Education and Science

Vol 39 (2014) No 174 1-32

A Research Methodology to Conduct Effective Research Syntheses: Meta-Analysis *

Ulaş Üstün [1], Ali Eryılmaz [2]

Abstract

The need for comprehensive and systematic research syntheses has been increasing as the number of primary studies has proliferated in educational sciences in recent years. That is the main reason why diverse applications of meta-analysis, which is one of the most effective ways of research synthesis, has been encouraged in many disciplines including social and educational sciences. The main purpose of this article is to provide a conceptual framework for meta-analysis by questioning its criticisms and strengths over other methods of research synthesis. In addition, some methodological and statistical considerations like comparison of fixed-effect and random-effects models, different measures of effect size, unit of analysis, validity issues including publication bias and quality of primary studies, heterogeneity, moderator and power analyses in meta-analysis are discussed in detail. Moreover, brief information about software for statistical analyses performed in meta-analysis and a summary of standards for reporting meta-analysis are provided within the scope of this article. How well meta-analyses are conducted and reported is crucial since it has essential roles not only for cumulative nature of science but for policy makers and practitioners as well. Thus, we aim to provide the meta-analysts with an introductory-level guideline for their meta-analysis research.

Keywords

Meta-analysis
Research synthesis
Effect size
Moderator analysis
Fixed-effect model
Random-effects model
Publication bias

Article Info

Received: 02.27.2014
Accepted: 07.14.2014
Online Published: 08.06.2014

DOI: 10.15390/EB.2014.3379

Introduction

Research synthesis is an indispensable part of scientific enterprise not only because of cumulative nature of scientific knowledge but also the role of research syntheses in providing policy makers with guidance in the light of powerful scientific evidence and its potential to explain the inconsistent data in the literature (Üstün, 2012).

* This study is mainly based on methodological part of the dissertation of Üstün (2012)

[1] Artvin Çoruh University, Faculty of Education, Department of Elementary Education, Turkey, ulasustun@artvin.edu.tr
[2] Middle East Technical University, Faculty of Education, Department of Secondary Science and Mathematics Education, Turkey, eryilmaz@metu.edu.tr

Being cumulative is one of the most important aspects of scientific process, which makes science grow exponentially as well. That was the logic behind what Isaac Newton stated over 300 years ago: "If I have seen further, it is by standing on the shoulders of giants". Although the idea has been obvious and almost noncontroversial throughout the history of science, it has been very recent that the responsibility of scientists in synthesizing old scientific knowledge to integrate into new ones has been acknowledged (Chalmers, Hedges, & Cooper, 2002). Today, it is widely accepted that research syntheses have a key role not only to create links between old and new scientific knowledge by giving an overall or more complete picture of existing paradigm but also to help broaden the scope of the existing knowledge (Card, 2012; Chalmers et al., 2002; Chan & Arvey, 2012; Hunter & Schmidt, 2004; Mulrow, 1994).

The contribution of research syntheses to cumulative nature of scientific endeavors is essential, yet growing academic recognition and popularity of this methodology results from what it serves for policy makers and practitioners (Chalmers et al., 2002). In this respect, Petticrew and Roberts (2006) make an analogy between a single study and a single respondent in a survey. The analogy is based on the necessity of many respondents to reach a conclusion in a survey. They claim that a single response is valuable but it is possible to get an opposite answer from the next respondent. Thus, any conclusion should be based on many responses from many participants. They infer that the decisions made by policy makers and practitioners should be constructed upon the consensus derived from many studies as well. Similarly, Davies (2000) emphasizes that a single experiment no matter how well designed and conducted, is limited by its unique properties like "time, sample and context specificity". Furthermore, emphasizing the function of research synthesis on the process of making decisions, Chalmers et al. (2002) assert that the forthcoming position of research synthesis will likely be created by those from outside academia, who face the reality that bits of information provided by single studies are of little help to the people who will make decisions based on the research findings.

In addition to contributions to cumulative scientific knowledge and the guidance to policy makers and practitioners, another reason why research synthesis is an essential part of scientific endeavor is its potential to assess the consistency of relationships and to explain any data inconsistencies and conflicts in the literature (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hunt, 1997; Hunter & Schmidt, 2004; Mulrow, 1994; Petticrew & Roberts, 2006). No matter which scientific discipline is in perspective, it is not uncommon to find contradictory results from similarly designed research studies on the same topic (Rosenthal & DiMatteo, 2001). However, in social and educational sciences, the situation becomes more complex since human behavior is more complicated and difficult to explain, and there exist many threats to internal validity of the study which are not easy to get rid of completely. In this sense, Berliner (2002) points out that "In my estimation, we (educational researchers) have the hardest-to-do science of them all! We do our science under conditions that physical scientists find intolerable". He claims that contexts include 10^{th} or 15^{th} order interactions during classroom teaching in an educational research studies like interaction between teacher behavior and socioeconomic factors, motivation to learn and many others, which results in many conflicting findings in educational research. Accordingly, educational research is highly criticized in recent years since much research has been unhelpful for policy makers and practitioners to determine what works and what does not work (J. Bennett, 2005). From this perspective, research synthesis should be highly encouraged in educational research as it may functionally serve to summarize overall findings and explain the reasons for any heterogeneity or contradictions in those findings.

The main purpose of this article is to introduce meta-analysis as an effective way of research synthesis by providing a framework including its criticisms and strengths over other methods of research synthesis. The details of how to conduct a meta-analysis with statistical and methodological considerations are explained in this article as well.

Within this context, firstly the unique properties and brief history of meta-analysis is explained in the next sections. Then, the strengths of meta-analysis over other methods of research synthesis are discussed, after which its criticism is questioned in detail. After comparison of two main models used in meta-analysis, how to handle with validity issues in meta-analysis studies is clarified by explaining the methods to identify, quantify, and adjust publication bias. Next, coding reliability, different types of effect size estimates, and the ways to perform heterogeneity, moderator and power analyses are analyzed. Afterwards, performing meta-analysis by using software programs is discussed. Finally, the standards developed for reporting meta-analysis studies are summarized within the scope of this article.

What is Meta-Analysis?

It is evident from the literature that there is almost a consensus on the fact that meta-analysis is one of the most widely-used methods of conducting research synthesis (Lipsey & Wilson, 2001; Schulze, 2007). However, there is no agreement about what "meta-analysis" actually refers to in the literature. Some researchers define "meta-analysis" as a research methodology while others refer to an analysis technique used within research synthesis (Shelby & Vaske, 2008). Cooper and Hedges (2009) claim that "meta-analysis" is often used as a synonym for research synthesis, namely as a research methodology. However, they choose to use the term as a statistical analysis in research synthesis rather than the entire enterprise of research synthesis. Similarly, Glass, the eponym of the term of "meta-analysis", uses the term to refer to "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (1976, p. 3). Nevertheless, he emphasizes that "the *sine qua non* of meta-analysis is the application of research methods to the characteristics and findings of reports of research studies" (1982, p. 93). In addition, Glass, McGaw, and Smith (1981) point out that "…it is not a technique; rather it is a perspective that uses many techniques of measurement and statistical analysis" (p. 21). Shelby and Vaske (2008) call attention to this dissensus about definition of meta-analysis stating that "What constitutes a true meta-analysis is debatable" (p. 97). On the other hand, Rosenthal and DiMatteo (2001) claims that with the work of Smith and Glass (1977), it became obvious that "meta-analysis is more than a statistical technique; it is a methodology for systematically examining a body of research…" (p. 62).

Furthermore, Glass (1976) identifies the relationship between primary analysis, secondary analysis, and meta-analysis. He defines primary analysis as "the original analysis of data in a research study" and secondary analysis as "the re-analysis of data for the purpose of answering the original research question with better statistical techniques, or answering new questions with old data" (p. 3). He claims that meta-analysis refers to "analysis of analyses" and aims to advance the practice of secondary analysis.

In this article, meta-analysis is used to refer to a total enterprise of research synthesis; that is to say, the term of "meta-analysis" is used as a research methodology throughout the article. It is mainly because meta-analysis has unique properties in some parts of research steps like coding for possible moderator variables; accordingly, defining it just as a statistical technique would exclude these characteristics. It is evident from the literature that some researchers define "meta-analysis" in a similar way (Fitz-Gibbon, 1985; Gliner, Morgan, & Harmon, 2003; Lundahl & Yaffe, 2007; Normand, 1999; Rosenthal & DiMatteo, 2001; Sánchez-Meca & Marín-Martínez, 2010a). Viewed in this light, meta-analysis can be defined as "a research methodology that aims to quantitatively integrate the results of a set of primary studies about a given topic in order to determine the state of the art on that topic" (Sánchez-Meca & Marín-Martínez, 2010a, p. 274).

A Brief History of Meta-Analysis

The study conducted by Karl Pearson (1904) to synthesize findings from different studies by using average correlation coefficients can be accepted as the starting point of research synthesis as we know it today (Chalmers et al., 2002; Lipsey & Wilson, 2001; O'Rourke, 2007). However, Lipsey and Wilson (2001) claim that the modern epoch of meta-analysis began with the works of Glass (1976), Rosenthal and Rubin (1978), Schmidt and Hunter (1977), Smith and Glass (1977), Rosenthal and Rubin (1978), and Smith, Glass, and Miller (1980). Although there has been some criticism about its use as a research synthesis methodology (Eysenck, 1978, 1984, 1994; Feinstein, 1995; Shapiro, 1994), the number of meta-analysis studies in different fields has been gradually grown up and meta-analysis has become increasingly more popular as a method of quantitative research synthesis since 1976 when Glass coined the term of "meta-analysis"(Berman & Parker, 2002; Dalton & Dalton, 2008; Fitzgerald & Rumrill, 2003, 2005; Hedges, 1992; Hunter & Schmidt, 2004; Marin-Martinez & Sanchez-Meca, 1999; Sánchez-Meca & Marín-Martínez, 1998; Shelby & Vaske, 2008).

The search for the key term "meta-analysis" as "topic" by using the databases of Web of Science, which covers Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (A&HCI) with Conference Proceedings Citation Index in Science (CPCI-S) and in Social Sciences and Humanities (CPCI-SSH), gives a total of 45,519 results published during the time interval from 1976 to 2012. Figure 1 shows how publication numbers in five years-time intervals increase from the beginning of the modern era of meta-analysis to 2011. In addition, a cited reference search via Web of Science for the keywords "meta-analysis" and "education" results in 38,806 citations for the same time interval with the previous search, which gives an idea about the impact of meta-analysis on educational studies. More interestingly, as illustrated in Figure 2, the number of citations increases exponentially especially in the last 20 years. The number of average citations per year, which is 384 for the time interval from 1991 to 2000, reaches to a very high value, 2898, for the next 11 years from 2001 to 2011. Finally, according to citation report based on this search, the number of average citations per study is 33.66 and the h-index is 94 meaning that, in the scope of Web of Science, there exist 94 meta-analysis studies about education having 94 or more citations, which shows how essential meta-analysis studies are for educational research.

