TO WHAT EXTENT IS PROBLEM-BASED LEARNING EFFECTIVE AS COMPARED TO TRADITIONAL TEACHING IN SCIENCE EDUCATION? A META-ANALYSIS STUDY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ULAŞ ÜSTÜN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
SECONDARY SCIENCE AND MATHEMATICS EDUCATION

SEPTEMBER 2012

Approval of the thesis:

TO WHAT EXTENT IS PROBLEM-BASED LEARNING EFFECTIVE AS COMPARED TO TRADITIONAL TEACHING IN SCIENCE EDUCATION? A META-ANALYSIS STUDY

submitted by **ULAŞ ÜSTÜN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Secondary Science and Mathematics Education Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Ömer Geban _____
Head of Department, **Secondary Science and Mathematics Edu.**

Assoc. Prof. Dr. Ali Eryılmaz _____
Supervisor, **Secondary Science and Mathematics Edu. Dept., METU**

Examining Committee Members:

Prof. Dr. Bilal Güneş _____
Secondary Science and Mathematics Education Dept., Gazi University

Assoc. Prof. Dr. Ali Eryılmaz _____
Secondary Science and Mathematics Education Dept., METU

Assoc. Prof. Dr. Esen Uzuntiryaki _____
Secondary Science and Mathematics Education Dept., METU

Assist. Prof. Dr. Yeşim Çapa Aydın _____
Educational Sciences Dept., METU

Assist. Prof. Dr. Ömer Faruk Özdemir _____
Secondary Science and Mathematics Education Dept., METU

Date: 12.09.2012

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Ulaş Üstün

Signature :

ABSTRACT

TO WHAT EXTENT IS PROBLEM-BASED LEARNING EFFECTIVE AS COMPARED TO TRADITIONAL TEACHING IN SCIENCE EDUCATION? A META-ANALYSIS STUDY

Üstün, Ulaş

Ph.D., Department of Secondary Science and Mathematics Education

Supervisor: Assoc. Prof. Dr. Ali Eryılmaz

September 2012, 274 pages

The main purpose of this meta-analysis was to investigate the effectiveness of PBL not only on student achievement and motivation in science, but also on attitudes towards science and skills in primary, secondary and higher educational levels. In addition, the effects of some moderator variables including publication type, research design, teacher effect, researcher effect, country, subject matter, school level, PBL mode, length of treatment, group size, type of questions and assessment instrument on the effectiveness of PBL were also examined in the scope of this meta-analysis. 147 effect sizes were revealed from 88 primary studies selected to be included in the meta-analysis based on the inclusion criteria. Random-effects model rather than fixed-effect model was chosen to be conducted to compute effect sizes indicating the effect of PBL on different outcomes while mixed-effect and fully random-effects model were used while performing analog ANOVA for moderator analysis. The results clearly show that PBL is more effective on different outcomes when compared to traditional teaching methods. The results indicate an overall medium mean effect size of 0.633 for PBL effectiveness. More specifically, PBL

has a large impact with a large effect size of 0.820 on students' achievement in science subjects in different levels and reveals medium effect sizes of 0.566, 0.616, and 0.565 for students' attitude towards science, motivation in science and different kinds of skills, respectively. Moderator analyses indicate that publication type, country, subject area, school level and length of treatment have a noteworthy impact on the effectiveness of PBL.

Keywords: Problem Based Learning, Meta-Analysis, Achievement and Motivation in Science, Attitude towards Science, Science Education

ÖZ

PROBLEME DAYALI ÖĞRENME GELENEKSEL ÖĞRETİM YÖNTEMİNE KIYASLA FEN EĞİTİMİNDE NE DERECE ETKİLİDİR? BİR META-ANALİZ ÇALIŞMASI

Üstün, Ulaş

Doktora, Ortaöğretim Fen ve Matematik Alanları Eğitimi Bölümü

Tez Yöneticisi: Doç. Dr. Ali Eryılmaz

Eylül 2012, 274 sayfa

Bu meta-analiz çalışmasının temel amacı, PDÖ'nün sadece öğrencilerin fendeki başarı ve motivasyonlarına değil aynı zamanda fene karşı tutumlarına ve becerilerine olan etkisini ilköğretim, ortaöğretim ve yüksek öğrenim düzeyinde araştırmaktır. Ayrıca, yayın türü, araştırma dizaynı, öğretmen etkisi, araştırmacı etkisi, ülke, konu alanı, okul düzeyi, PDÖ'nün kapsamı, uygulamanın süresi, grup büyüklüğü, ölçmede kullanılan soru çeşitleri ve ölçme aracının çeşidi gibi ara değişkenlerin PDÖ'nün etkinliği üzerindeki etkisi de bu meta-analiz çalışması kapsamında araştırılmaktadır. Dâhil edilme kriterleri temel alınarak seçilen 88 birincil çalışmadan 147 etki büyüklüğü elde edilmiştir. PDÖ'nün farklı öğrenme ürünleri üzerindeki etkisini gösteren etki büyüklüklerini hesaplayabilmek için sabit etki modeli yerine rastgele etki modeli kullanılmış, ara değişken analizi sırasında analog ANOVA yapılırken ise birleşik etki modeli veya tamamen rastgele etki modelinden yararlanılmıştır. Sonuçlar PDÖ'nün geleneksel yöntemlere göre daha etkili olduğunu açıkça göstermektedir. Rastgele etki modeli kullanılarak PDÖ'nün farklı değişkenler üzerindeki genel verimliliği için etki büyüklüğü 0.633 olarak

hesaplanmıştır. PDÖ'nün öğrencilerin fendeki başarıları üzerindeki etkisini gösteren etki büyüklüğü, büyük etki büyüklüğü kabul edilen 0.820 olarak hesaplanılırken, öğrencilerin fene karşı tutumları, fendeki motivasyonları ve becerileri üzerindeki etkisini gösteren etki büyüklükleri ise sırasıyla 0.566, 0.616 ve 0.565 olarak bulunmuştur. Ara değişken analizleri sonucunda ise yayın türleri, ülke, konu alanı, okul seviyesi ve uygulama süresi değişkenlerinin PDÖ'nün verimliliği üzerinde önemli bir etkisi olduğu görülmüştür.

