this may be an issue in educational research, these sources of bias are subsumed under the category of grey literature bias.

Place of publication bias, language bias, database bias, and retrieval bias may all impact effect sizes as they are reported. However, with modern indexing of journal articles, these sources of bias are less likely to account for systemic bias in educational research.

Media attention bias is a subset of hot stuff bias and has been discussed above.

5.2 Increasing Effect Sizes Represent Educational Reality

There is another explanation for the phenomenon of longitudinally increasing effect sizes in educational research. It is possible that effect sizes seem to be increasing because they actually are. This is a hopeful notion that as educational researchers have begun to more rigorously conduct research and educational practitioners have received better training in the utilization of research-based educational techniques, that educational practices have become more effective. This would be supported by the fact that over the past 40 years, which is the timeframe of this study, that many states have implemented tougher teacher training and licensure laws and departments of education at universities have taken a more rigorously quantitative approach. However, when one assesses the outcomes of large-scale assessments of student learning across this time period, no similarly significant gains are apparent. It is beyond the scope of this research project to adequately assess the growth of students in comparison to the perceived growth of teacher effectiveness. However, it does seem less

likely that this is the case and more likely that the correct explanation for the phenomenon of longitudinally increasing effect sizes is publication bias.

5.3 Potential Solutions

If, as this study suggests, effect sizes are in fact increasing over time, then this potentially indicates that there is a problem in the publication process that should be corrected by researchers and publishers. Failure to do so may cause misperceptions regarding the efficacy of a host of educational interventions that may diminish the impact of schooling for students, which is an outcome that is patently undesirable. Below is a set of potential solutions to help alleviate this problem.

First, educational researchers should strive to conduct meta-analyses and other research in the most methodologically sound manner possible. Narrative literature reviews should be only used when a research question is either very limited in scope or is so new that very little literature is available such that it would be possible for a researcher to adequately summarize findings from the literature base without quantitative methods. It may also be useful to provide narrative literature reviews as an element of a meta-analysis. Meta-analytic techniques should be included in most literature reviews and these techniques should follow the guidelines set forth by the Cochrane and Campbell Collaborations. These organizations have initiated programming to assist researchers with developing the most accurate summarizations of literature possible. Following their recommendations globally would create a less biased body of educational literature that would be more useful to practitioners and researchers alike.

The other element that would need to change in order for this phenomenon to be ameliorated is to change how educational research is published. First and foremost, there

must be a journal dedicated to publishing only null or statistically insignificant findings. This journal must be indexed properly in major educational research databases and should draw from as many countries and languages as possible. By doing so, researchers who wish to properly conduct meta-analyses will be able to more readily access these results and then conduct a more methodologically sound and less biased meta-analysis. Additionally, a comprehensive effort should be made to index the wide body of grey literature that is generated globally each year. Conference presentations, dissertations, theses, working papers, action research and other forms of grey literature may provide important insight into research questions and should not be ignored. Moreover, publishers should be conservative when announcing special issues or accepting papers on topics that are very new. While this is difficult to do and may not always be advisable, this would help alleviate the problems associated with hot stuff bias that were described above.

5.4 Limitations of the Present Study

This study has two key limitations that should be discussed. First, the studies included in this study came from a very limited subset of educational studies. Many studies were not included if they did not meet the criteria for inclusion. Hence, a more inclusive literature search may invalidate or temper the results found here. Moreover, due to financial constraints of this project, dissertations were excluded from analysis. This presents an unfortunate source of bias that must have some degree of impact on the results.

Second, it is considered best practice for meta-analysis to be conducted using a team of reviewers who would make decisions regarding which studies to include together. This process creates a less biased result. It is possible that had this research been conducted

utilizing a team of researchers or assistants to help determine which studies should be included that the results of this project may have been different.

5.5 Recommendations for Future Research

This project opens up the possibility of expanding this research question more fully in the future. It would be highly worthwhile to begin a new phase of this project by loosening the literature inclusion criteria so that more studies could be included. Specifically, many studies were excluded since they dealt with university classrooms. These studies should certainly be included in any future research. Also, as discussed in the limitations section above, future research should involve the utilization of at least one other co-researcher to diminish bias inherent in the study selection process. Additionally, parsing out the results of this study may provide interesting points of consideration. For example, is there a difference between studies focusing on the English classroom as opposed to the science classroom? Are there differences between studies focusing on early grades and those focusing on later grades? Other distinctions would be possible and may provide fascinating sub-texts to the larger questions.

Beyond this, however, the larger question remains as to the cause of the observed phenomenon. Is this phenomenon caused by pervasive publication biases that should be immediately addressed and remedied or have effect sizes increased because educators have become better at their jobs over the past 40 years? This causal question is truly vexing and should be a primary focus of future research. In general, publication biases are not widely studied in education and should be a source of concern for the community of educational researchers and for those who utilize that research.