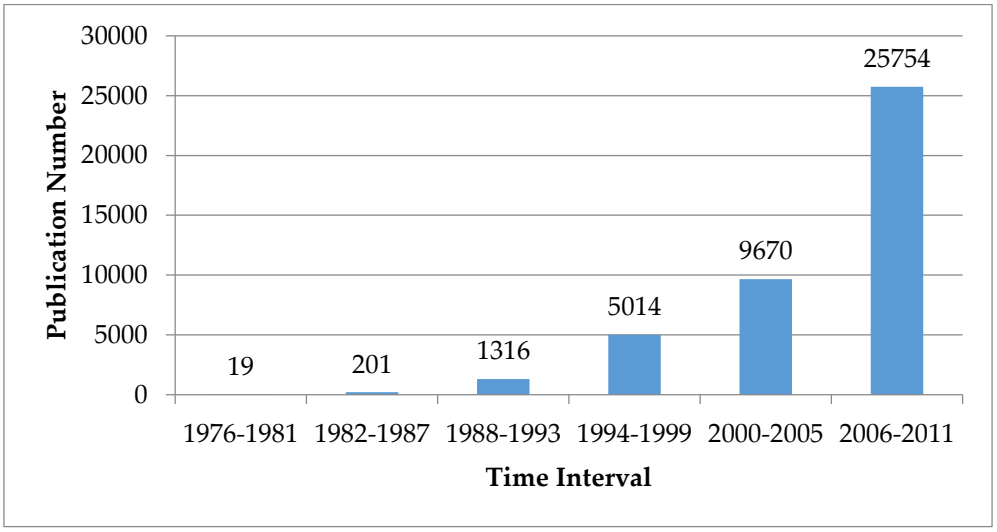

Figure 1. Results of the Search for the Key Term 'Meta-Analysis' from 1976 to 2011

Why Meta-Analysis rather than Other Research Synthesis Methods?

Research synthesis, which aims "to integrate empirical research for the purpose of creating generalizations" (Cooper & Hedges, 2009, p. 6), can be conducted by means of qualitative, quantitative or mixed methods including conventional (traditional, or narrative) review, conventional vote-counting method, combined significance test and meta-analysis. In the next sections, meta-analysis is compared with other methods of research synthesis.

Meta-analysis versus Vote-counting Method and Combined Significant Test

Conventional vote-counting method and combined significance test are two quantitative methods that can be used in the scope of research syntheses. Conventional vote-counting method is simply based on tally of significant and nonsignificant results and the overall decision is made by

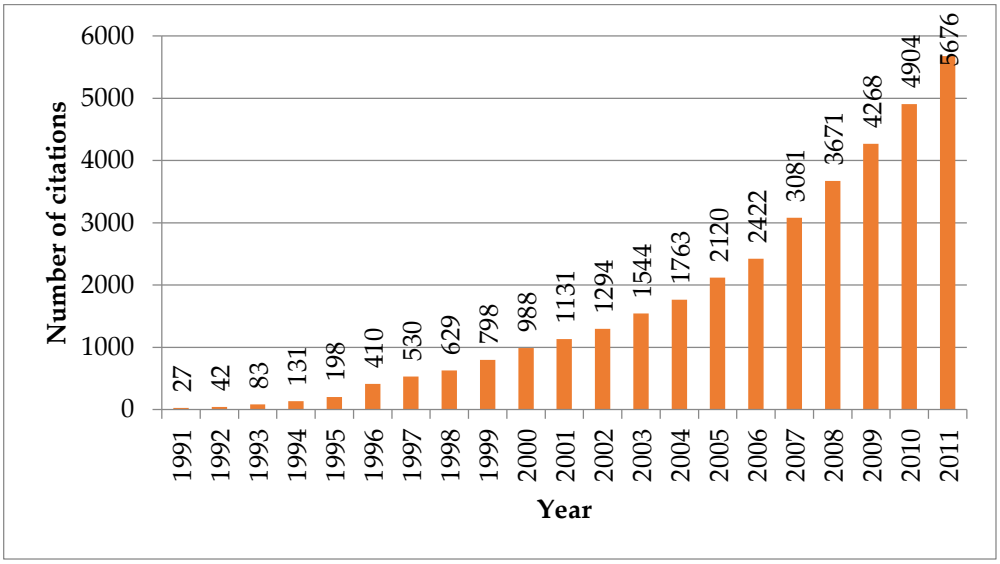

Figure 2. Results of the Cited Reference Search for the Keywords 'Meta-Analysis' and 'Education' between 1991 and 2011.

counting the votes of each category (Borenstein et al., 2009; Bushman & Wang, 2009; Davies, 2000) while combined significance test aims to statistically test the combined probabilities of results of the studies to be reviewed for significance (Bligh, 2000; Fitzgerald & Rumrill, 2003, 2005). Although these methods have a common advantage of being more objective than conventional reviews by minimizing subjective judgment, both suffer from the problems originated from statistical significance test (Fitzgerald & Rumrill, 2003, 2005). In addition, Hedges and Olkin (1980) show that as the number of studies having statistical power less than .50 increases, the probability of making false decisions using vote counting method increases as well if a true effect exists. Thus, Hunter and Schmidt (2004) state "the traditional voting method is fatally flawed statistically and logically" (p. 447). Furthermore, as conventional reviews, both vote counting method and combined significance test are criticized that neither of them allow researchers to investigate the effects of study characteristics (Fitzgerald & Rumrill, 2003, 2005).

It is clearly evident from the literature that faulty use of statistical significance, which gives us the extent to which the results are different from what would be expected due to chance, leads to flawed and conflicting results (Ellis, 2010; Fan, 2001; Hunter & Schmidt, 2004; Kirk, 1996, 2001; Olejnik & Algina, 2000; Schmidt, 1992, 1996; Vacha-Haase, 2001). It is mainly due to the fact that researchers rarely distinguish between the statistical and practical significance, which provides us with an idea about how useful the results are in the real world (Ellis, 2010; Kirk, 1996). The more problematic situation emerges when the results shown to be statistically significant are interpreted as if they are practically significant because it is not uncommon in the literature to reach a result, which is statistically significant but trivial as well (Ellis, 2010; Olejnik & Algina, 2000). Thus, some researchers

suggest that statistical testing should be abandoned (Hunter & Schmidt, 2004; Schmidt, 1996), still some others argue that these tests should be used but effect size should be more emphasized (Cohen, 1990; Kirk, 1996, 2001; Vacha-Haase, 2001). Although how to utilize from statistical significance tests is a controversial issue, a consensus on the fact that statistical significance does not always guarantee practical significance has already been constructed in the literature (Borenstein et al., 2009; Cohen, 1990; Ellis, 2010; Gravetter & Walnau, 2007; Hunter & Schmidt, 2004; Kirk, 1996, 2001; Schmidt, 1996; Vacha-Haase, 2001). Thus, Cohen underlines that "I have learnt and taught that the primary product of a research is one or more measures of effect size, not p values" (1990, p. 1310). Cohen emphasizes another point, in the same paper:

> I am happy to say that the long neglect of attention to effect size seems to be coming to a close. The clumsy and fundamentally invalid box-score method of literature review based on p values is being replaced by effect-size-based meta-analysis as formulated by Gene Glass (1977)…Meta-analysis makes me very happy (1990, pp.1309-1310).

As pointed out by Cohen, the strength of meta-analysis over other quantitative methods of research synthesis come from the fact that it is not based on statistical significance, rather it uses effect size measures of the results (Borenstein et al., 2009; Shelby & Vaske, 2008). Thus, Hunter and Schmidt (2004) recommend two alternatives to the statistical significance tests, which are confidence interval for primary studies and meta-analysis at the level of secondary studies.

In addition to the strength of being based on practical significance rather than p values, another advantage of meta-analysis is that it allows researchers to investigate the effect of moderator variables like study characteristics, which is almost impossible to be performed by other qualitative or quantitative methods of research synthesis (Borenstein et al., 2009; Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001). The opportunity of handling large amount of data from primary studies, increased power and enhanced precision are just some of the other reasons why meta-analysis is labeled as one of the most useful way of conducting a research synthesis (Borenstein et al., 2009; Cohn & Becker, 2003; Gliner et al., 2003; Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001).

Meta-analysis versus Conventional (Traditional) Review

Conventional review, which is a traditional, non-systematic alternative of research synthesis, suffers from serious disadvantages and limitations (Borenstein et al., 2009; Bushman & Wells, 2001; Carlton & Strawderman, 1996; Cooper & Rosenthal, 1980; Fitzgerald & Rumrill, 2003, 2005; Littell, Corcoran, & Pillai, 2008; Petticrew & Roberts, 2006; Torgerson, 2003). Conventional review, also called as traditional, or narrative review, is often conducted by an expert on the specific topic of the review, which, unfortunately, does not guarantee to produce an unbiased and reliable summary of primary studies (Petticrew & Roberts, 2006). Subjective judgment the degree to which is hardly ever explained, biased and unrepresentative sample of studies due to unsystematic way of inclusion of studies, and no explicit reasoning for weighting procedure are pointed out as some of the problems in conventional review (Bushman & Wells, 2001; Carlton & Strawderman, 1996; Cooper & Rosenthal, 1980; Fitzgerald & Rumrill, 2003, 2005; Littell et al., 2008; Oakley, 2002; Petticrew & Roberts, 2006; Torgerson, 2003). Other limitations of conventional reviews are that they are unable to investigate the effects of study characteristics and to establish overall magnitude of effect (Bushman & Wells, 2001; Fitzgerald & Rumrill, 2003, 2005). Finally, traditional narrative reviews become less useful as increasing number of studies leads to enormous information to be synthesized (Borenstein et al., 2009; Glass, 2006; Hunter & Schmidt, 2004). As a result of these weaknesses, it is not an exceptional situation for different researchers conducting conventional reviews on the same research question to reach different and misleading conclusions (Fitzgerald & Rumrill, 2005) as illustrated by Cooper and Rosenthal (1980), Oakley (2002), and Bushman and Wells (2001).

Initially, Cooper and Rosenthal (1980) designed an experimental study to compare statistical combining procedures to traditional narrative review, in which 41 researchers were randomly assigned to statistical combining or narrative group to conduct a review of the same seven studies investigating sex differences in the psychological trait of "persistence". As a result of the study, the researchers using statistical combining procedures identified more support for the hypothesis stating females are more persistent. In addition, they reported a larger effect size than did traditional reviewers. This conclusion may result from the fact that statistical combining procedures increase the power, which provides the researchers with the ability to detect even small effects and more precise results (Petticrew & Roberts, 2006).

Similarly, Oakley (2002) investigated six traditional reviews of older people and accident prevention covering 137 studies totally to examine how many primary studies were in common to all six reviews. The results were surprising: there were only 33 studies common to at least two reviews while only two studies were common to all six studies, only one of which was treated consistently in all six reviews. She also compared two reviews including totally 27 studies of anti-smoking education for young people and identified only three studies common to both reviews. Furthermore, she claimed that there were at least 70 more studies which met the inclusion criteria of the reviews in the literature.

Finally, Bushman and Wells (2001) illustrated the corrective properties of meta-analysis against biased and subjective decisions based on narrative reviews in another study conducted with 280 participants. First of all, they created 20 fictional heterogeneous research results examining the relation between similarity and attraction to be reviewed by the participants, which identified an overall positive relationship with d=0.2. Then, they manipulated the salience of the studies and the order, in which the studies were presented, in both meta-analysis and narrative review groups. Consequently, the judgments of the participants in narrative review group were affected by salient titles significantly ($p < .007$, d=0.50) while title saliency did not affect the conclusions in the meta-analysis group ($p=.71$, d=-0.07). An interesting point to be underlined was that title saliency affected memory robustly for both narrative review and meta-analysis participants, but the effect size estimates of meta-analysis participants were unaffected by salience manipulation while it was not the case for narrative reviewers. Furthermore, they concluded that meta-analysis resulted in very close estimation of effect size while narrative reviewers underestimated the strength of the effect.