Anahtar Kelimeler: Probleme Dayalı Öğrenme, Meta-analiz, Fen Başarısı, Fene Karşı Tutum ve Motivasyon, Fen Eğitimi

To three valuable women, without whom, nothing in my life would be as beautiful as it is now;
my late mom, my wife and my beautiful daughter…

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere appreciation to my supervisor Assoc. Prof. Dr. Ali Eryılmaz for his guidance, support and patience throughout all my doctorate education. It is hard to explain how much I have learnt from him during my higher education starting from undergraduate years.

I would like to extend my appreciation to my committee members, Assoc. Prof. Dr. Esen Uzuntiryaki and Assist. Prof. Dr. Yeşim Çapa for their supportive guidance. I am also grateful to my friends, Cezmi Ünal, Haki Peşman, Demet Kırbulut, Ayla Çetin Dindar, H. Özge Arslan, M. Şahin Bülbül, Özlem Ateş and Kübra Eryurt for their support during my doctorate education.

I also would like to thank The Scientific and Technological Research Council of Turkey (TUBITAK), which has provided me with partial financial support during my doctorate education.

Finally, I wish to express my deepest gratitude to my wife, Neslihan Üstün, who was extremely patient while listening to all my explanations about the details of my dissertation process, and my little daughter, Ceylin Üstün, who always provided me with the extra energy I needed while studying for long hours.

Thank you all…

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PBL: Problem-based Learning

PDÖ: Probleme Dayalı Öğrenme

CMA: Comprehensive Meta-analysis

ANOVA: Analysis of Variance

TFM: Trim and Fill Method

AR: Agreement Rate

FSN: Fail-safe N

NBME: National Board of Medical Examiners

SCI-EXPANDED: Science Citation Index Expanded

SSCI: Social Sciences Citation Index

A&HCI: Arts and Humanities Citation Index

CPCI-S: Conference Proceedings Citation Index in Science

CPCI-SSH: Conference Proceedings Citation Index in Social Sciences and Humanities

PQDT: ProQuest Dissertations and Theses

NTC: National Thesis Center

CHAPTER 1

INTRODUCTION

Being cumulative is one of the most important aspects of scientific enterprise, which is what makes science grow exponentially as well. That was the same idea 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). Pillemer and Light (1980) call attention to the role of research synthesis in terms of cumulative aspects of science approximately 30 years ago, noting that "the need for research synthesis can only be realized when one understands that in order for gains of scholarship to be cumulative, there must be link between past and future research. Often the need for a new study is not as great as the need for assimilation of already existing studies" (p. 2). 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 assist with broadening the scope of the existing knowledge (Card, 2012; Chalmers et al., 2002; Chan & Arvey, 2012; Hunter & Schmidt, 2004; Mulrow, 1994).

1.1 Background and Rationale of the Study

The contribution of research syntheses to cumulative nature of scientific endeavors is essential, yet the 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 based on the necessity of many respondents to reach a conclusion in a survey. He claims 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. He infers that the decisions 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 the ones from outside academic circles, 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.

Besides 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 the 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 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 is unhelpful for policy makers and practitioners to determine what works and what does not work (J. Bennett, 2005). So, research synthesis should be highly encouraged in educational research as it may functionally serve to summarize the overall findings and to explain the reasons for any heterogeneity or contradictions in that results.

It is possible to come across a group of terms including research synthesis, research review and systematic review, which are generally used interchangeably with similar meanings in the literature (Cooper & Hedges, 2009). Although, as stated by Cooper and Hedges, there is no consensus about the differences between these terms, research synthesis is consistently used throughout this dissertation since, first of all, I agree with the idea that the word "synthesis" represents the process better than "review" does. Another reason for choosing "research synthesis" is that "research review" stands for the evaluation process of an article to judge its quality for some purposes like deciding to be published in a journal as well. On the other hand, "systematic review" may cause some confusion by evoking another term "literature review" in the reader. Petticrew and Roberts (2006) claim that because of the similarity between the terms of "systematic reviews" and "literature reviews", "research synthesis" has been becoming gradually more widespread. Finally, as Cooper and Hedges posit while they are explaining the reason why they use "research synthesis" rather than "systematic review", "research synthesis" is more familiar to social scientists comparing to other two terms although the term of "systematic review" is widely used in medical research.

Similar to lack of consensus on the use of the term "research synthesis", there is no agreement about what "meta-analysis" refers to in the literature. Some researchers define "meta-analysis" as a research methodology while others refers 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", explains that he 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 with his colleagues 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). In my dissertation, however, meta-analysis is used to refer a total enterprise of research synthesis; that is to say, the term of "meta-analysis" is used as a research methodology throughout the dissertation. It is mainly because, I believe, 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).