APPENDICES

Appendix A

EBSCOHost (Includes Academic Search Complete, Education Research Complete, ERIC, PsycInfo, Social Sciences Full Text, Education Full Text, Psychology and Behavioral Sciences Collection)

- Searched under subject terms: meta-analysis and education; meta-analysis and teaching; meta-analysis and learning. Dates were restricted to the range 1970 – 2012. Only studies published in English were considered. Only peer – reviewed scholarship was considered.
- Searching under the subjects of meta-analysis and education returned 292 results. Searching under the subjects of meta-analysis and teaching returned 132 results. Searching under the subjects of meta-analysis and learning returned 142 results.
- These results were evaluated. Studies were chosen for further consideration if they:
    - Dealt with some sort of pedagogical intervention or technique aimed at improving a cognitive or academic domain.
    - Dealt with primary or secondary education.
    - Did not deal with assessment
    - Did not deal with policy or school improvement.
    - Did not deal with research methodology concerns.
    - Did not deal with classroom management.
    - Did not deal with physical education, medical education, dental education, driver's education, music education, arts education, or distance learning. It was felt that these were specialized forms of education that should be considered separately.

- These searches yielded 129 articles that continued on to secondary consideration.

APA PsycNET

- Searched for following terms in all fields: meta-analysis and education; meta-analysis and teaching; meta-analysis and learning. Dates were restricted to the range 1970 – 2012. Only studies published in English were considered. Only peer – reviewed scholarship was considered.
- Searching under the subjects of meta-analysis and education returned 0 results. Searching under the subjects of meta-analysis and teaching returned 0 results. Searching under the subjects of meta-analysis and learning returned 0 results.
- These results were evaluated using the same criteria as described in the EBSCOHost entry. This process resulted in 0 articles that continued on to the secondary level of consideration.

ArticleFirst

- Searched for following terms in the keyword field: meta-analysis and education; meta-analysis and teaching; meta-analysis and learning. Dates were restricted to the range 1970 – 2012. Only studies published in English were considered. Only peer – reviewed scholarship was considered.
- Searching under the subjects of meta-analysis and education returned 431 results. Searching under the subjects of meta-analysis and teaching returned 28 results. Searching under the subjects of meta-analysis and learning returned 154 results.

- These results were evaluated using the same criteria as described in the EBSCOHost entry. This process resulted in 88 articles that continued on to the secondary level of consideration.

Dissertation Abstracts

- Searched for following terms in the keyword field: meta-analysis and education; meta-analysis and teaching; meta-analysis and learning. Dates were restricted to the range 1970 – 2012. Only studies published in English were considered.
- Searching under the subjects of meta-analysis and education returned 13 results. Searching under the subjects of meta-analysis and teaching returned 3 results. Searching under the subjects of meta-analysis and learning returned 10 results.
- These results were evaluated using the same criteria as described in the EBSCOHost entry. This process resulted in 12 dissertations that continued on to the secondary level of consideration.

Electronic Journal Center

- Searched for following terms in the keyword field: meta-analysis and education; meta-analysis and teaching; meta-analysis and learning. Dates were restricted to the range 1970 – 2012. Only studies published in English were considered. Only peer – reviewed scholarship was considered.
- Searching under the subjects of meta-analysis and education returned 50 results. Searching under the subjects of meta-analysis and teaching returned 48 results. Searching under the subjects of meta-analysis and learning returned 6 results.

- These results were evaluated using the same criteria as described in the EBSCOHost entry. This process resulted in 22 articles that continued on to the secondary level of consideration.

Electronic Dissertation and Theses Center

- Searched for meta-analysis in the keyword field. This database only allowed for search on one term at a time. Dates were restricted to the range 2001 – 2012. Only studies published in English were considered.
- Search returned 63 results.
- These results were evaluated using the same criteria as described in the EBSCOHost entry. This process resulted in 4 dissertations that continued on to the secondary level of consideration.

Expanded Academic ASAP

- Searched for following terms in the keyword field: meta-analysis and education; meta-analysis and teaching; meta-analysis and learning. Dates were restricted to the range 1970 – 2012. Only studies published in English were considered. Only peer – reviewed scholarship was considered.
- Searching under the subjects of meta-analysis and education returned 330 results. Searching under the subjects of meta-analysis and teaching returned 76 results. Searching under the subjects of meta-analysis and learning returned 194 results.

- These results were evaluated using the same criteria as described in the EBSCOHost entry. This process resulted in 165 articles that continued on to the secondary level of consideration.