To sum up, one of the most important strengths of meta-analysis is that it is immune to the limitations that traditional narrative reviews suffer from like biased and subjective judgments, and unrepresentative sampling. Furthermore, in meta-analysis, increasing number of primary studies to be synthesized results not only in increased statistical power and precision but flexibility to examine the inconsistencies in the results (if exist) while it may be chaotic and impractical for narrative reviewers because of inability of the human to handle massive amount of data reliably and validly at the same time (Borenstein et al., 2009; Glass, 2006; Glass et al., 1981; Hunter & Schmidt, 2004; Petticrew, 2003; Petticrew & Roberts, 2006; Wolf, 1986).

Summary of Advantages of Meta-Analysis over Traditional Research Synthesis Methods

Lipsey and Wilson (2001) point out four reasons why we should use meta-analysis rather than conventional research review methods to summarize and analyze a body of research studies. These reasons also constitute the primary advantages of meta-analysis. Firstly, meta-analysis procedures compel a useful discipline on the process of synthesizing research findings. Meta-analysis has prearranged steps similar to primary research studies and meta-analysts are expected to report each step followed during research synthesis explicitly so that it becomes open to scrutiny and replication. The second reason is that meta-analysis summarizes main study findings in a manner that is more effective and sophisticated than conventional reviews that are based on qualitative summaries or 'vote-counting' method relying on statistical significance, which is highly criticized as being very sensitive to sampling error mainly shaped by sample size. Third important reason to prefer meta-analysis over other reviews is that meta-analysis provides us with the capability of finding effects or relationships that are unclear in other approaches to summarizing research. Finally, meta-analysis gives us the ability of handling large amount of study findings under review in a very organized way.

In addition, Glass (1982), the eponym of the term of "meta-analysis", claims that labeling meta-analysis as "averaging effect sizes" is a misinterpretation, which is not less faulty than describing analysis of variance as "adding and multiplying". Moreover, he indicates three essential character specifications of meta-analysis. Firstly, it is quantitative, in which a set of statistical methods are employed to synthesize very large amount of data. Then, meta-analysis does not prejudge research findings in terms of research quality, which makes meta-analysis different from other approaches to research synthesis. Finally, meta-analysis seeks overall conclusions; that is, it aims to derive a meaningful generalization.

Furthermore, Rosenthal and DiMatteo (2001) emphasize that meta-analysis provides the researchers with the conclusions that are more accurate and more credible than can be achieved by any primary study or by narrative review. Then, they summarize the advantages of conducting meta-analysis as seeing the landscape of a research enterprise, keeping statistical significance in perspective, wasting no data, intimacy with data, focused research hypothesis, and identifying moderator variables.

Criticism of Meta-Analysis

In the previous part, the reasons why meta-analysis is encouraged to be used as a method of research synthesis rather than other qualitative and quantitative methods are explained by summarizing the strengths of meta-analysis stated by different researchers. However, there is also some criticism about meta-analysis in the literature, which is categorized by Glass (1982) into four groups. The first group represents the 'apples and oranges problem'. This criticism is based on the idea that meta-analysis approach to research synthesis mixes apples and oranges. It is asserted that reasonable generalizations cannot be made by comparing studies, the results of which depend on different measuring techniques, definitions of variables, and subjects since they are too unlike. However, Glass explains that there is no need to compare the studies that are the same in all respects since they would clearly provide us with very similar results within the statistical error. He emphasizes the point that "the only studies which need to be compared or integrated are different studies" (p.102). In addition, he also affirms that it is not incompatible with getting data in a primary research study from different persons and performing data analysis by lumping together since these persons are also as different as much like apple and oranges.

The second criticism is the assertion that meta-analysis method 'advocates low standards of judgment' of the quality of primary studies. That is, results from poorly-designed studies are included into the meta-analysis to be synthesized along with results from well-designed studies. Glass claims that eliminating a research study when it fails to meet the conditions based on subjective judgment may result in also unhealthy conclusions. He suggests alternative ways to overcome this problem. For example, description of design and analysis features and study of their covariance with research findings offers a way to diminish this criticism, which provides us with the capability of examining whether there are differences between sizes of the experimental effect of different modes of design issues. Furthermore, Glass examined the findings of 12 meta-analyses studies to check whether there exist a relationship between design quality and the findings of the studies. According to results, he indicates that "there is seldom much more than one-tenth standard deviation difference between average effects for high validity and low validity experiments" (p.104). On the other hand, it is evident in the literature that the opportunity of conducting moderator analysis gives the meta-analysts the chance of examining the extent to which poorly and well-designed studies differ each other in terms of effect size measures (Borenstein et al., 2009; Card, 2012; Wolf, 1986).

The third criticism is the publication bias, which is "the term for what occurs whenever the research that appears in the published literature is systematically unrepresentative of the population of completed studies" (Rothstein, Sutton, & Borenstein, 2005, p. 1). It is claimed that published research is biased in favor of statistically significant results because statistically non-significant results are rarely accepted to be published; this consequently results in biased meta-analysis results. Rosenthal (1979) called this phenomenon as 'file drawer problem' since the problem results from the fact that nonsignificant findings are banished to file drawers while significant ones are sent to be published (Rosenthal & DiMatteo, 2001). Glass, as in the previous criticism, inspected several meta-analyses and concluded that "…findings reported in journals are, on the average, one-third standard deviation more favorably disposed toward the favored hypotheses of the investigators than findings reported in theses and dissertations" (p. 106). Furthermore, Rothstein et al. (2005) assert that publication biased presents possibly the most noteworthy threat to the validity of research synthesis. However, they draw attention to two important points about this phenomenon: firstly this problem is not unique to meta-analysis but a common issue for all types of reviews or syntheses, which is stated by other researchers several times as well (Borenstein et al., 2009; Card, 2012; Rosenthal & DiMatteo, 2001; Sutton, 2009). Next, publication bias is not a problem caused by meta-analysis, or any other method of research synthesis, rather it exists as a phenomenon in the literature irrespective of whether research syntheses are conducted to summarize the results or not. Thus, the existence of publication bias in the literature should not be an argument against the research synthesis remembering that it also affects the primary studies, which draw conclusions from the literature as well (Rothstein et al., 2005; Sutton, 2009).

In fact, meta-analysis is not source of this problem but it is a part of solution since it offers several approaches for diagnosis of publication bias and to estimate the extent to which it affects the results (Glass, 1982; Sutton, 2009). Analyzing the results separately by types of publication, conducting moderator analysis or using funnel plot for diagnosis purposes are only some ways to examine publication bias in a meta-analysis study. Several methods for not only diagnosis but also adjustment purposes will be explained in detail within the scope of "publication bias" section in this article.

Finally, the fourth criticism is the 'lumpiness (non-independent data)'. That is, multiple results from the same study are often used, which may bias or invalidate the meta-analysis and make the results appear to be more reliable than they really are, since the results are not independent. For example, if a study has the effect sizes of 0.3, 0.3, 0.3 and another study has the effect sizes of 0.5, 0.5, and 0.5 in the same meta-analysis, which means that true degrees of freedom taken into account in the meta-analysis must be two, the number of studies, rather than six, the number of effect sizes. Glass (1982) proposes that a simplistic solution to this problem is to average all findings within a study. In addition, we should be careful about the journal articles based on theses or dissertations; no study should be included in the meta-analysis more than once. It is also possible to use more sophisticated ways for averaging dependent effect sizes as explained by other researchers (Gleser & Olkin, 2009; Hedges & Olkin, 1985; Marin-Martinez & Sanchez-Meca, 1999; Rosenthal & Rubin, 1986).

Similarly, Rosenthal and DiMatteo (2001) explain the criticism of meta-analysis by categorizing them into five groups. These are bias in sampling the findings, "garbage in and garbage out", singularity and non-independence of effects, overemphasis on individual effects, and combining apples and oranges. The groups, through which the criticism of meta-analysis is summarized, are similar to the ones stated by Glass (1982). Additionally, they mention that meta-analysis is criticized since it systematically assesses only individual effects between independent and dependent variables. However, they argue that before investigating the interaction of different variables, meta-analysis provides us with a clear picture of straightforward operation of each individual component. Finally, they point out that much of the criticism of meta-analysis is based on simple misunderstanding of how it is actually conducted.

How to Conduct Meta-analysis

Rosenthal and DiMatteo (2001) highlight that meta-analysis is a methodology to conduct systematic research synthesis carefully following the steps similar to the ones for primary research studies rather than being just a statistical technique. Then, they explain the basic steps of doing meta-analysis as follows:

- Define the independent and dependent variables of interest, e.g. the effects of problem based learning on students' achievement, motivation in science, and attitudes towards science.
- Collect and select the primary studies in a systematic way and read each article very carefully.
- Investigate the heterogeneity among the obtained affect sizes by means of graphs and charts or chi-square test of significance, which should be interpreted cautiously since it is, as other significance tests, dependent upon the sample size; i.e. number of studies included in the meta-analysis. In addition, the effect of relevant moderator variables on the variability among the effect sizes should be explored.
- Combine the effect sizes obtained from the primary studies using the measures of central tendency like weighted means.
- Examine the significance level of the indices of central tendency.
- Evaluate the importance of the obtained mean effect size.

Similarly, Glass (2006) also summarizes the main steps in a meta-analysis as defining problem, retrieving the literature, coding the studies, transforming findings to a common scale, and statistically analyzing the findings.

In terms of statistical models, there are two main approaches with different assumptions, which can be used within meta-analysis procedure. These are fixed-effect and random-effects models, both of which have been developed for inference about average effect size from a collection of studies (Borenstein et al., 2009; Hedges & Vevea, 1998; Hunter & Schmidt, 2000, 2004; Tweedie, Smelser, & Baltes, 2004). In the following section, these models are compared in detail with respect to different aspects.

Comparison of Fixed-Effect and Random-Effects Model

The most important assumption of fixed-effect model is that there is only one true effect size for all studies in the meta-analysis. This assumption also results in the fact that all differences in observed effects are due to only sampling error. On the other hand, the random-effects model is based on the idea that true effect size could vary from study to study because of some moderator variables like the age of participants, education level, and class size. Thus, true effect size is distributed about some mean. The effect sizes from the studies included in the meta-analysis are assumed to be a random sample of this distribution.

Since all factors that may influence the effect size are assumed to be constant in fixed-effect model, the observed effect for each study is calculated by population mean and sampling error while random-effects model assumes that true effects are distributed, which allows for inter-study variation. Thus, the observed effect for each study is calculated by adding another error resulting from between study variance in random-effects model (Borenstein et al., 2009).

In both models, to obtain more precise estimate of the summary effect (population mean for fixed-effect and overall mean for random-effects model), i.e. to minimize the variance, a weighted mean is calculated by assigning more weight to more precise studies. To decide which studies are more precise, the study variance is taken into account. In other words, more weight is assigned to the studies with less variance in both models.

Furthermore, there exists an important distinction between fixed-effect and random-effects models in terms of estimating the summary effect. Since the main purpose in fixed-effect model is to predict one true effect size, the information from small sample studies is underestimated, assigning more weight to larger sample studies. On the contrary, in random-effects model, the main goal is to

estimate the mean of distributions of effects, which results in the fact that each study, either with small or large sample, has to be represented in the summary effect. Thus, relative weights assigned under random effects become more balanced (Borenstein et al., 2009).

The amount of standard error and confidence interval constitutes another difference between two models. Since random-effects model assumes there is between-studies variance, in addition to the within-study variance, standard error and confidence interval for summary effect are expected to be always larger under random-effects model than under fixed-effect model for the meta-analysis of the same studies.