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.

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. 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 (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 evidence (Petticrew & Roberts, 2006). Subjective judgments 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). In this sense, Cooper and Rosenthal (1980), in their experimental study, show the inconsistency of the conclusions drawn by different researchers conducting traditional narrative reviews using the same articles to be reviewed. Similarly, Oakley (2002) takes attention the biased and unrepresentative sample selection of the narrative reviewers by presenting examples from the literature.

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 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; F. L. Schmidt, 1992, 1996; Vacha-Haase, 2001). It is mainly because 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 for a result to be statistically significant but trivial as well (Ellis, 2010; Olejnik & Algina, 2000). Thus, some researchers suggest that statistical testing should be abandoned (Hunter & Schmidt, 2004; F. L. 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 about the idea 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; F. L. 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.

Besides 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).

A meta-analysis is conducted by following similar steps as primary research. The first step of meta-analysis studies is defining the purpose of the review, and developing related research questions. Second, the meta-analyst collects data by searching for relevant studies and selects research studies that meet the specified criteria. Then, the data collected is synthesized by transforming study outcomes to a common metric so that they can be compared. The most commonly used metric is the effect size (d), which is "degree to which a phenomenon exists"

(Cohen, 1977, p. 9). Finally, overall effect size is obtained and the relations between study characteristics and findings are investigated (Bayraktar, 2000).

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), F. L. Schmidt and Hunter (1977), M. L. Smith and Glass (1977), and Rosenthal and Rubin (1978); M. L. Smith, Glass, and Miller (1980). Since 1976 when Glass coined the term of "meta analysis", 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 (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) although there has been some criticism about its use as a research synthesis methodology (Eysenck, 1978, 1984, 1994; Feinstein, 1995; Shapiro, 1994).
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 totally 45,519 results published during the time interval from 1976 to 2012. Figure 1.1 shows how publication number in five years-time intervals increases from the beginning of the modern era of meta-analysis to today. 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 1.2, 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.

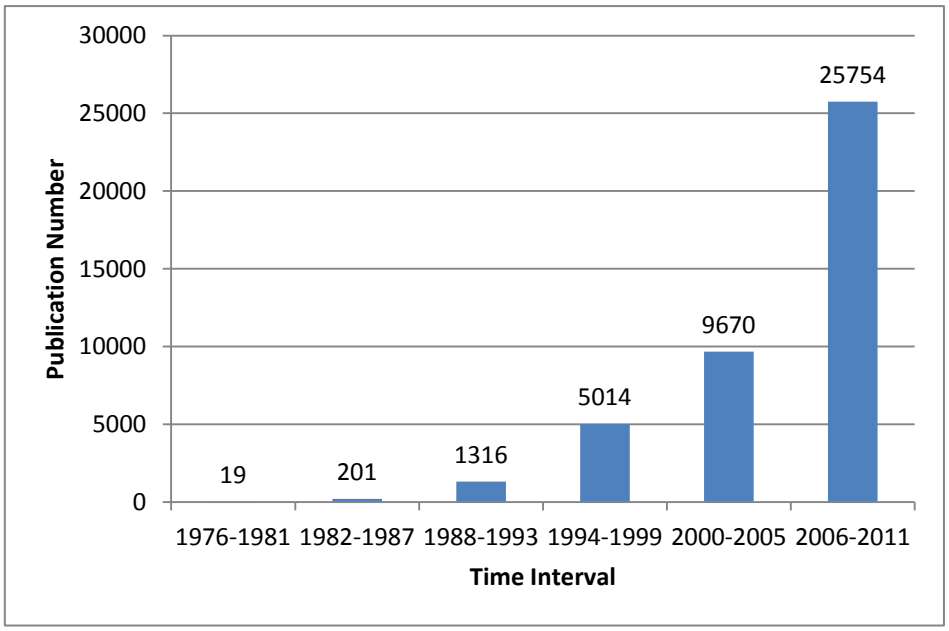

Figure 1. 1 Results of the search for the key term 'meta-analysis' for corresponding time period from 1976 to 2011

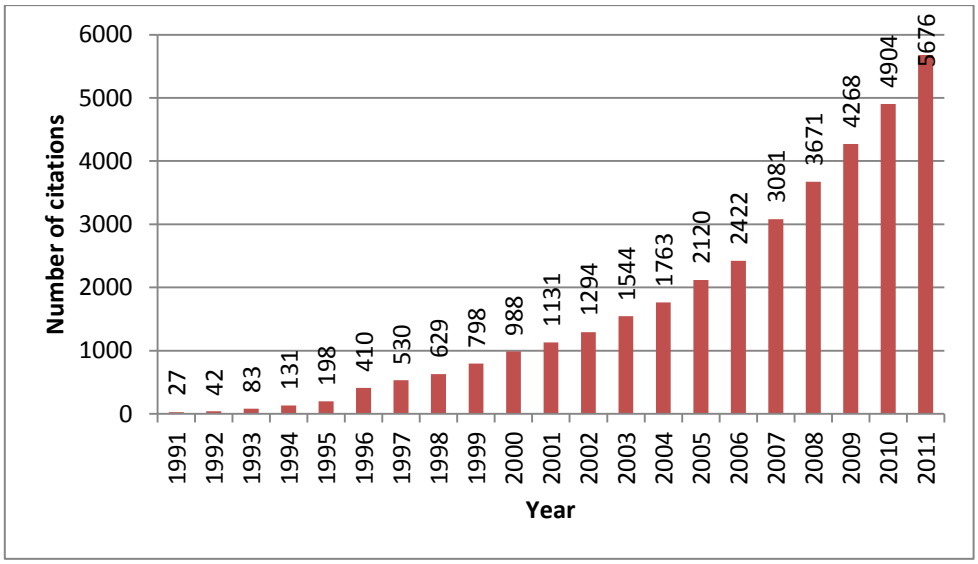

Figure 1. 2 Results of the cited reference search for the keywords 'meta-analysis' and 'education' for the last 20 years