JSTOR

- Searched for following terms in the abstract field: meta-analysis and education; meta-analysis and teaching; meta-analysis and learning. Dates were restricted to the range 1970 – 2012. Only studies published in English were considered. Journal results were restricted to the fields of education and psychology. Only peer – reviewed scholarship was considered.
- Searching under the subjects of meta-analysis and education returned 28 results. Searching under the subjects of meta-analysis and teaching returned 43 results. Searching under the subjects of meta-analysis and learning returned 22 results.
- These results were evaluated using the same criteria as described in the EBSCOHost entry. This process resulted in 44 articles that continued on to the secondary level of consideration.

Appendix B

Instructions: Review the title and publication information for the study. If a negative answer is generated for questions 1 - 6, then exclude these studies from further consideration.

1. Is the study peer-reviewed from an academic journal?
2. Does the study fall in the date range of 1970 – 2012?
3. Is the study published in English?
4. Does the study deal with some sort of pedagogical concern?
5. Is the pedagogical concern in question primarily a cognitive one?
6. Does the study deal solely with subjects from the primary and secondary grade level?

If the answers to questions 7 – 12 have a Yes answer, then exclude these studies from further consideration.

7. Does the study deal with assessment?
8. Does the study deal with policy or school improvement?
9. Does the study deal with theoretical or methodological concerns?
10. Does the study deal with classroom management issues?
11. Does the study use data taken from any of the following specialized educational environments:
    a. Physical Education
    b. Medical Education
    c. Dental Education

d. Driver's Education

e. Music/ Art Education

f. Distance Education

APPENDIX C
Unobtainable Literature

Adesope, O., Lavin, T., Thompson, T., Ungerleider, C. (2011). Pedagogical strategies for teaching literacy to ESL immigrant students: a meta-analysis. *British journal of educational psychology, 81,* 629 – 653.

Henriksson, W. (1994). Meta – analysis as a method for integrating results of studies about the effects of practice and coaching on test scores. *The British journal of educational psychology, 64,* 319.

Johnson, D. & Johnson, R. (1994). Learning together and alone: overview and meta-analysis. *Asia Pacific journal of education, 64,* 95 – 105.

APPENDIX D

TERTIARY LITERATURE CONSIDERATION FORM

1. Does article provide the following information:
    a. At least one mean or general effect size?
    b. Provide a table or other means of relating the following information taken from the studies that were included in the meta-analysis:
        i. Year of publication
        ii. Sample size
        iii. Effect size

If the study provides the above information, code it into SPSS.

APPENDIX E
Studies Chosen for Meta - Analysis

Abraham, L. (2008). Computer – mediated glosses in second language reading comprehension and vocabulary learning. *Computer assisted language learning, 21,* 199 – 226.

Asher, J., Feldhusen, J. & Vaughn, V. (1991). Meta – analyses and review of research on pull – out programs in gifted education. *Gifted child quarterly, 35,* 92 - 98.

Baker, S., Gersten, R. & Graham, S. (2003). Teaching expressive writing to students with learning disabilities: research – based applications and examples. *Journal of learning disabilities, 36,* 109 - 123.

Blok, H., Oostdam, R., Otter, M. & Overmaat, M. (2002). Computer – assisted instruction in support of beginning reading instruction: a review. *Review of educational research, 72,* 101 - 130

Blok, H. (1999). Reading to young children in educational settings: a meta – analysis of recent research. *Language learning, 49* (2)*,* 343 – 371.

Cable, A., Edmonds, M., Reutebach, C., Schnakenberg, J., Tackett, K., Vaughn, S., & Wexler, J. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of educational research, 79*, 262 – 300.

Dexter, D. & Hughes, C. (2011). Graphic organizers and students with learning disabilities: a meta – analysis. *Learning disabilities quarterly*, *34* (1), 51 – 72.

Ehri, L., Nunes, S., Stahl, S. & Willows, D. (2001). Systematic phonics instruction helps students learn to read: evidence from the National Reading Panel's meta – analysis. *Review of educational research, 71,* 393.

Ehri, L., Nunes, S., Willows, D., Schuster, B., Yaghoub – Zadeh, Z. & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: evidence from the National Reading Panel's meta – analysis. *Reading research quarterly, 36* (3), 250 – 287.

Elbaum, B., Hughes, M., Moody, S. & Vaughn, S. (1999). Grouping practices and reading outcomes for students with disabilities. *Exceptional children, 65*, 399 - 415.

Fukkink, R., Glopper, K. (1998). Effects of instruction in deriving word meaning from context: a meta – analysis. *Review of educational research, 68* (4), 450 – 469.

Geoghegan, D. & O'Neill, S. (2012). Pre-service teachers' comparative analyses of teacher –

parent – child talk: making literacy teaching explicit and children's literacy learning visible. *International journal of English studies, 12,* 97 - 127.