Fixed effect model is highly criticized since it underestimates sampling error resulting in narrower confidence intervals for mean effect sizes than their actual width, which also leads to overestimation of precision when basic assumptions of the model, which seem to be unrealistic for many situations, are violated (Borenstein et al., 2009; Erez, Bloom, & Wells, 1996; Hunter & Schmidt, 2000; Overton, 1998; Schmidt, Oh, & Hayes, 2009). And yet, it is evident from the literature that fixed-effect model has been much more widely used in the meta-analyses conducted until recently (Cooper, 1997; Hunter & Schmidt, 2000; National Research Council, 1992; Overton, 1998; Schmidt et al., 2009). The reason why many meta-analysts prefer fixed-effect model rather than random-effects model is that it is easier to manage (Cooper, 1997) and much simpler in terms of conceptual background and computational analysis (National Research Council, 1992).

However, while it is easy to manage, many researchers (Field, 2003; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Overton, 1998; Schmidt et al., 2009) take attention to the fact that fixed-effect model leads to escalated Type I error rates for statistical tests when homogeneity assumption is not met. Hunter and Schmidt (2000) explain how Type I error rate is affected by heterogeneity and average sample size of the studies included in the meta-analysis emphasizing that Type I error rate increases as the homogeneity assumption is violated more seriously. Counter-intuitively, the probability of doing Type I error raises dramatically as the average sample size of the studies included in the meta-analysis. As a result, for the average sample size of 100 and standard deviation of 0.25, the alpha value rises to .46, which means almost one of two meta-analyses in these conditions erroneously leads to significant results. Furthermore, increasing number of studies included in the meta-analysis does not decrease this inflated error rate. Thus, Hunter and Schmidt conclude that "FE (fixed-effect) models and procedures are rarely, if ever, appropriate for real data in meta-analyses and that therefore RE (random-effects) models and procedures should be used in preference to FE models and procedures" (p. 284), which is parallel to the recommendations of National Research Council (1992).

On the other hand, Hedges and Vevea (1998) aim to clarify the conceptual distinction between the models and argue that the most important issue in determining suitable model should be the nature of inference desired. They suggest that fixed-effect model is used to make inferences about the parameters only in the studies included in the meta-analysis while it is not suitable for unconditional inferences, i.e. inferences about the population from which the studies included in the meta-analysis are sampled, for which, random-effect model is suggested to be conducted. However, Borenstein et al. (2009) and Erez, Bloom and Wells (1996) claim that the basic assumption of fixed-effect model, which predicts only one true effect size for all studies in the meta-analysis, seems to be unrealistic for many situations. Similarly, Schmidt et al. (2009) indicate that the circumstances in which fixed-effect model would be appropriate are very limited. Thus, many researchers recommend using random-effects model rather than fixed-effect model for meta-analysis studies (Borenstein et al., 2009; Field, 2003; Hunter & Schmidt, 2000; National Research Council, 1992).

Effect Size Index

There are several indices of effect size, which can be defined as "the extent to which the phenomenon investigated is present in the study results, regardless of the sample size and the result of the statistical tests" (Sánchez-Meca & Marín-Martínez, 2010b, p. 274). Table 1 illustrates some of the common effect size indices, details of which are presented in many resources in the literature (Borenstein, 2009; Borenstein et al., 2009; Ellis, 2010; Fleiss & Berlin, 2009; Olejnik & Algina, 2000). Furthermore, Huberty (2002) provides detailed information about the history of effect size indices.

Presenting comprehensive explanation of all types of effect size goes beyond the purpose of this article. However, the essence of the concept can be explained by means of some examples. In the scope of educational research, it is quite common to compare groups on a continuous dependent variable, so Cohen's d, which is the most familiar one of the effect sizes representing groups compared on continuous outcomes, could be a good starting point for exemplifying the concept of effect size.

Cohen's d is an uncorrected standardized mean difference between two groups based on the pooled standard deviation, which can be presented as:

$$\text{Cohen's } d=\frac{X_e- X_c}{S_p}$$

where X_e is the experimental group mean, X_c is the control group mean, and S_p is the pooled standard deviation of two groups, which can be calculated by the formula:

$$S_p^2=\frac{(N_e-1)S_e^2+(N_c-1)S_c^2}{(N_e+N_c-2)}$$

where N_e is the number of subjects in experimental group, N_c is the number of subjects in control group, S_e^2 is the experimental group variance, and S_c^2 is the control group variance (Borenstein, 2009).

Finally, variance of d is given by;

$$v_d=\frac{N_e+ N_c}{N_eN_c}+\frac{d^2}{2(N_e+ N_c)}$$

Glass Δ is another uncorrected standardized mean difference between two groups based on, however, the standard deviation of control group, which can be presented as:

$$\text{Glass } \Delta=\frac{X_e-X_c}{S_c}$$

where Sc is the standard deviation of control group.

Table 1. Some of the common effect size indices

			Interpretation of effect sizes*		
			Small	Medium	Large
d family	Groups compared on dichotomous outcomes	Risk difference (RD)			
		Relative risk (RR)			
		Odds ratio (OR)			
	Groups compared on continuous outcomes	Cohen's d	0.20	0.50	0.80
		Glass delta (Δ)	0.20	0.50	0.80
		Hedge's g	0.20	0.50	0.80
		Response ratios I			
r family	Correlation indices	Pearson correlation r	0.10	0.30	0.50
		Kendall's tau (τ)			
		Phi coefficient (φ)	0.10	0.30	0.50
		Kruskal's lambda (λ)			
	Proportion of variance indices	Coefficient of determination (r^2)	0.01	0.09	0.25
		R squared (R^2)	0.02	0.13	0.26
		Cohen's f	0.10	0.25	0.40
		Eta squared (η^2)	0.01	0.06	0.14
		Epsilon squared (ε^2)			
		Omega squared (ω^2)			

*Interpretations of effect sizes are based on Cohen (1988).

That is, both Cohen's d and Glass Δ are uncorrected; i.e. biased, estimate of population effect size while only difference lies behind which standard deviation is used to standardize the mean difference. In Glass Δ, standard deviation of control group is used rather than a pooled standard deviation, which is based on the idea that control group is assumed to be more representative for population standard deviation since it is untainted by treatment effects (Ellis, 2010).

Both Cohen's d and Glass Δ have a slight bias in estimation of the population effect size especially in small samples. They slightly overestimate the parameter, which is corrected in Hedge's g by using a correction factor called as J. It can be calculated as follows:

$$J=1-\frac{3}{4d_f-1}$$

where d_f is the degrees of freedom for estimation S_{within}.

Then, g and corresponding variance (v_g) and standard error (SE_g) are given by,

$$g=J.d$$

$$v_g=J^2.v_d$$

$$SE_g=\sqrt{v_g}$$

J is always smaller than one, therefore, Hedge's g is always slightly smaller than Cohen's d, which is also correct for variance of Hedge's g comparing to Cohen's d. The difference increases with decreasing sample size (Borenstein et al., 2009).

Interpretation of effect size revealed from a research study is not an easy task, which actually depends on the context in which treatment effect is evaluated (Ellis, 2010). However, to interpret effect size values in an easier way, some threshold values are proposed by Cohen (1988), who outlines three cut-off points for small, medium and large effect sizes as 0.20, 0.50 and 0.80 respectively, which are valid for all three types of effect size indices mentioned above including Hedge's g. Although it is

simple to use these cut-off points and Cohen states that they are sufficiently grounded in logic, using Cohen's criteria to interpret the magnitude of effect size is still a controversial issue. Glass et al. (1981) speculate that "Depending on what benefits can be achieved at what cost, an effect size of 2.0 might be 'poor' and one of 0.1 might be 'good'" (p. 104). However, these cut-off points stated by Cohen are still the most-widely used criteria to interpret the effect sizes in the literature and they are suggested to be referred while interpreting the results but considering the importance of context and assessing the effect size in terms of its contribution to knowledge as well (Ellis, 2010).

Unit of Analysis

Each of primary studies included in the meta-analysis or each of effect sizes provided by these studies can be accepted as unit of analysis in a meta-analysis study. For both cases, some precautions should be taken to prevent lumpiness as a result of dependent data. Either primary studies or effect sizes are assumed to be unit of analysis, it should be checked whether each of primary studies is independent from each other; that is, in the scope of the same meta-analysis, there should be no studies sharing the same sample because of the fact that some articles published in a journal may also be included in the sample of primary studies as dissertations or theses. When each of effect sizes is accepted as unit of analysis in a meta-analysis, we also should be careful about that some studies may provide more than one effect size for the same outcome as a result of using several instruments to assess the same construct.

Validity Issues in Meta-Analysis

Publication bias and quality of primary studies constitute main concerns about the validity of a meta-analysis study (Borenstein et al., 2009; Lipsey & Wilson, 2001; Rendina-Gobioff, 2006). In the following sections, detailed explanations are provided about what 'publication bias' and 'quality of studies' mean, why they are potential threats to validity of meta-analysis.

Publication Bias

It is evident from the literature that publication bias, or "file-drawer problem", is one of the most serious problems in locating relevant studies (D. A. Bennett, Latham, Stretton, & Anderson, 2004; Borenstein et al., 2009; Rendina-Gobioff, 2006; Rothstein et al., 2005; Song, Khan, Dinnes, & Sutton, 2002; Thornton & Lee, 2000; Tweedie et al., 2004). Rothstein et al. (2005) underline that no matter how flawless in other methodological issues, the validity of the results of a meta-analysis study is threatened if the studies included in the meta-analysis is biased. The specific concern is the tendency of journals to reject the studies with negative (non-significant) results. In other words, studies with significant results are more likely to be published, which results in a bias in the published literature and then carries over to meta-analysis based on the literature (Borenstein et al., 2009). Table 2 illustrates how Rendina-Gobioff (2006) explains the impact of variance and effect size observed in a study on the likelihood of publication. As it is clearly seen in the table, statistical significance is dependent upon not only the effect size of the treatment but also the variance, which is inversely related to sample size of the study. Many researchers accepts its dependency of sample size as one of the weaknesses of statistical tests, which may result in statistically significant results although it has no practical significance (Borenstein et al., 2009; Cohen, 1990; Ellis, 2010; Gravetter & Walnau, 2007; Hunter & Schmidt, 2004; Kirk, 1996, 2001; Schmidt, 1996; Vacha-Haase, 2001). However, that is not the case which results in publication bias. What causes biased results is the fact that statistically non-significant studies tend to have small effect sizes since statistical test depend on effect size as well. That is, since the studies with non-significant results, which are more likely to have small effect sizes, are less likely to be published, any meta-analysis covering only published studies probably would indicate an overestimated mean effect size values.

Table 2. Impact of Variance and Effect Size Observed in a Study on the Likelihood of Publication (Rendina-Gobioff, 2006)

		Effect Size	
		Small	Large
Variance	Small (N=large)	Published (Statistical Significance)	Published (Statistical Significance)
	Large (N= small)	Not Published (No Statistical Significance)	Published (Statistical Significance)

* N stands for number of participants in the study

Publication bias threat is not particular to the method of meta-analysis but also a problem for narrative reviews and for any type of review method of the literature (Borenstein et al., 2009; Rosenthal & DiMatteo, 2001). Indeed, meta-analysis is not source of this problem but it is a part of solution since it provides meta-analysts with opportunity of using several methods to detect and control likely impact of publication bias as mentioned before. Forest plots, funnel plots, Rosenthal's fail-safe N (FSN), Duval and Tweedie's Trim and Fill are some of the methods that have been much cited in the literature (Duval & Tweedie, 2000a, 2000b; Egger, Smith, Schneider, & Minder, 1997; Lewis & Clarke, 2001; J. A. C. Sterne & Egger, 2001; J. A. C. Sterne & Harbord, 2004; Thornton & Lee, 2000; Tweedie et al., 2004; Yeh & D'Amico, 2004). However, it is crucial to emphasize that the most efficient way of protecting from the harmful effects of publication bias is the prevention, which is only possible by including both unpublished and published studies in the meta-analysis. Nevertheless, having unpublished studies does not guarantee the lack of publication bias, therefore, methods to diagnose and remediate the effects of biased results should be used to provide evidence that the results of the meta-analysis is sufficiently robust for additional studies with negative results.