1.2 Problem Based Learning as an Alternative Teaching Method

Different terminologies are used in the literature for similar meaning like teaching methods (Danielson, 2008), teaching strategies (Schroeder, Scott, Tolson, Huang, & Lee, 2007), teaching techniques (Wise & Okey, 1983), learning methods and strategies (Hartley, 2001), instructional methods and strategies (Treagust, 2007), instructional technology (Smaldino, Russell, Heinich, & Molenda, 2005), and instructional systems (Willett, Yamashita, & Anderson, 1983). Although 'learning methods' may be the most appropriate term to use, in this dissertation, 'teaching method' is consistently used for all other similar terms since 'teaching' is used more frequently than 'learning' in the literature.

Alternative teaching methods is one of the most popular topics in science education literature. The search only for the keyword 'teaching methods' by using the databases of Academic Search Complete, Education Research Complete, ERIC, and PsycINFO gives totally 256,154 results, which shows that there are many studies related to teaching methods in the literature. However, findings of educational research often cause contradictions. Even replication studies can produce different results (Berliner, 2002). It is important to underline one of the strengths of meta-analysis that it provides the researcher with the opportunity of identifying and analyzing the heterogeneity of the results on a particular topic, generally presenting the moderator variables that cause the inconsistency as well (Bangert-Drowns & Rudner, 1991; Borenstein et al., 2009; Field, 2003a; Field & Gillett, 2010; Fitzgerald & Rumrill, 2003; Lipsey & Wilson, 2001).

In consequence of the incongruous results, it is possible to find many examples of the meta-analysis studies in the literature aiming to integrate the results of different studies investigating the effectiveness of alternative teaching methods. In this sense, some meta-analysis studies focus on comparing the effects of various teaching methods. They aim to synthesize the results of multiple studies to find out which teaching methods have overall effectiveness on student achievement (D. R. Anderson, Kahl, Glass, & Smith, 1983; Haas, 2005; Marcucci, 1980; Marzano, 1998; Marzano, Pickering, & Pollock, 2001; Schroeder et al., 2007; Wise, 1996; Wise & Okey, 1983). The meta-analyses which have already been conducted to investigate relative effectiveness of teaching methods provide the literature with

comprehensive meta-analyses, and exemplify how to conduct a meta-analysis to investigate the overall effectiveness of different teaching methods and instructional systems in science education. However, they have some limitations in terms of their scope and classification of strategies. Firstly, their scope is limited to the studies conducted in the United States and generally with K-12 grade level. In addition, there exist serious problems with the classification of teaching methods. Firstly, none of the classifications have been developed in a systematic way. Next, the methods involved in the classifications are not familiar with the ones stated in the literature. It should be noted that none of these meta-analyses except for Haas (2005) cover problem based learning as a teaching method in any of their classifications.

On the other hand, some other meta-analyses aim to investigate the overall effectiveness of a particular alternative teaching method. For example, it is evident in the literature that computer-based instruction (CBI) is one of the teaching methods studied very often in meta-analysis (Bayraktar, 2000; Burns & Bozeman, 1981; Christmann, 1997; Clark, 1985; Flinn & Gravatt, 1995; C. C. Kulik & Kulik, 1991; J. A. Kulik, 1983, 1985; J. A. Kulik, Bangert, & Williams, 1983; J. A. Kulik, Kulik, & Cohen, 1980; Liao, 1999; Niemiec & Walberg, 1985), which is mainly because of the conflicting results presented by hundreds of studies investigating the effectiveness of CBI as a teaching method on student achievement. In addition, there are also many meta-analysis studies related to the effectiveness of other teaching methods in the literature like cooperative learning (Igel, 2010; Jonhson, Johnson, & Stanne, 2000; Qin, Johnson, & Johnson, 1995), concept mapping as an instructional tool (Campbell, 2009; Horton & Hamelin, 1993; Nesbit & Olusola, 2006), conceptual change strategies (Guzzetti, Snyder, Glass, & Gamas, 1993), and inquiry based learning (Lott, 1983; Minner, Levy, & Century, 2009; D. Smith, 1996; Sweitzer & Anderson, 1983).

Another alternative teaching method about which many research synthesis studies including meta-analyses have been conducted is problem based learning (PBL) in view of the fact that the results of primary studies investigating the effectiveness of PBL shows too much heterogeneity incorporating both significant (Sungur, Tekkaya, & Geban, 2006) and nonsignificant statistical results (Carrio,

Larramona, Banos, & Perez, 2011; Dobbs, 2008). Although there exist many review studies synthesizing the effectiveness of PBL on different outcomes in the literature (Albanese & Mitchell, 1993; Berkson, 1993; Colliver, 2000; Dochy, Segers, Van den Bossche, & Gijbels, 2003; Gijbels, Dochy, Van den Bossche, & Segers, 2005; Kalaian, Mulllan, & Kasim, 1999; R. A. Smith, 2003; Smits, Verbeek, & De Buisonje, 2002; Vernon & Blake, 1993; Walker & Leary, 2009), none of them focuses on the primary studies in science education. It is evident from the literature that research syntheses about PBL mainly cover the studies from medical education (Strobel & Van Barneveld, 2009). For example, among the research syntheses cited here, Walker and Leary (2009), which aim to compare the effectiveness of PBL across different disciplines, includes only eight studies from science education while others covers almost no studies from science education. As a result of this, there is very limited information about the overall impact of PBL on the dependent variables of achievement, attitude and motivation in science education.