Goodwin, A. & Ahn, S. (2010). A meta – analysis of morphological interventions: effects on literacy achievements of children with literacy difficulties. *Annals of dyslexia, 60* (2), 183 – 208.

Graham, S. & Perin, D. (2007). A meta – analysis of writing instruction for adolescent students. *Journal of educational psychology, 99* (3), 445 – 476.

Hattie, J., Biggs, J. & Purdie, N. (1996). Effects of learning skills interventions on student learning: a meta – analysis. *Review of educational research, 66* (2), 99 – 136.

Jeynes, W. (2008). A meta – analysis of the relationship between phonics instruction and minority elementary school student academic achievement. *Education & urban society, 40* (2), 151 – 166.

Jeynes, W. & Littell, S. (2000). A meta – analysis of studies examining the effect of whole language instruction on the literacy of low – SES students. *The elementary school journal, 101,* 21 – 33.

Klauer, K. (1984). Intentional and incidental learning with instructional texts: a meta – analysis for 1970 – 1980. *American educational research journal, 21* (2), 323 – 339.

Klauer, K.(2008). Inductive reasoning: a training approach. *Review of educational research, 78* (1), 85 – 123.

Lake, C. & Slavin, R. (2008). Effective programs in elementary mathematics: a best – evidence synthesis. *Review of educational research, 78,* 427 - 515.

Sencibaugh, J. (2007). Meta – analysis of reading comprehension interventions for students with learning disabilities: strategies and implications. *Reading improvement,* 44 (1), 6 - 22

Spada, N. & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: a meta – analysis. *Language learning, 60* (2), 263 – 308.

APPENDIX F
CODING FORM FOR META-ANALYSIS LEVEL

All of the following information was entered into SPSS about the meta-analyses included in the study.

Identification of Study

1. Study ID: This is a unique identifying number assigned to each study.
2. Type of Source:
    1. Journal
    2. Book
    3. Book Chapter
    4. Doctoral Dissertation
2. Publication Year
3. Sample size (number of studies used in the meta-analysis)
4. Mean effect sizes (this could result in more than one entry for each study)
5. General content area of study:
    1. Special Ed
    2. Language Arts
    3. Math
    4. Technology
    5. ELL/ESL
    6. Blend of two or more of the above categories
    7. Other
6. Type of studies being meta-analyzed:
    1. Experimental
    2. Quasi-experimental
    3. Blend of both experimental and quasi-experimental

APPENDIX G
CODING FORM FOR INDIVIDUAL STUDY LEVEL

1. Study ID
2. Year of publication
3. Sample size
4. Effect size

References

Alatalo, R. V., Mappes, J. & Elgar, M. (1997). Heritabilities and paradigm shifts. *Nature, 385,* 402 – 403.

Anderson, C.A., & Lindsay, J. J. (1998). The development, perseverance, and change of naïve theories. *Social Cognition, 16,* 8 – 30.

Anderson, C. A., & Sechler, E. S. (1986). Effects of explanation and counter-explanation on the development and use of social theories. *Journal of Personality and Social Psychology, 50,* 23 – 34.

Andrews, G., Guitar, B., & Howie, P. (1980). Meta-analysis of the effects of stuttering treatment. *Journal of Speech and Hearing Disorders, 45,* 287 – 307.

Angell, M. & Relman, A. S. (1989). Redundant publication. *New England journal of medicine, 320,* 1212 – 1213.

Armijo-Olivo, S. (2012). Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. Journal of Evaluation in Clinical Practice, 18(1), 12-18.

Asher, W. (1990). Education Psychology, Research Methodology, and Meta-Analysis. *Educational Psychologist, 25(2),* 143.

Auger, C. P. (1998). *Information sources in grey literature, 4th ed.* London: Bowker Saur.

Bacon, F. (1621/ 1960). *Novum organum*. New York: Bobbs – Merrill.

Bakewell, D. Publish in English, or peril? *Nature, 356,* 648.

Ben-Shlomo, Y., Davey-Smith, G. (1994). Place of publication bias. *British medical journal, 309,* 274.

Birnbaum, R. (2000). *Management fads in higher education.* San Francisco: Jossey-Bass.

Begg, C. B. (1994). Publication bias. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399 – 409). New York: Russel Sage Foundation.

Begg, C. B. & Berlin, J. A. (1988). Publication bias: a problem in interpreting medical data. *Journal of the royal statistical society - series A – statistics in society, 151,* 419 – 463.

Bozarth, J. D., Roberts, R. R. (1972). Signifying significant significance. *American psychologist, 27,* 774 – 775.

Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personal and Social Psychology Bulletin, 27,* 1123 – 1130.