Since each method has its own unique strengths and weaknesses, several methods should be used within the scope of meta-analysis studies for diagnosis of publication bias and to estimate the extent to which it affects the results. Considered from this angle, forest plots, funnel plots, Egger's linear regression method, Rosenthal's FSN, Orwin's FSN, and Duval and Tweedie's trim and fill method developed for diagnosis and adjustment of publication bias are clarified in the following sections.

Forest plots

Borenstein (2005) asserts that forest plot as the visual representation of the data is a key element in any meta-analysis. Figure 3 shows an example of forest plot with Hedge's g, which is a corrected standardized mean difference estimate for effect size, from 16 studies investigating the effect of problem based learning on critical thinking skills (Üstün, 2012). In Figure 3, the individual squares symbolize each study's effect size estimate and the lines extending from the squares signify the 95% confidence interval for the estimate. The area of each square corresponds to the weight that the individual study contributed to the meta-analysis. Larger squares also indicate the studies of larger samples because the larger sample size and precision is; the more weight is assigned for each study in the meta-analysis. Finally, the overall estimate from the meta-analysis and its confidence interval are represented by a diamond with extending lines put at the bottom.

While the forest plot seems to be more associated with the core of meta-analysis than with the publication bias, examining this plot is a logical first step in any analysis (Borenstein, 2005) because a forest plot not only provides the readers with information of individual studies in the meta-analysis at a glance but also summarizes overall effect with a pooled result. Furthermore, how much variation exists among studies can easily be seen by means of forest plots (Yeh & D'Amico, 2004).

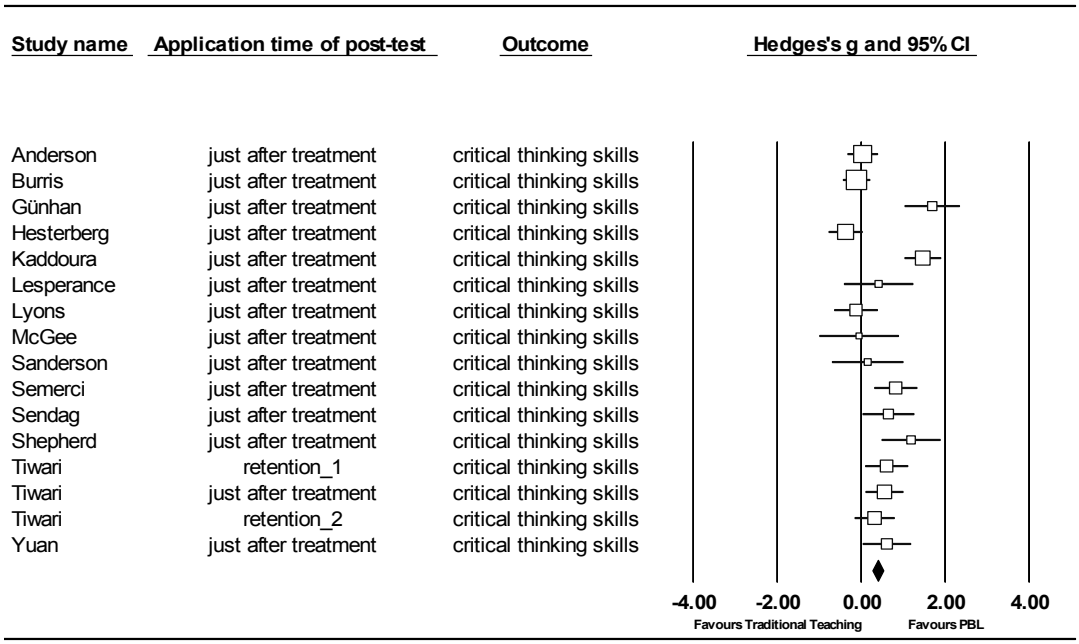

Figure 3. An Example of Forest Plot Showing Hedge's g With 95% Confidence Intervals for 16 Studies Investigating the Effect of PBL on Critical Thinking Skills

Funnel plots

Funnel plots are simple scatterplots of effect sizes estimated from each study against a measure of study sample size. Conventionally, funnel plot is constructed in such a way that X axis of the plot shows effect size values while Y axis illustrates sample size, variance or standard error. The name of "funnel plot" comes from the idea that precision in estimation of effect size of treatment increases as the sample size of component studies increases (Sterne & Harbord, 2004). Results from small sample studies will scatter widely at the bottom of the plot with smaller spread at the top as a result of larger sample studies. Thus, in the absence of any bias, the plot is expected to resemble a symmetrical inverted funnel as shown in Figure 4 (Üstün, 2012).

Conversely, if there is a publication bias, generally a skewed and asymmetrical spread is expected on the funnel plots as shown in Figure 5. In this situation, the overall effect estimated in meta-analysis overestimates the treatment's effect by resulting in an effect size of 0.38, which would be expected to be 0.09, as calculated by trim and fill method, if there would be no bias. (Üstün, 2012). However, it is highly emphasized in the literature that funnel plots should be interpreted cautiously because shape of funnel plot may be misleading for researchers and because publication bias is only one of the reasons for funnel plot asymmetry (Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006; J. A. C. Sterne & Harbord, 2004; Terrin, Schmid, & Lau, 2005). In addition, Tang and Liu (2000) claim that

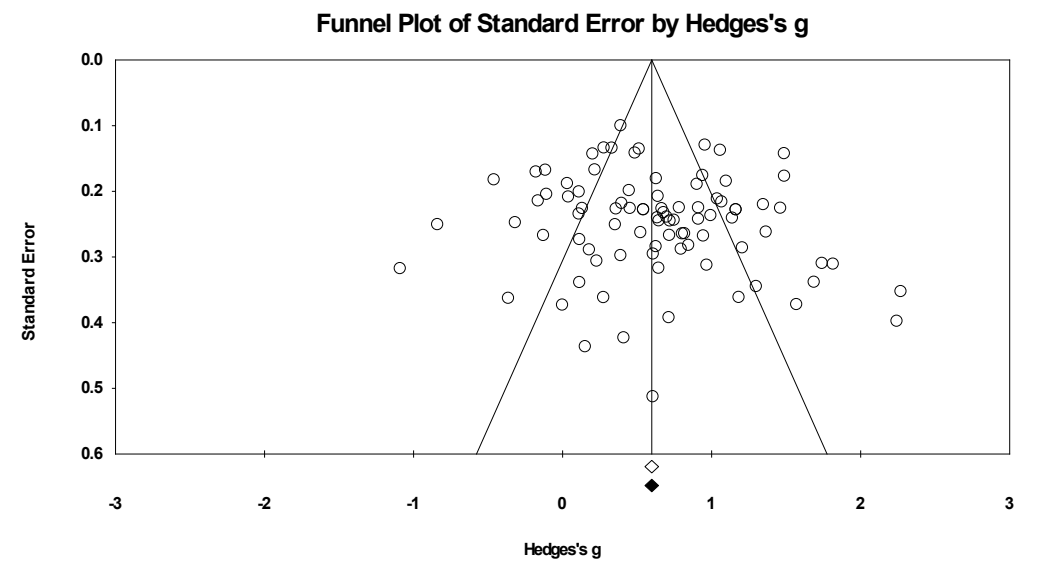

Figure 4. A Symmetrical Funnel Plot without Bias

when a different definition of precision and/or effect size measure is used, the shape of funnel plot may change significantly. They also indicate that any asymmetry of the funnel plot may result from a true heterogeneity.

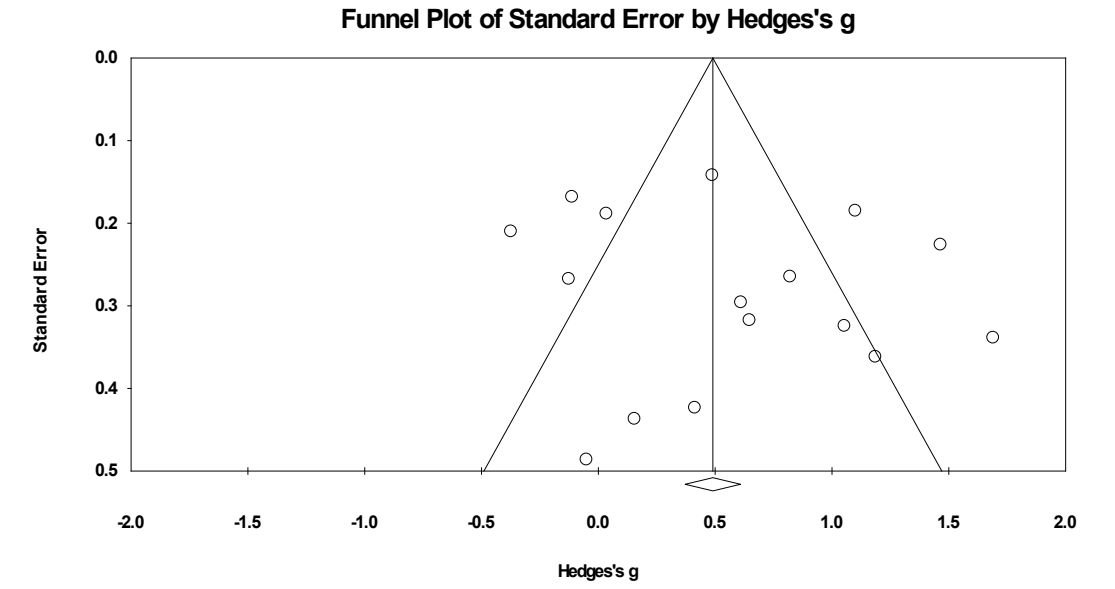

Figure 5. An Asymmetrical Funnel Plot with a Possible Bias

Egger et al. (1997) and Sterne and Harbord (2004) summarize possible sources of asymmetry in funnel plots as, selection bias (publication bias, location bias), true heterogeneity, data irregularities, artifact, that is heterogeneity due to poor choice of effect measure, and chance alone, to emphasize that funnel plot asymmetry need not result from bias.

Egger's linear regression method

Funnel plots are useful visuals to getting a sense of data about publication bias. However, it does not provide a quantitative way to detect biased results. On the other hand, Egger et al. (1997) suggest a linear regression approach to test statistically whether there exist any bias in the data included in meta-analysis. The statistical test is based on the model in which the standard normal deviate (z=effect size estimate/standard error) is regressed against its precision (*prec*= 1/standard error) (Sterne & Egger, 2005).

$$E[z] = \beta_0 + \beta_1 prec$$

For a symmetrical funnel plot, the regression line is expected to run through the origin, yielding $\beta_0 = 0$. On the other hand, if there is an asymmetry on the funnel plot, the intercept β_0 gives a measure of asymmetry. Thus, statistical test is used to check the null hypothesis of "$\beta_0 = 0$".

It is important to note that Egger's Linear Regression test still suffers from the limitations of statistical significance test. Furthermore, Borenstein (2005) highlights that the Egger test is suitable for the data which includes studies of different sample sizes and at least one of medium effect size.