Furthermore, there are some meta-synthesis studies, in the scope of which, the researchers try to combine the results of meta-analyses conducted to synthesize primary studies for effectiveness of PBL (Hattie, 2009; Strobel & Van Barneveld, 2009). The meta-synthesis of Strobel and Van Barneveld (2009) is based on eight PBL meta-analyses while Hattie (2009) synthesizes over 800 meta-analyses relating to achievement, six of which are related to effectiveness of PBL on the achievement. Although these meta-syntheses provide us with a very big picture of the impact of PBL on achievement, both still have the same limitation with the meta-analyses included in these meta-syntheses; they are based on the data largely from medical education but barely from science education.

1.3 Purpose of the Study

The main purpose of this meta-analysis study is to investigate the effectiveness of PBL on not only student achievement and motivation in science, but also attitudes toward science and different types of skills in the school level of elementary, secondary, college, and university. In addition, the effects of some moderator variables including publication type (doctoral dissertations, master theses

and journal articles), research design (true experimental or quasi experimental), teacher effect (same teacher or different teachers for control and experimental conditions), researcher effect (whether researcher is any of teachers in experimental or control conditions), location (different countries), subject matter (physics, chemistry, biology, or general science), school level (primary, secondary or higher education), PBL mode (curriculum model or teaching method), length of treatment, group size, type of questions (open-ended or objective type) and assessment instrument (pre-existing tests or researcher-developed tests) on the effectiveness of PBL are also examined in the scope of this meta-analysis.

1.4 Research Questions

The dependent variables in the following research questions are student achievement and motivation in science (physics, chemistry, biology, or general science), attitudes towards science, and different types of skills; along with the inclusion of the studies is limited by the ones conducted in the school level of elementary, secondary, college, and university; in the time interval of January 1, 1990 and June 1, 2012.

1. To what extent is PBL effective on different outcomes when compared to traditional teaching methods?
2. What is the effectiveness of PBL on science achievement when compared to traditional teaching methods?
3. What is the effectiveness of PBL on students' attitudes toward science when compared to traditional teaching methods?
4. What is the effectiveness of PBL on motivational constructs in science when compared to traditional teaching methods?
5. What is the effectiveness of PBL on different types of skills when compared to traditional teaching methods?
6. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by publication type (doctoral dissertations, master theses and journal articles)?
7. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of research design (true

experimental, quasi experimental with randomly assigned clusters and quasi experimental without randomly assigned clusters)?

8. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of 'teacher effect' (same teacher or different teachers for control and experimental conditions)?
9. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of 'researcher effect' (whether researcher is any of teachers in experimental or control conditions)?
10. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by the countries where the studies are conducted (Turkey, USA and others)?
11. Does the effectiveness of PBL on different outcomes in science when compared to traditional teaching methods differ by subject matter (physics, chemistry or biology)?
12. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by school level (primary, secondary and higher education)?
13. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of PBL mode (curriculum model or teaching method)?
14. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by length of treatment?
15. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by group size?
16. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of questions in the assessment instrument?
17. Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of assessment instrument (pre-existing, researcher developed or adapted)?

1.5 Definition of Important Terms

Meta-Analysis was firstly introduced by Glass (1976). It is one of the ways of doing research synthesis and described 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).

Research Synthesis refers to a group of terms with similar meaning like "systematic review" and "research review" and can be defined as "the application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic" (Last, 2001, p. 176).

Effect Size can be defined as "a measure of the magnitude of a relationship, either in the units of the original measures such as B_{YX} or mean differences, or standardized units such as r, r^2, R, β, or R^{2}" (Cohen, Cohen, West, & Aiken, 2003, p. 673) or "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). Effect size provides a common metric for practical significance of the results of the study independently from the sample size.

Student Achievement in Science and Attitudes toward Science are operationally measured by the effect size values calculated in the studies for the corresponding variables.

Motivational Constructs in Science are defined operationally by the effect size values calculated in the primary studies for a group of related variables like motivation, self-efficacy, self-concept, self-regulated learning skills and meta-cognitive skills.

Skills are operationally measured by the effect size values calculated in the primary studies for different types of skills like critical thinking skills, problem solving skills, science process skills, self-directed learning skills, meta-cognitive skills, inquiry learning skills, logical thinking skills, and self-regulation skills, which are not exclusive completely.

Teaching Methods, in this dissertation, refers to the related terms with a similar meaning like teaching strategies (Danielson, 2008), teaching techniques (Wise & Okey, 1983), learning methods and strategies (Hartley, 2001),

instructional methods and strategies (Treagust, 2007), instructional technology (Smaldino et al., 2005), and instructional systems (Willett et al., 1983).

Traditional Teaching refers to a variety of direct instruction which excludes any type of alternative teaching methods.

Problem-based Learning can be defined as "an instructional (and curricular) learner-centered approach that empowers learners to conduct research, integrate theory and practice, and apply knowledge and skills to develop a viable solution to a defined problem" (Savery, 2006, p. 12). Similarly, in this dissertation, PBL is used as an alternative teaching method in which "relevant problems are introduced at the beginning of the instruction cycle and used to provide the context and motivation for the learning that follows" (Prince, 2004, p. 1).