Carlton, P. L., & Strawderman, W. E. (1996). Evaluating cumulated research I: The inadequacy of traditional methods. Biological Psychiatry, 39, 65−72.

Carson, K., Schriesheim, C., Kinicki, A. (1990). The Usefulness of the "Fail-Safe" Statistic in Meta-Analysis. Educational and Psychological Measurement, 50(2), 233-243.

Chalmers, I. Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation & The Health Professions, 25 (1),* 12 – 37.

Chalmers, I., Adams, M., Dickersin, K., Hetherington, J., Tamow-Mordi, W., Meinert, C. (1990). A cohort study of summary reports of controlled trials. *Journal of the American Medical Association, 263,* 1401 – 1405.

Chan, A. W., Hrobjartsson, A., Haar, M. T., Gotzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association, 291,* 2457 – 2465.

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

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Light, R. J., Louis, T. A. & Mosteller, F. (1992). *Meta-analysis for explanation.* New York: Russell Sage Foundation.

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

Cooper, H. & Rosenthal, R. (1980). Statistical vs. traditional procedures for summarizing research findings. *Psychological bulletin, 87,* 442 – 449.

Cordes, A. K. (1998). Current status of the stuttering treatment literature. In A. K. Cordes & R. J. Ingham (Eds.), *Treatment efficacy for stuttering: A search for empirical bases* (pp. 117 – 144). San Diego, CA: Singular.

Coursol, A. & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: a note on meta-analysis bias. *Professional psychology, 17,* 136 – 137.

Csada, R. D., James, P. C., Espie, R. H. M. (1996). The file drawer problem of nonsignificant results – does it apply to biological research? *Oikos, 76,* 591 – 593.

Cuban, L. (2004). "The open classroom: schools without walls became all the rage during the early 1970s. Were they just another fad? - Whatever Happened to …?". *Education Next. 4(2).* Retrieved from: http://educationnext.org/theopenclassroom/ on March 3, 2011.

Davidson, R. A. (1986). Source of funding and outcome of clinical trials. *Journal of general internal medicine, 1,* 155 – 158.

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford review of education, 26,* 365 – 378.

DeBellefeuille, C., Morrison, C. A., Tannock, I. F. (1992). The fate of abstracts submitted to a cancer meeting: factors which influence presentation and subsequent publication. *Annals of oncology, 3,* 187 – 191.

Detsky, A., Baker, J., O'Rourke, K., Goel, V. (1987). Perioperative parenteral nutrition: a meta-analysis. *Annals of Internal Medicine,* 107, 195 – 203.

Devine, E. C. (1999). Empirical assessment of publication bias: lessons from two meta-analyses. In: *Proceedings of the 7th Cochrane Colloquium;* Oct 5 – 9; Rome. Rome: Universitas Tommaso D' Aquino: 60.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association, 263,* 1385 – 1389.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H.R. Rothstein, A. J. Sutton, & M. Bornstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 11 – 33). Chichester, UK: John Wiley & Sons.

Dickersin, K., Hewitt, P., Mutch, L., Chalmers, I., Chalmers, T. C. (1985). Perusing the literature: comparison of MEDLINE searching with a perinatal trials database. *Controlled clinical trials, 6,* 306 – 317.

Easterbrook, P. J., Berlin, J. A., Gopalan, R. Matthews, D. R. (1991). Publication bias in clinical research. *Lancet, 337,* 867 – 872.

Egger, M. & Davey-Smith, G. Bias in location and selection of studies. *British medical journal, 316,* 61 – 66.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychology, 33,* 517.

Felson, D. T. (1992). Bias in meta-analytic research. *Journal of Clinical Epidemiology, 45,* 885 – 892.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology, 48,* 71 – 79.

Gavaghan, D., Moore, A., McQay, H. (2000). An evaluation of homogeneity tests in meta-analysis in pain using simulations of patient data. *Pain, 85*, 415 – 424.

Gibbs, L. E. (2003).*Evidence-based practice for the helping professions: A practical guide with integrated multimedia.* Pacific Grove, CA: Brooks/ Cole-Thompson Learning.

Glass, G. (1976). Priamry, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3 – 8.

Glass, G., McGaw, B., & Smith, M.L. (1981) *Meta-analysis in social research*. Beverly Hills: SAGE Publications.

Gontard-Danek, M. C. & Moller, A. P. (1999). The strength of sexual selection: a meta-analysis of bird studies. *Behavioral ecology, 10,* 476 – 486.

Gotzsche, P.C., Hrogjartsson, A., Maric, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *Journal of the American Medical Association, 298,* 430 – 437.

Gotzsche, P. C., Lange, B. (1991). Comparison of search strategies for recalling doublie-blind trials from Medline. *Danish medical bulletin, 38,* 476 – 478.