Rosenthal's fail-safe N method

Fail-safe N (FSN), or "file-drawer number", suggested by Rosenthal (1979) is one of the earliest and still one of the most popular approaches in social sciences to deal with the problem of publication bias (Becker, 2005). Rosenthal's FSN can be described as the number of new studies in a meta-analysis that would be necessary to "nullify" the effect (Borenstein et al., 2009); that is, to reverse the overall probability obtained from the combined test to a value higher than the critical value for statistical significance of, usually .05 or .01 (Rosenthal, 1991). Rosenthal claims that if FSN is quite large comparing to number of observed studies, the results can be assumed to be robust to publication bias. Although there is no exact rule to decide how big N is enough to be far from publication bias, based on the Rosenthal's suggestion of rule of thumb, Mullen, Muellerleile, and Bryant (2001) propose that if N/(5k+10) (where k is the number of individual studies in the meta-analysis) exceeds 1, the result of the meta-analysis seems to be sufficiently robust for future studies.

Table 3 illustrates an example of Rosenthal's FSN calculations conducted for six studies investigating the effect of problem based learning on creativity (Üstün, 2012). The ratio of N/(5k+10) is calculated as 1.95, which indicates that the results of the meta-analysis is sufficiently tolerant for future studies although the number of the studies included in the meta-analysis is very small.

Table 3. An Example of Output for Rosenthal's FSN Calculations Conducted for Six Studies Investigating the Effect of Problem Based Learning on Creativity

Z-value for observed studies	7.29293
p-value for observed studies	0.00000
Alpha	0.05
Tails	2
Z for alpha	1.95996
Number of observed studies	6
Fail safe N	78

Orwin's fail-safe N method

Although Rosenthal's FSN provides us with a clear and quantitative way of detecting publication bias, it is criticized to be dependent on statistical significance and to assume that the mean effect sizes of missing studies is zero by default (Borenstein, 2005). Alternatively, Orwin's FSN is calculated on the basis of practical significance and allows meta-analysts to specify not only the effect size of missing studies but the specific effect size value that the overall effect would reduce with addition of missing studies as well, which would provide us with modeling a series of distributions for missing studies (Becker, 2005; Borenstein et al., 2009). Table 4 illustrates an example of Orwin's FSN calculations conducted for the same studies in the previous example for Rosenthal's FSN (Üstün, 2012). Results show that 370 additional studies with null effect are needed to bring the overall effect to the effect size value of 0.1, which is decided to be trivial. If the effect size value for the additional studies are changed from null to 0.005, the number of additional studies increases to 740. It is possible to obtain different numbers of additional studies to be needed for different specified values.

Table 4. An Example of Output for Orwin's FSN Calculations Conducted for Six Studies Investigating The Effect of Problem Based Learning on Creativity

Hedge's g in observed studies	0.62592
Criterion for a 'trivial' Hedge's g	0.10000
Mean Hedge's g in missing studies	0.00000
Fail safe N	370

Duval and Tweedie's trim and fill method.

Trim and Fill was developed by Duval and Tweedie (2000a, 2000b) to estimate the number of missing studies that may exist in meta-analysis and the effect of the missing studies on overall outcome. It is an iterative procedure in which asymmetric outlying part of the funnel plot is firstly trimmed off to calculate a theoretically unbiased estimate of effect size called as "adjusted effect size". However, this procedure affects the variance of the effects as well, resulting in a too narrow confidence interval. Thus, the trimmed studies are added back into the analysis but virtual symmetrical studies are imputed to create an unbiased sample of studies. These imputed virtual studies do not change the adjusted estimate of overall effect size (Borenstein et al., 2009; Duval, Rothstein, Sutton, & Borenstein, 2005; Duval & Tweedie, 2000a, 2000b).

Figure 6 represents an example of funnel plot in which Trim and Fill adjustment is taken into account for a similar data in Figure 5. Five imputed studies are shown as filled circles and filled diamond indicates the adjusted overall estimate. For this example, the adjusted estimate is fairly close to the null effect (Üstün, 2012).

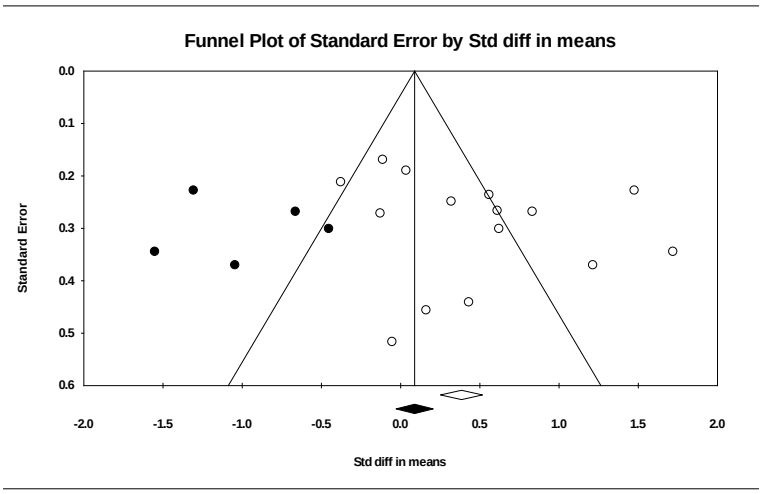

Figure 6. An Example of Funnel Plot with the Studies Imputed by TFM, Resulting in an Adjusted Effect Size

Quality of Primary Studies

Quality of primary studies is another important concern about the validity of meta-analysis results (Lipsey & Wilson, 2001; Rendina-Gobioff, 2006). However, both judgment of study quality and how to incorporate this judgment into meta-analysis cause some tensions in terms of different aspects. Firstly, the term "quality" is not easy to define since what makes a study more qualified depends on the "why the judgment is being made", which makes the construct multifaceted (Jüni, Altman, & Egger, 2001; Valentine, 2009). Difficulty in assessment of study quality as a result of multifaceted nature of the construct results in another tension, which makes researchers obtain different quality scores for the same study by using different standardized quality scales (Herbison, Hay-Smith, & Gillespie, 2006). Another issue related to the judgment of the study quality results from the interference of study quality and reporting quality (Wells & Littell, 2009). In many cases, information essential to a meta-analyst for coding the elements of study quality is not present and there is no clear procedure that meta-analyst should follow in these situations (Valentine, 2009). The final tension arises from how to use study quality in a meta-analysis. One ordinary approach to addressing study quality in a meta-analysis is simply to exclude studies with low standards (Lipsey & Wilson, 2001; Valentine, 2009). However, Glass (1982, 2006) does not agree with the idea of using study quality as one of the exclusion criteria since excluding a primary study due to quality concerns is based on subjective judgment, which may result in unhealthy conclusions. Another approach to addressing study quality in a meta-analysis is to include all available primary studies irrespective of quality concerns and then to conduct moderator or sub group analysis for study quality indicators (Littell et al.,2008).

It is evident from the literature that there exist many tools including sets of standards or criteria lists to evaluate the study quality for research syntheses (Herbison et al., 2006; Littell et al., 2008; Valentine, 2009). There are also a number of studies to review these assessment tools for study quality in the literature (Deeks et al., 2003; Herbison et al., 2006; Jüni et al., 2001; Jüni, Witschi, Bloch, & Egger, 1999; Wells & Littell, 2009). For example, Deeks et al. (2003) examine 194 tools to evaluate study quality of nonrandomized studies and conclude that none of the studies are completely suitable without revision for this aim. Similarly, Herbison et al. (2006) empirically investigate the validity of 45 scales to obtain a study quality score and they underline that "contemporary quality scores have little or no value in improving the utility of meta-analysis. Indeed, they may introduce bias, because you get different answers depending upon which quality score you use" (p. 1251). They admit that study quality is obviously important, however, they also highlight that quality scores cannot offer a solution for this situation.

As a result, it is widely-accepted in the literature that assigning a summative score based on the study quality scales should be abandoned in meta-analyses (Herbison et al., 2006; Jüni et al., 1999; Littell et al., 2008; Wells & Littell, 2009). Instead, it is suggested to examine specific dimensions of study quality by means of moderator analysis in meta-analysis studies (Herbison et al., 2006; Jüni et al., 2001; Littell et al., 2008). However, Wells and Littell (2009) claim that publication status is not a suitable indicator for study quality and also stress that reporting quality should not be confused with study quality.

Coding Reliability

Coding reliability is essential to be established in a meta-analysis since how to code the items in the coding sheet may show some variability as a result of the judgment process that the coder unavoidably applies while coding primary studies. There are two aspects of coding reliability, one of which is the consistency of coding by a single coder from study to study; i.e. "coder reliability" and the second one is the consistency between different coders; i.e. "inter-coder reliability" (Lipsey & Wilson, 2001).

As a measure of coder and inter-coder reliability 'agreement rate' (AR) can be used, which is calculated by the following formula (Orwin & Vevea, 2009):

$$\text{AR} = \frac{\text{number of observations agreed upon}}{\text{total number of observations}}$$

Orwin and Vevea (2009) provide further information about how to evaluate coding decisions.

Heterogeneity Analysis

Huedo-Medina, Sanchez-Meca, Marin-Martinez, and Botella (2006) affirm that there are three main goals of meta-analysis, which are to get an overall index about the effect size of studied relation with a confidence interval and its statistical significance, to test the heterogeneity of the effect sizes and to identify possible moderator variables that affect the results if there exists heterogeneity among the effect sizes obtained from the primary studies. That is, testing heterogeneity is one of the major aims of meta-analysis not only because it indicates the existence of moderator variables but also as it is one of the assumptions lies behind the idea of random-effects model.

The difficulty to identify the heterogeneity between true effect sizes, which mean the effect sizes in the underlying populations, arises from the fact that we try to estimate true heterogeneity by means of observed variance, which covers random error as well (Borenstein et al., 2009). In other words, there are two sources of variability, which are sampling error, also named as within-study variability and between-studies variability. The former is always present in the meta-analyses while the latter only exists when there is true heterogeneity between the population effect sizes estimated by observed ones (Huedo-Medina et al., 2006). It is the between-studies heterogeneity that we want to quantify but excluding sampling error.

There are different ways of identifying and quantifying the heterogeneity in meta-analysis. The advantages and shortcomings of Q statistic and its corresponding chi-square significance test, which is the usual way of assessing heterogeneity, are presented in the next section while alternatives like I^2 and τ^2 are explained briefly in the following sections.

Q Statistic and Corresponding Chi-squared Significance Test

Q statistic is simply the weighted sum of squares, in which deviations from mean effect size are weighted by the inverse-variance. Thus, it provides a measure of total variance including within-study variance. True heterogeneity is estimated by excluding df, which is k-1 (k is the number of studies), from Q statistic. However, it should be noted that it is not a mean but sum of deviations, thus it is not an intuitive measure. Therefore, Q statistic is used to test the null hypothesis that all studies share a common effect size by means of chi-squared distribution (Borenstein et al., 2009).

Yet, the test of significance still shares the limitations of any other statistical significance test, being highly dependent on sample size, which also refers number of studies in a meta-analysis. Huedo-Medina et al. (2006) claim that Q-test, which is the statistical test using Q statistic, suffers from low power when number of studies and/or average sample size is low in a meta-analysis. They also emphasize that Q test only indicates the presence or absence of heterogeneity while it does not quantify the extent of such heterogeneity.