1.6 Significance of the Study

It is highly emphasized in the literature that reliable and valid research syntheses of individual studies investigating similar research questions are strongly suggested for all disciplines of science (Davies, 2000; Torgerson, 2003), which is mainly based on the idea that it is very rare for a single experiment to provide adequately definitive results upon which to make policy (Chalmers et al., 2002; Davies, 2000; Hedges & Olkin, 1985). Another reason why conducting research synthesis is greatly encouraged is the key role it plays for cumulative nature of scientific enterprise (Chalmers et al., 2002; Chan & Arvey, 2012; Hunter & Schmidt, 2004; Mulrow, 1994). That is the essential idea which motives Card (2012) to argue that "many areas of social science research in less need of further research than they are in need of organization of the existing research" (p.4).

The "crisis" situation as result of contradictory results especially in social and behavioral sciences constitutes another reason for the essentialness and significance of research synthesis (Berliner, 2002; Glass, 1977; Rosenthal, 1991; Rosenthal & DiMatteo, 2001). Berliner (2002) underlines the difficulties to do research in social sciences describing educational studies as hardest-to-do science of all disciplines due to the power of context embedded in complex and unstable networks of social interaction. Similarly, Glass (1976) claims:

> In education, the findings are fragile; they vary in confusing irregularity across contexts, classes of subjects, and countless other factors. Where ten studies might suffice to resolve a matter in biology, ten studies on computer assisted instruction or reading may fail to show the same pattern of results twice. (p. 3)

He also indicates that there is a clear need for meta-analysis since the literature on different topics is growing very fast in education. Furthermore, Petticrew and Roberts (2006) summarize the phenomenon of conflicting results in the literature, even on the same research question, remembering an old scientific joke drawing on Newton's Third Law of Motion: "For every expert there is an equal and opposite expert" (p. 5).

Research studies investigating the effectiveness of PBL as a teaching method provide us with a typical example of what Petticrew and Roberts claim about the contradictory nature of the literature. It is obvious in the literature that the results of some studies significantly favor PBL over traditional method in terms of science achievement (Sungur et al., 2006), attitude towards science (Akınoğlu & Tandoğan, 2007), motivation (Sungur & Tekkaya, 2006) and critical thinking skills (Semerci, 2006) while some others show no statistically significant difference between them (Carrio et al., 2011; Dobbs, 2008). It is also not surprising to come across research studies presenting again statistically significant results, however, indicating reverse direction; that is, traditional method is more effective than PBL (Scott, 2005). So, based on these heterogeneous results of the primary studies, what should be the overall decision of a teacher, an administrator, a curriculum developer or a policy maker who struggles with coming to a decision whether PBL would work in the classrooms or not?

In fact, meta-analysis emerged while Glass was trying to answer a similar question about overall practical significance of psychotherapy in 1976 (Glass, 2000) and has grown up influentially since then. Subsequently, meta-analysis has been widely-used method of synthesizing the results of empirical studies within many disciplines of science as a result of its superiority over other approaches to

research synthesis like conventional narrative reviews, vote-counting method or combined significance tests (Lipsey & Wilson, 2001).

Thus, it is not surprising that many research syntheses, mostly using meta-analysis as research methodology, have already been conducted to combine the conflicting results for the effectiveness of PBL on different outcomes (Albanese & Mitchell, 1993; Berkson, 1993; Colliver, 2000; Dochy et al., 2003; Gijbels et al., 2005; Kalaian et al., 1999; R. A. Smith, 2003; Smits et al., 2002; Vernon & Blake, 1993; Walker & Leary, 2009). However, almost all of them focus on medical education including very few primary studies from science education. Thus, it can be easily deduced that there is an apparent need in the literature for meta-analyses investigating overall impact of PBL in science education.

Since it is the first to investigate the overall effectiveness of PBL focusing on science education, this meta-analysis study has important functions to fill an important gap in the literature. It not only gives us the opportunity of constructing a more complete picture by synthesizing empirical studies conducted with the students from elementary to higher levels in different countries but attempts to reveal the variables that moderate the effectiveness of PBL on different outcomes as well.

Beside the contributions to science education literature, this meta-analysis aims to be helpful for policy makers, educational administrators, science curriculum developers, and textbook authors by providing them with an evidence-based answers to the questions about whether or in which conditions PBL works as a teaching method. In this manner, the results of the study will also provide great guidance to the science teachers who want to improve their instruction by using PBL for the specific conditions (context, subject matter, student characteristics, grade level, etc…) in which it works better.

CHAPTER 2

LITERATURE REVIEW

The main purpose of this chapter is to summarize related literature about meta-analysis, advantages and criticism of meta-analysis, previous meta-analysis studies comparing the effectiveness of different teaching methods with theoretical background, advantages and limitations, and effectiveness of PBL.

2.1 Meta-Analysis as a Method of Research Synthesis

Lord Rayleigh underlined an important point in his presidential address to the 54th meeting of the British Association for the Advancement of Science as follows:

> The work which deserves, but I am afraid does not always receive, the most credit is that in which discovery and explanation go hand in hand, in which not only new facts presented, but their relation to old ones is pointed out (As cited in Chalmers et al., 2002, p. 30).