Gregoire, G., Derderian, F., Lorier, J. L. (1995). Selecting the language of the publications included in a meta-analysis: is there a tower of Babel bias? *Journal of Clinical Epidemiology, 48,* 159 – 163.

Hargreaves, D. H. (1997). In defence of research for evidence-based teaching: a rejoinder to Martyn Hammersley. *British Educational Research Journal, (23), 4,* 405 – 419.

Hedges, L.V., Laine, R., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Research, 23,* 5 – 14.

Hedges. L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychology Bulletin, 88,* 359 – 369.

Helfenstein, U. (2002). Data and models determine treatment proposals – an illustration from meta-analysis. *Postgrad Medical Journal, 78,* 131 – 134.

Higgins, J., & Green, S. (Eds). (2005).Cochrane Handbook for Systematic Reviews of Interventions, Version 4.2.5 In: The Cochrane Library, Issue 3, 2005. Chichester, UK: John Wiley & Sons, Ltd. Accessed March 1, 2011 at: http://www.cochrane.org/resources/handbook/hbook.htm.

Higgins, J., & Thompson, S. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine, 21,* 1539 – 1558.

Higgins, J., Thompson, S., Deeks, J., & Altman, D. (2003) Measuring inconsistency in meta-analyses. *British medical journal, 327,* 557 – 560.

Hillage, J., Pearson, R., Anderson, A., & Tamkin, P. (1998). *Excellence in research on schools: Research report RR74.* Sudbury, UK: DfEE Publications.

Hilmer, F. G., & Donaldson, L. (1996). *Management redeemend: debunking the fads that undermine corporate performance.* New York: Free Press.

Hilton, J. L., & von Hippell, W. (1990). The role of consistency in the study of stereotype-relevant behaviors. *Personality and Social Psychology Bulletin, 16,* 430 – 448.

Hopewell, S., Clarke, M., Lefebvre, C., & Scherer, R. (2006). Handsearching versus electronic searching to identify reports of randomized trials. The Cochrane Database of Systematic Reviews, 2006, Issue 4 Chichester, UK: John Wiley & Sons, Ltd.

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2006). Grey literature in meta-analyses of randomized trials of health care interventions. The Cochrane Database of Systematic Reviews, 2006, Issue 2 Chichester, UK: John Wiley & Sons, Ltd.

Horder, T. J. (2001). The organizer concept and modern embryology: Anglo-American perspectives. *International Journal of Developmental Biology, 45,* 97 – 132.

Hubbard, R., Armstrong, J. S. Publication bias against null results. *Psychological reports, 80,* 337 – 338.

Ingham, R. J. (1984). *Stuttering and behavior therapy: Current status and empirical foundations.* San Diego, CA: College-Hill Press.

Ingham, R. J. (2002). Yet another "exercise in mega-silliness?" *Journal of Fluency Disorders, (27),* 169 – 174.

Ioannidis, J. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Journal of the American Medical Association, 279,* 281 – 286.

Jadad, A. R. & Rennie, D. (1998). The randomized controlled trial gets a middle-aged checkup. *Journal of the American Medical Association, 279,* 319 – 320.

Jorgensen, A. W., Hilden, J., & Gotzsche, P. G. (2006). Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: Systematic review. *British Medical Journal, 333,* 782 – 785.

Koren, G., Klein, N. (1991). Bias against negative studies in newspaper reports of medical research. *Journal of the American Medical Association, 266,* 1824 – 1826.

Kozloff, M. (2002) http://people.uncw.edu/kozloffm/fads.html

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological bulletin, 91,* 404 – 412.

Kulik, J. A. & Kulik, C-L. C. (1989). Meta-analysis in education. *International journal of educational research, (13),* 221 – 340.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: procedures for resolving contradictions among different research studies. *Harvard educational review, (41),* 429 – 471.

Lipsey, M. W. & Wilson, D. B. (1993). The efficacy of psychological, educational and behavioural treatment: Confirmation from meta-analysis. *American psychologist, (48),* 1181 – 1209.

Littell, J. H. (2008). Evidence-based or biased? The quality of published reviews of evidence-based practices. *Children & Youth Services Review, 30(11),* 1299-1317.

Littell, J. H. (2005). Lessons from a systematic review of effects of Multisystemic Therapy. *Children and Youth Services Review, 27,* 445 – 463.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York, NY: Oxford University Press.

Littell, J. & Shlonsky, A. (2011). Making Sense of Meta-Analysis: A Critique of 'Effectiveness of Long-Term Psychodynamic Psychotherapy'. Clinical Social Work Journal, 39(4), 340-346.

Landr, V. L. (1990). The publication outcome for the papers presented at the 1990 ABA conference. *Journal of burn care rehabilitation, 17,* 23A – 26A.