Estimation of τ^2

τ^2 is a parameter, which refers to the variance of the true effect size. In a meta-analysis, τ^2 is estimated by the variance of observed effect sizes, denoted by T^2. This estimate depends on the value of (Q-df), but differently it quantifies the extent of true variation by providing an absolute value in the same metric as the effect size (Borenstein et al., 2009). Furthermore, its square root gives an estimate of tau the standard deviation of population.

Both τ^2 and τ are informative to provide the extent to which the effect sizes are heterogeneous, which cannot be inferred from Q statistic directly. However, the discussion about magnitude of τ^2 is dependent upon type of effect size estimate since τ^2 quantifies the between study variance in the same metric with effect size estimate. To exemplify how τ^2 can be used to quantify the heterogeneity of true effect sizes, let's assume a distribution of Hedge's g estimate of 0.566 with a given τ^2 of 0.212, which results in a τ value of 0.461. In this example, it is easy to reach a conclusion that 95% of cases, the true effect size in a new study would fall inside in the interval of -0.416 to 1.548 by appropriate calculations, the details of which is available on Borenstein et al. (2009).

The I^2 Statistic

Another way of quantifying heterogeneity is to establish I^2 statistic, which is the ratio of true variance to total variance across the observed effect sizes. Although I^2 also depends on Q statistic, it provides us with a measure of heterogeneity in a more intuitive scale than Q statistic does. Unlike τ^2 and τ, which present absolute measures on the same scale as the effect size index, I^2 statistic offers a ratio on relative scale, which is not dependent on the effect size scale.

Higgins, Thompson, Deeks, and Altman (2003) summarize some of the advantages of using I^2 as a measure of heterogeneity in meta-analyses as follows:

- Its interpretation is intuitive since it provides a ratio
- It is simple to calculate
- It does not inherently depend on sample size
- It is possible to be interpreted similarly irrespective of effect size scale.

They also suggest cut-off points for low, moderate and high level of heterogeneity as 25%, 50% and 75% respectively and claim that I^2 is preferable as a measure of heterogeneity in meta-analysis.

However, Borenstein et al. (2009) underline that I^2 reflects only a proportion of between study variance to total variance and does not provide an absolute value of true variance. Thus, a significant amount of true variance can be easily masked by high amount of random error as a result of poor precision; i.e. wide confidence intervals. They also suggest that both a measure of the magnitude of heterogeneity, which can be indicated by T^2 as an estimate of τ^2 or I^2, and a measure of uncertainty, which can be presented by Q-test or confidence intervals for T^2 or I^2, should be reported for an informative presentation of true heterogeneity.

Moderator Analysis

One of the major aims of conducting a meta-analysis is to analyze the variation among the effect sizes obtained from primary studies included in the meta-analysis by comparing the mean effect for different subgroups of studies (Borenstein et al., 2009; Huedo-Medina et al., 2006). However, analysis of variance (ANOVA), which is used to compare subgroups in primary studies, is not applied directly in a meta-analysis since effect sizes revealed from each of the primary studies take the place of individual scores of participants in a primary study. Thus, an analog to ANOVA based on Q-test is conducted in meta-analyses as statistical test to compare subgroups.

Analog to ANOVA test can be conducted on the basis of different models including fixed-effect, random-effects (also called as fully random-effects) and mixed-effects model, each of which has different assumptions about the variation of effect sizes within subgroups and variability of subgroups. In the "within subgroups" level, the difference between fixed-effect and random-effects model is the same with ones used to calculate the overall effect size. That is, fixed-effect model assumes that there is only one true effect size representing one true population and the variation of effect sizes within the subgroups results from simply sampling error. However, random-effects model allow different true effect sizes representing different populations, dividing total variance into two components, which are between and within study variances. On the other hand, in the "between subgroups" level, fixed and random refers to different meanings. Fixed means that subgroups are fixed or the same for any researchers who would perform similar analysis. For example, the subgroups of moderator variable of gender can be assigned as fixed while the country variable can be assigned as random at the between subgroups level to be able to make generalization to other countries not included in the subgroups of a specific meta-analysis (Borenstein et al., 2009).

Fixed-effect model for moderator analysis assumes only one true effect size within subgroups and fixed subgroup categories at the between subgroup level while random-effects model uses random variability at both levels. There is also another model called as mixed-effect model, which uses random-effects models within subgroups but assumes fixed subgroups categories.

Finally, Lipsey and Wilson (2001) assert that the ANOVA analog should be conducted to test a limited number of priori hypotheses regarding moderator variables. They underline that it is a common but incorrect application that a vast number of categorical variables are tested by analog to ANOVA, which inflates Type I error rates.

It should also be noted that meta-regression, which is similar to regression or multiple regression conducted within primary studies, provides the meta-analysts with another option to run moderator analysis in the scope of meta-analysis. The only difference between the logic of regression and meta-regression lies behind the idea that covariates are defined at the level of primary studies rather than the level of subject participated in the primary studies. However, unlike regression, each primary study should be weighted while conducting meta-regression. Thus, it is not possible to use standard regression models constructed by statistical packages for general purposes like SPSS or STATA without appropriate macros.

The Proportion of Variance Explained

Analog to ANOVA test suffers from the weaknesses inherent to statistical significance test. In addition, it is evident from the literature that significance tests conducted for moderator analysis generally have low statistical powers (Borenstein et al., 2009; Pigott, 2012). Thus, non-significant results from this significance test should be interpreted in caution. Finally, this test only checks whether the difference between mean effect sizes of subgroups is statistically significant but does not quantify the magnitude of difference.

In primary studies, R^2, which is an index defined as the ratio of explained variance to total variance, is used to quantify the impact of covariate on the dependent variable. However, it cannot be directly applied in meta-analysis due to within study variance, which is impossible to be excluded

completely. Thus, R^2 is redefined in meta-analysis in a way that it only focuses on true variance, which is τ^2. That is, R^2 is redefined as the proportion of true variance, rather than total variance, explained by the covariate (Borenstein et al., 2009). The index can be calculated as;

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}}$$

where $T^2_{within}$is the pooled variance across subgroups, which is given by;

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}}$$

where C is a scaling factor, which is provided by Comprehensive Meta-Analysis (CMA) or can be calculated by;

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

where W is the weight of each study (Ellis, 2010). T^2_{within}can result in a negative value due to sampling issues, then it should be set to zero (Borenstein et al., 2009).

The index of R^2 is not provided by CMA but it can be calculated by these formulas using C given by CMA. Cohen (1988) suggests thresholds points of .02, .13 and .26 as small, medium and large respectively for R^2 index.

Power Analysis

Statistical power describes "the probability that a test will correctly identify a genuine effect. Technically, the power of a test is defined as the probability that it will reject a false null hypothesis" (Ellis, 2010, p. 52). There are four factors affecting statistical power in a primary study, which are the magnitude of effect size, the alpha level set by the researcher, the number of tails; i.e. one-tailed or two-tailed test, and finally the sample size (Gravetter & Walnau, 2007). Direction of effects can be summarized as statistical power increases with increasing treatment effect and increasing precision of study, which is exactly true for statistical power of meta-analyses as well (Borenstein et al., 2009). Thus, it is not surprising that statistical power of a meta-analysis under fixed-effect model is always higher than the power of each primary study included in the meta-analysis. It can be easily predicted from confidence interval of mean effect size, which is always narrower than the ones for primary studies in a fixed-effect model, indicating very high precision as a result of substantial sample size.

However, the situation is quite different for the meta-analyses using random-effects model, in which, as explained previously, there are two sources of error. Between-study variance, which is an indicator of heterogeneity, affects statistical power as well; therefore, it is possible for a meta-analysis to have a lower power than primary studies in a random-effects model.

Power analysis for the statistical tests conducted for main effect is very similar to the ones for primary studies. The only difference results from the calculations of the variance of mean effect size, which increases with increasing heterogeneity in random-effects model. Once the variance is calculated, the parameter lambda (λ) can be calculated as follows:

$$\lambda = \frac{\delta}{\sqrt{V_\delta}}$$

where δ is the true effect size and V_δ is corresponding variance. Then, power is given by:

$$\text{Power} = 1 - \Phi(c_\alpha - \lambda) + \Phi(-c_\alpha - \lambda)$$

where c_α is the critical value of Z associated with significance level α, which is 1.96 for α of 0.05. $\Phi(x)$ can be calculated in EXCEL by using NORMSDIST function (Pigott, 2012).

Software for Statistical Analyses

Statistical packages designed for general purposes such as SPSS, SAS, STATA and R have no inbuilt support for meta-analysis. It is not easy to assign weights as required especially for random-effects model in any of these software packages and in the case of subgroup analysis (analog to ANOVA) and meta-regression, they produce incorrect p-values because of different rules for assigning degrees of freedom in meta-analysis (Borenstein et al., 2009). However, by means of some macros developed for these statistical packages, it is possible to conduct meta-analysis with specific limitations and strengths.

On the other hand, there are some statistical software developed specifically for meta-analysis like Comprehensive Meta-Analysis (CMA), RevMan and MIX. Bax, Yu, Ikeda, and Moons (2007) compares six statistical programs dedicated to meta-analysis and conclude that the most appropriate meta-analysis software may change for each user depending on his or her expectations while indicating that CMA is the most versatile software and MIX and CMA are the most user-friendly ones designed for meta-analysis. CMA is commercial software, which allows running many statistical analyses including the ones to calculate main effects in both fixed-effect and random-effects models, to perform subgroup analyses and meta-regression besides different types of heterogeneity and publication bias analyses. It is also possible to create forest and funnel plots and to make some changes on these graphs. Another important advantage of CMA is that it provides the researcher with 100 different formats for data entry. Still, neither CMA nor MIX gives the researcher the opportunity of studying with multilevel and Bayesian models or conducting multivariate analysis. Finally, Wallace, Schmid, Lau, and Trikalinos (2009) also present a detailed comparison of meta-analysis software including Stata and R with their macros, MIX, CMA, RevMan and Meta-Analyst.

How to Report Meta-analysis

Up to this point, we have attempted to construct conceptual framework of meta-analysis as a research methodology for research synthesis and provide explanations about how to conduct meta-analysis. However, how to perform a research study is only one facet of a scholar work, the other of which covers reporting issues. Since what other researchers know about a research study is commonly limited by what is reported in the related article or thesis, reporting quality has a vital significance for any kind of scholar work including meta-analyses (Clarke, 2009).

Within this respect, Ahn, Ames, and Myers (2012) take attention to the fact that meta-analyses conducted in education have several methodological weaknesses especially in data evaluation and analysis. Within the scope of this study, they investigate 56 meta-analyses published recently in the field of education and underline that both research and reporting quality should be improved in meta-analysis studies especially in the use of statistical methods. Thus, it is highly suggested for every meta-analyst that appropriate one(s) of these checklists should be studied carefully before starting to conduct a meta-analysis so that they can design their meta-analysis as well as possible.

From this point of view, it is evident from the literature that there exists considerable effort to construct reporting standards for meta-analysis including "The Quality of Reporting of Meta-analysis" (QUOROM) statement (Moher et al., 1999) and its revised version, "Preferred Reporting Items for Systematic Reviews and Meta-analyses" (PRISMA) (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009), and "Meta-analysis of Observational Studies in Epidemiology" (MOOSE) (Stroup et al., 2000). Moreover, APA Publication and Communications Board Working Group on Journal Article Reporting Standards has published another reporting standard for meta-analysis, called as "Meta-analysis Reporting Standards" (MARS) (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). MARS, which is available as an online document on the website indicated on the related reference, explains in detail what are expected to be reported in each part of meta-analysis explicitly including title, abstract, introduction, method covering inclusion and exclusion criteria, moderator analysis, search strategies, coding

procedures, and statistical methods, and results and discussion sections. MARS seems to be most appropriate one for meta-analysis of primary studies in the field of education research.