What Lord Rayleigh highlights in his speech is the importance of creating links between new and existing scientific knowledge by assimilating new discoveries into the framework constructed upon the literature shaped by previous scientific endeavor, which lies behind the idea about how science develops: "the advancement of scientific knowledge is based on the systematic building of one study on top of a foundation of prior studies, the accumulation of which takes our understanding to ever increasing heights" (Card, 2012, p. 3).

Rosenthal and DiMatteo (2001) call attention to another important issue that scientific studies in nearly every field are increasing almost explosively in the twenty first century. They also stress the conflicting situation in findings in terms of

central issues of theory and practice in different domains like psychology, education, medicine and other related disciplines. Thus, they claim a resolution of this conflicting situation is necessary for further advance of any field and for any related practical application. Furthermore, they posit that there is also a high demand for more accurate estimation of descriptive statistics giving some examples of how much variability it may show, which makes it "challenging and precarious" to take decisions based on these data.

Similarly, Hunter and Schmidt (2004) assert that ultimate goal of any scientific enterprise is to produce cumulative knowledge describing, on the other hand, a pathetic situation of the research literatures, most of which, they claim, present highly contradictory results with a split of approximately 50-50 for many cases. They are not the only researchers depicting these challenging limitations of the literature but many others (J. Bennett, 2005; Borenstein et al., 2009; Davies, 2000; Glass, 1976; Hunt, 1997; Mulrow, 1994; Oakley, 2002; Petticrew & Roberts, 2006; Torgerson, 2003) underline these well-known problems and encourage scientists to synthesize existing literature about specific research problems in order to not only cumulate the scientific knowledge already been constructed but provide meaningful explanations for conflicting results as well.

Mulrow (1994) goes over the main premises that the rationale of research synthesis is grounded by describing ten reasons why it is a fundamental scientific activity. The first one is that research synthesis gives us the opportunity of handling large amount of information reducing into small and meaningful pieces of knowledge. Next, it has the potential of providing practitioners and policy makers with important contributions by presenting an integrated knowledge in much broader aspects comparing to primary studies. Furthermore, she summarizes other strengths of research synthesis as being an efficient scientific method, having increased power, precision, accuracy, and the ability of broadening the generalizability of the scientific findings, assessing consistency and explaining any inconsistent or conflicting result, and finally presenting the methods followed by researchers explicitly.

Meta-analysis, which 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, 2010b, p. 274), is one of the most widely-used methods of research synthesis (Lipsey & Wilson, 2001). As stated in Dieckmann, Malle, and Bodner (2009), traditionally, narrative reviews have been used to synthesize research findings but in recent years, more quantitative approaches have become more popular. They also argue that research syntheses, and specifically meta-analyses, have an essential role in scientific literatures.

The study of Pearson (1904) can be accepted as the starting point of the quantitative synthesis of the research findings (Lemeshow, Blum, Berlin, Stoto, & Colditz, 2005; Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001). He used correlation coefficients from different studies to determine the extent of the relationship between inoculation against smallpox and survival rate. Based on the data included in the study, he concluded that there is a large effect (above .60 either for weighted or unweighted case) with substantial clinical significance.

Although there were examples of the quantitative synthesis of different research studies formerly, the term of "meta-analysis" was firstly introduced by Glass (1976) as "the analysis of analyses" (p.3). Glass classifies the data analysis as primary and secondary analysis, and places meta-analysis as an advancement of practice of secondary analysis.

After the beginning of 1980s, around when the modern era of meta-analysis has began, meta-analysis has gained popularity with an increasing rate in various domains especially in the field of biomedicine and behavioral sciences (Rosenthal & DiMatteo, 2001; Schulze, 2004, 2007; Viechtbauer, 2007). Furthermore, Schulze (2007) asserts "more than 30 years after the term was coined, meta-analysis has earned its place in the pantheon of scientific methods" (p. 87). He also indicates that meta-analysis has become a standard method of research synthesis in many fields, but especially in the social sciences.

2.2 Why Meta-Analysis rather than Other Research Synthesis Methods?

Meta-analysis, being one of the most comprehensive and systematic ways of conducting research synthesis, has obvious advantages over conventional review,

which suffers from serious limitations as illustrated by Cooper and Rosenthal (1980), Oakley (2002), and Bushman and Wells (2001).

Firstly, 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 only 3 studies were common to both reviews. Furthermore, she claimed that there were at least 70 more studies met the inclusion criteria of the reviews in the literature.

Finally, Bushman and Wells (2001) illustrated the corrective properties of meta-analysis against the biased and subjective decisions based on narrative reviews in another study conducted with 280 participants. 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 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 the limitations that traditional narrative reviews suffer from like biased and subjective judgments and unrepresentative sampling. Furthermore, increasing number of primary studies to be synthesized results in not only increased power and precision but flexibility to examine the inconsistencies in the results (if exist, of course) 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).

Another advantage of meta-analysis stem from the idea lying behind it, which is practical significance. Although other quantitative alternatives of meta-analysis, like vote counting method and combined significance tests, are based on statistical significance, which does not guarantee practical significance at all since it does not mean it is not trivial in terms of impact in real life (Borenstein et al., 2009; Cohen, 1990; Ellis, 2010; Hunter & Schmidt, 2004; Kirk, 1996, 2001; F. L. Schmidt, 1996; Vacha-Haase, 2001), meta-analysis mainly is constructed upon measures of effect size, an indicator of how big the effect is, i.e. practical significance.