Lord, C. G., Ross, L. & Lepper, M. (1970). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 47,* 1231 – 1243.

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., et al. (1999). The researcher's own therapy allegiances: A 'wild card' in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice, 6,* 95 – 106.

Mahoney, M. J. (1977). Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161 – 175.

McAuley, L., Moher, D., Pham, B., Tugwell, P. (1999). Evaluation of the impact of grey literature in meta-analysis. In: Proceedings of the 7th Cochrane Colloquium; Oct 5 – 9; Rome. Rome: Universitas Tommaso D' Aquino: 17.

Misakian, A. L., Bero, L. A. (1998). Publication bias and research on passive smoking. Comparison of published and unpublished studies. *Journal of the American Medical Association, 280,* 250 – 253.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D. F., et al. (1999). Improving the quality of reports of meta-analyses of randomized controlled trials: The QUORUM statement. *The Lancet, 354,* 1896 – 1900.

Moher, D., Dulberg, C. S., Wells, G. A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association, 272,* 122 – 124.

Moscati, R., Jehle, D., Ellis, D., Fiorello, A., Landi, M. (1994). Positive – outcome bias: Comparison of emergency medicine and general medicine literatures. *Academic emergency medicine, 1,* 267 – 271.

Mulward, S. & Gotzsche, P. C. (1996). Sample-size of randomized double – blind trials 1976 – 1991. *Danish Medical Bulletin, 43,* 96 – 98.

Ottenbacher, K., Difabio, R. P. (1985). Efficacy of spinal manipulation/ mobilization therapy. A meta-analysis. *Spine, 10,* 833 – 837.

Palmer, A. R. (2000). Quasireplication and the contract of error: Lessons from sex ratios, heritabilities and fluctuating asymmetry. *Ann Rev Ecol Sys, 31,* 441 – 480.

Petrosino, A., Turpin-Petrosino, C., & Buehler, J. (2005). Scared straight and other juvenile awareness programs for preventing juvenile delinquency. *Scientific Review of Mental Health Practice, 4(1)*, 48-54.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide.* Oxford, UK: Blackwell Publishing.

Pillemer, D. B. (1984). Conceptual issues in research synthesis. *Journal of Special Education, 18,* 27 – 40.

Poulin, R. (2000). Manipulation of host behaviour by parasites: a weakening paradigm? *Proceedings of the Royal Society of London, 267,* 787 – 792.

Rigby, D. (1998). *Management tools and techniques.* Boston: Bain and Company.

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The Effectiveness of Volunteer Tutoring Programs for Elementary and Middle School Students: A Meta-Analysis. *Review of Educational Research, 79(1),* 3-38.

Roe, E. *Narrative policy analysis: Theory and practice.* Durham, N.C.: Duke University Press.

Rosenberg, M., Adams, D., & Gurevitch, J. (2000). *Meta-win: statistical software for meta-analysis, v. 2.0.* Sunderland, MA: Sinauer.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological bulletin, 85,* 185 – 193.

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

Rosenthal, R. (1988). Publication bias, retrieval bias and pipeline effects. Discussion of the paper by Begg and Berlin. *Journal of the royal statistical society series A – statistics in society, 151,* 419 – 463.

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

Rosenthal, M. C. (1994). The fugitive literature. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 85−94). New York: Russell Sage Foundation.

Rothstein, H. R., Turner, H. M., & Lavenberg, J. G. (2004). The Campbell Collaboration Information Retrieval Policy Brief. Retrieved June 12, 2006, from http://www.campbellcollaboration.org/MG/IRMGPolicyBriefRevised.pdf

Rothstein, H., Sutton, A. J., & Bornstein, M. (2005). Publication bias in meta-analysis: Prevention, assessment, and adjustments Chichester, UK: Wiley.

Rotton, J. Foos, P. W., Vanmeek, L., Levitt, M. (1995). Publication practices and the file drawer problem – a survey of published authors. *Journal of social behavior and personality*, *10,* 1 – 13.

Rothwell, P. M., & Robertson, G. (1997). Meta-analyses of randomized controlled trials. *Lancet, 350,* 1181 – 1182.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Disease, 32,* 51 – 53.

Sackett, D.L., Haynes, R.B., Guyatt, G.H. & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Boston: Little, Brown.

Scher, L. S., Maynard, R. A., & Stagner, M. (2006). Interventions intended to reduce pregnancy-related outcomes among teenagers. In *The Campbell Collaboratio Library.* Retrieved March 1, 2011, from http://www.campbellcollaboration.org/doc-pdf/teenpregreview_dec2006.pdf.