Conclusion

In this article, we have attempted to highlight the essential role of research synthesis in scientific enterprise and to clarify the concept of meta-analysis by discussing its position in the framework of research synthesis. Then, we have explored meta-analysis in different aspects covering its strengths and criticisms, and partially methodological and statistical foundations. Finally, the standards developed about how to report meta-analyses have been summarized while the essentiality of following these types of standards for meta-analysts have been emphasized to construct better-designed meta-analyses and to report them in a such a way that all readers could understand of the details of the study. As an effective method of research synthesis, the number of meta-analysis studies has proliferated by recent years in many disciplines including educational researches. However, remarkable number of them are disadvantaged by methodological, statistical and reporting weaknesses (Ahn et al., 2012). It is for this reason that both research and reporting quality should be improved in meta-analysis studies especially in statistical method parts. In this respect, we hope that this article will provide meta-analysts with a general framework and guidance while they are conducting and reporting their meta-analysis.

Acknowledgments

Special thanks to Assist. Prof. Dr. Ertuğrul Özdemir and Assist. Prof. Dr. Süleyman Davut Göker for their valuable feedbacks about this article.

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82*(4), 436-476. doi: 10.3102/0034654312458162

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*(9), 839-851. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2957094/

Bax, L., Yu, L. M., Ikeda, N., & Moons, K. G. (2007). A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology, 7*(1), 40.

Becker, B. J. (2005). Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex, England: John Wiley & Sons, Ltd.

Bennett, D. A., Latham, N. K., Stretton, C., & Anderson, C. S. (2004). Capture-recapture is a potentially useful method for assessing publication bias. *Journal of Clinical Epidemiology, 57*(4), 349-357.

Bennett, J. (2005). Systematic reviews of research in science education: Rigour or rigidity? *International Journal of Science Education, 27*(4), 387-406.

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18.

Berman, N., & Parker, R. (2002). Meta-analysis: Neither quick nor easy. *BMC Medical Research Methodology, 2*(1), 10.

Bligh, J. (2000). Problem-based learning: The story continues to unfold. *Medical Education, 34*(9), 688-689.

Borenstein, M. (2005). Software for Publication Bias. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication Bias for Meta-Analysis: Prevention, Assessment and Adjustments*. West Sussex, England: John Wiley & Sons Ltd.

Borenstein, M. (2009). Effect size for continuous data. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons, Ltd.

Bushman, B. J., & Wang, M. C. (2009). Vote-counting procedures in meta-analysis. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 207-220). New York: Russell Sage Foundation

Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personality and Social Psychology Bulletin, 27*(9), 1123-1130.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: The Guilford Press.

Carlton, P. L., & Strawderman, W. E. (1996). Evaluating cumulated research I: The inadequacy of traditional methods. *Biological Psychiatry, 39*(1), 65-72.

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

Chan, M. L. E., & Arvey, R. D. (2012). Meta-analysis and the development of knowledge. *Perspectives on Psychological Science, 7*(1), 79-92.

Clarke, M. (2009). Reporting format. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2 ed., pp. 521-534). New York: Russell Sage Foundation.

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

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

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*(3), 243-253.

Cooper, H. (1997). Some finer points in the meta-analysis. In M. Hunt (Ed.), *How science takes stock: The story of meta-analysis*. New York: Russell Sage Foundation.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3-16). New York: Russell Sage Foundation

Cooper, H., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87*(3), 442.

Dalton, D. R., & Dalton, C. M. (2008). Meta-analyses. *Organizational Research Methods, 11*(1), 127-147.

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education, 26*(3-4), 365-378.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., . . . Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27).

Duval, S., Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). The Trim and Fill Method *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. West Sussex, England: John Wiley & Sons, Ltd.

Duval, S., & Tweedie, R. (2000a). A nonparametric" trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology, 49*(2), 275-306.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33*(5), 517.

Eysenck, H. J. (1984). Meta-analysis: An abuse of research integration. *The Journal of Special Education, 18*(1), 41-59.

Eysenck, H. J. (1994). Systematic reviews: Meta-analysis and its problems. *British Medical Journal, 309*, 789-792.

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

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century *Journal of Clinical Epidemiology, 48*(1), 71-79.

Field, A. P. (2003). The problem in using fixed-effects models of meta-analysis on real world data. *Understanding Statistics, 2*, 77-96.

Fitz-Gibbon, C. T. (1985). The implications of meta-analysis for educational research. *British Educational Research Journal, 11*(1), 45-49.

Fitzgerald, S. M., & Rumrill, P. D. (2003). Meta-analysis as a tool for understanding existing research literature. *Work: A Journal of Prevention, Assessment and Rehabilitation, 21*(1), 97-103.

Fitzgerald, S. M., & Rumrill, P. D. (2005). Quantitative alternatives to narrative reviews for understanding existing research literature. *Work: A Journal of Prevention, Assessment and Rehabilitation, 24*(3), 317-323.

Fleiss, J. L., & Berlin, J. A. (2009). Effect size for dichotomous data. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8.

Glass, G. V. (1982). Meta-analysis: An approach to the synthesis of research results. *Journal of Research in Science Teaching, 19*(2), 93-112.

Glass, G. V. (2006). Meta-analysis: The quantitative synthesis of research findings. In J. L. Green, P. B. Elmore & G. Camilli (Eds.), *Handbook of Complementary Methods in Education Research*. Mahwah: Lawrence Erlbaum Associates.

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

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

Gliner, J. A., Morgan, G. A., & Harmon, R. J. (2003). Meta-analysis: Formulation and interpretation. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*(11), 1376.

Gravetter, F. J., & Walnau, L. B. (2007). *Statistics for behavioral sciences*. Belmont, CA: Thomson Learning, Inc.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational and Behavioral Statistics, 17*(4), 279-296.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88*(2), 359-369.

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Herbison, P., Hay-Smith, J., & Gillespie, W. J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology, 59*, 1249-1256.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557-560.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*(2), 227.

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-anlaysis: Q statistic or I^2 index? *Psychological Methods, 11*(2), 193-206.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. NY: The Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. Random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*(4), 275-292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2 ed.). California: Sage Publications.

Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. *British Medical Journal, 323*, 42-46.

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association, 282*, 1054-1060.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*(2), 213-218.

Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., & Olkin, I. (2006). Evidence based medicine: The case of the misleading funnel plot. *BMJ: British Medical Journal, 333*(7568), 597.

Lewis, S., & Clarke, M. (2001). Forest plots: trying to see the wood and the trees. *British Medical Journal, 322*(7300), 1479.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. California: Sage Publications.

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

Lundahl, B., & Yaffe, J. (2007). Use of meta-analysis in social work and allied disciplines. *Journal of Social Service Research, 33*(3), 1-11.

Marin-Martinez, F., & Sanchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology, 2*(1), 32-38.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D., & The QUOROM group. (1999). Improving the quality of reporting of meta-analysis of randomized controlled trials: The QUOROM statement. *Lancet, 354*(9193), 1896-1900.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *Annals of Internal Medicine, 151*(4), 264-269.

Mullen, B., Muellerleile, P., & Bryant, B. (2001). Cumulative meta-analysis: a consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin, 27*(11), 1450.

Mulrow, C. D. (1994). Systematic reviews: Rationale for systematic reviews. *British Medical Journal, 309*, 597-599.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research* Washington, DC: National Academy Press.

Normand, S. L. T. (1999). Tutorial in biostatistics meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine, 18*(3), 321-359.

O'Rourke, K. (2007). An historical perspective on meta-analysis: Dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine, 100*(12), 579-582.

Oakley, A. (2002). Social science and evidence-based everything: The case of education. *Educational Review, 54*(3), 277-286.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations and limitations. *Contemporary Educational Psychology, 25*(3), 241-286.

Orwin, R. G., & Vevea, J. L. (2009). Evaluating Coding Decisions. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*(3), 354-379.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal, 3*, 1243-1246.

Petticrew, M. (2003). Why certain systematic reviews reach uncertain conclusions. *British Medical Journal, 326*(7392), 756-758.

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

Pigott, T. D. (2012). *Advances in meta-analysis*. NY: Springer Verlag.

Rendina-Gobioff, G. (2006). *Detecting publication bias in random-effects meta-analysis: An emprical comparison of statistical methods* Unpublished doctoral dissertation. University of South Florida. Florida.

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

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* (Vol. 6). CA: Sage Publication.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*(1), 59-82.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3*, 377-386.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*, 400-406.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex, England: John Wiley & Sons.

Sánchez-Meca, J., & Marín-Martínez, F. (1998). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical and Statistical Psychology, 51*(2), 311-326.

Sánchez-Meca, J., & Marín-Martínez, F. (2010a). Meta-analysis in psychological research. *International Journal of Psychological Research, 3*(1), 150-162.

Sánchez-Meca, J., & Marín-Martínez, F. (2010b). Meta Analysis. In P. Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education* (Vol. 7, pp. 274-282). Oxford: Elsevier.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*(10), 1173-1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115-129.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*(5), 529-540.

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed- versus random effects models in meta-analysis: Model properties and an emprical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62*, 97-128.

Schulze, R. (2007). The state and the art of meta-analysis. *Zeitschrift für Psychologie/Journal of Psychology, 215*(2), 87-89.

Shapiro, S. (1994). Meta-analysis/Shmeta-analysis. *American Journal of Epidemiology, 140*(9), 771-778.

Shelby, L. B., & Vaske, J. J. (2008). Understanding meta-analysis: A review of the methodological literature. *Leisure Sciences, 30*(2), 96-110.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752-760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Song, F., Khan, K. S., Dinnes, J., & Sutton, A. J. (2002). Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *International Journal of Epidemiology, 31*(1), 88.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54*(10), 1046-1055.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex, England: John Wiley & Sons, Ltd.

Sterne, J. A. C., & Harbord, R. M. (2004). Funnel plots in meta-analysis. *The Stata Journal, 4*(2), 127-141.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., . . . Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *The Journal of the American Medical Association, 283*(15), 2008-2012.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435-452). New York: Russell Sage Foundation.

Tang, J. L., & Liu, J. L. Y. (2000). Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology, 53*(5), 477-484.

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology, 58*(9), 894-901.

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis its causes and consequences. *Journal of Clinical Epidemiology, 53*(2), 207-216.

Torgerson, C. (2003). *Systematic reviews*. London: Continuum International Publishing Group.

Tweedie, R. L., Smelser, N. J., & Baltes, P. B. (2004). Meta-analysis: Overview *International Encyclopedia of the Social & Behavioral Sciences.* (pp. 9717-9724): Elsevier Science Ltd.

Üstün, U. (2012). *To what extent is problem-based learning effective as compared to traditional teaching in science education? A meta-analysis study*. Unpublished doctoral dissertation. METU. Ankara.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61*(2), 219-224.

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Wallace, B. C., Schmid, C. H., Lau, J., & Trikalinos, T. A. (2009). Meta-Analyst: Software for meta-analysis of binary, continuous and diagnostic data. *BMC Medical Research Methodology, 9*(80). doi: 10.1186/1471-2288-9-80

Wells, K., & Littell, J. H. (2009). Study quality assessment in systematic reviews of research on intervention effects. *Research on Social Work Practice, 19*(1), 52-62.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. California: Sage Publications Inc.

Yeh, J., & D'Amico, F. (2004). Forest plots: data summaries at a glance. *The Journal of Family Practice, 53*, 1007.