Scherer, R. W., Langenberg, P., & von Elm, E. (2007). Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews, 2.*

Shadish, W., & Myers, D. (2004). Campbell Collaboration Research Design Policy Brief. Retrieved March 1, 2011 from http://www.campbellcollaboration.org/MG/resdespolicybrief.pdf

Shapiro, S. (1994). Meta-analysis/ shmeta-analysis. *American Journal of Epidemiology, 140,* 771 – 778.

Shlonsky, A., Noonan, E., Littell, J., Montgomery, P. (2011). The Role of Systematic Reviews and the Campbell Collaboration in the Realization of Evidence-Informed Practice. Clinical Social Work Journal, 39(4), 362-368.

Silagy, C. A. (1993). Developing a register of randomised controlled trials in primary care. *British medical journal, 306,* 897 – 900.

Simes, R. J. (1986). Publiation bias: The case for an international registry of clinical trials. *Journal of clinical oncology, 4,* 1529 – 1541.

Simes, R. J. (1987). Confronting publication bias: a cohort design for meta analysis. *Statistics in medicine, 6,* 11 – 29.

Simmons, L. W., Tomkins, J. L., Kotiaho, J. S. & Hunt, J. (1999). Fluctuating paradigm. *Proceedings of the Royal Society of London, 266,* 593 – 595.

Slavin, R. E. (1984). Meta-analysis in education: How has it been used? *Educational researcher, 13,* 6 – 15.

Slavin, R. E. (1986). Best evidence synthesis: An alternative to meta-analysis and traditional reviews. *Educational researcher, 15,* 5 – 11.

Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychology, 5,* 225 – 232.

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

Sohn, D. (1996). Publications bias and the evaluation of psychotherapy efficacy in reviews of the research literature. *Clinical psychology review, 16,* 147 – 156.

Song, F., Eastwood, A. J., Gilbody, S., Duley, L. & Sutton, A. J. (2000). Publication and related biases. *Health technology assessment, 4, (10),* 1 – 125.

Song, F & Gilbody, S. (1998). Increase in studies of publication bias coincided with increasing use of meta-analysis. *British medical journal, 316,* 471.

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

Sterling, T. D., Rosenbaum, W. L., Weinkam, J. J. (1995). Publication decisions revisitied – the effect of the outcome of statistical tests on the decision to publish and vice – versa. *Am Stat, 49,* 108 – 112. – need good cite

Stern, J. M. & Simes, R. J. (1997). Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. *British medical journal, 315,* 640 – 645.

Thomas, C., & Howell, P. (2001). Assessing efficacy of stuttering treatments. *Journal of Fluency Disorders, 26,* 311 – 333.

Tornatzky, L. G., & Fleischer, M. *Technological innovation.* San Francisco: New Lexington Press.

Tooley, J. & Darby, D. (1998). *Educational research: An Ofsted critique.* London, UK: OFSTED.

Torgerson, C. J. (2006). Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies, 54,* 89 – 102.

Tramer, M. R. Reynolds, D. J. M., Moore, R. A., McQuay, H. J. (1997). Impact of covert duplicate publication on meta-analysis: a case study. *British medical journal, 315,* 635 – 640.

Tregenza, T. & Wedell, N. (1997). Natural selection bias. *Nature, 386,* 234.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207 – 302.

Vandenbrouche, J. P. (1989). On not being born a native speaker of English. *British Medical Journal, 289,* 1461 – 1462.

Vickers, A., Goyal, N., Harland, R., Rees, R. (1998). Do certain countries produce only positive results? A systematic review of controlled trials. *Controlled Clinical Trials, 19,* 159 – 166.

Wachter, K W (Sept 16, 1988). Disturbed by meta-analysis? Science, 241, n4872. p.1407(2). Retrieved December 23, 2010, from Expanded Academic ASAP via Gale:

http://find.galegroup.com/gtx/start.do?prodId=EAIM&userGroupName=clev91827

Wells, H. G. (1938). *World brain*. Garden City, N.Y: Doubleday, Doran & Co.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91,* 461 – 481.

Williamson, P. R., & Gamble, C. (2005). Identification and impact of outcome selection bias in meta-analysis. *Statistics in Medicine, 24,* 1457 – 1561.

Williamson, P., Altman, D., Gamble, C., Dodd, S., Dwan, K., & Kirkham, J., (2006, February). *Outcome reporting bias in meta-analysis.* Paper presented at the Fourteenth Cochrane Colloquium, Dublin, Ireland.

Wolf, F. (1986). "Meta-analysis." Sage University Paper series on Quantitative Applications in the Social Sciences. Beverly Hills: Sage Publications.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 97−109). New York: Russell Sage Foundation.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985).Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Research, 27,* 336 – 371.

Zar, J. (1984). *Biostastical analysis,* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.