Lipsey and Wilson (2001) point out four reasons why we should use meta-analysis to summarize and analyze a body of research studies rather than conventional research review techniques. These reasons also constitute the primary advantages of meta-analysis. First, 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 narrative 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 and 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.

2.3 Criticisms 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 is explained by summarizing the strengths of meta-analysis stated by different researchers. However, there are also some criticisms about meta-analysis in the literature, which are 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 studies. That is, results from poorly designed studies are included into the meta-analysis to be synthesized along with results from good 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 significant results because 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. 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. Another approach is to estimate the number of additional studies with non-significant results that would be necessary to bring the overall treatment effect to nonsignificance, which presents an estimate of the robustness and validity of the findings. 'Rosenthal's Fail-safe N (FSN)' can be used to specify 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, usually .05 or .01 (Rosenthal, 1991). Table 2.1 exemplifies the FSN tabulation.

Table 2.1 *An example for FSN computation from Schroeder et al. (2007) (N_{fs}: Rosenthal's FSN)*

Data	ES	N	N_{fs}
Overall	0.67	61	756
Questioning Strategies	0.74	3	42
Manipulation Strategies	0.57	8	84
Enhanced Materials Strategies	0.29	12	58
Assessment Strategies	0.51	2	19
Inquiry Strategies	0.65	12	145
Enhanced Context Strategies	1.48	6	172
Instructional Technology Strategies	0.48	15	130
Collaborative Learning Strategies	0.96	3	55

While Rosenthal's FSN focuses on p values, i.e. statistical significance, an alternative approach developed by Orwin (1983) results in another FSN value, which is calculated on the basis of practical significance. Orwin's FSN value gives the number of the primary studies with a specific effect size would be needed to reduce the calculated mean effect size to a particular effect size value chosen by the researcher (Becker, 2005). Besides these diagnosis procedures, Trim and Fill Method (TFM) developed by Duval and Tweedie (2000a, 2000b) provides meta-analysts with estimation and adjustment of the impact of publication bias (Becker, 2005; Borenstein et al., 2009; Sutton, 2009). Several methods for not only diagnosis but adjustment purposes as well shows clearly that meta-analysis is not a source of publication bias but it is a part of solution (Glass, 1982; Sutton, 2009).

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 is closer to 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.

2.4 Previous Meta-Analyses Comparing the Effectiveness of Different Teaching Methods

It is evident in the literature that there exist both earlier and recent meta-analysis studies comparing the effectiveness of alternative teaching methods in the scope of comprehensive research projects. For example, in 1983, only seven years later than Glass coined the term "meta-analysis", the results of a broad meta-analysis study were reported in several research papers contained in a special issue in the Journal of Research in Science Teaching (D. R. Anderson et al., 1983; Druva & Anderson, 1983; Fleming & Malone, 1983; Lott, 1983; Shymansky, Kyle Jr, & Alport, 1983; Sweitzer & Anderson, 1983; Willett et al., 1983; Wise & Okey, 1983). In the study, it was aimed to synthesize the results of the research studies investigating the major science education research questions.

Anderson et al. (1983) provide general information including the purpose and the scope of the study, and research questions to be meta-analyzed. In the meta-analysis project, the studies are limited to those conducted in the context of K-12 grade, within the United States and published between 1950 and research date. Two of research questions selected in this study by empirical process of identifying the most frequently researched questions in the literature are as follows:

- What are the effects of different instructional system used in science teaching (e.g., programmed instruction, mastery learning, and departmentalized instruction)?
- What are the effects of different teaching techniques (e.g., questioning behaviors, wait-time, advance organizers, and testing practices)? (p. 381)

Wise and Okey (1983), in the scope of this project, examined the effect of different teaching techniques on achievement. Teaching techniques are grouped into 12 categories in this meta-analysis, which are: questioning, wait-time, testing, focusing, manipulative, presentation approach, inquiry or discovery, audio-visual, teacher direction, grading, modified, and miscellaneous. 400 effect sizes representing 160 studies conducted in the United States are included in the meta-analysis. The average of the overall effect sizes for all teaching techniques is calculated as 0.34, and more than 20 moderator variables are introduced and tabulated with effect size, such as class size, community type, and science subject area. Table 2.2 gives an example of the effects of moderator variables; tabulates mean effect sizes obtained in classes of different size. Results of the meta-analysis indicate that the teaching techniques with a mean effect size more than 0.50 are wait-time (0.90), focusing (0.57), manipulative (0.56), and modified (0.52). For only one type of teaching technique, negative mean effect size is reported; grading (-0.15). They conclude that the effective teaching techniques are the ones in which students are kept aware of instructional objectives and receive feedback on their progress. In addition, students should be given the opportunity of physically interacting with instructional materials and participating in varied kinds of activities. They also claim that the effect sizes associated with different class sizes provide policy makers who advocate smaller classes with strong evidence.

Table 2.2 *An example of how moderator variables affect the magnitude of effect size: Mean effect sizes obtained in classes of different size (Wise & Okey, 1983)*

Size of Class	Mean Effect Size	SD	Number of Cases	% of Cases
Fewer than 15 Students	0.74	0.86	32	11
15-24	0.37	0.60	119	39
25-34	0.23	0.46	114	37
35 or more Students	0.23	0.57	38	13

Willett et al. (1983), in the scope of the same project with Wise and Okey (1983), investigate the effectiveness of different instructional systems used in science teaching. They specify 12 categories of instructional systems, which are; audio-tutorial, computer linked (also reported separately in three categories as computer assisted instruction, computer managed instruction, and computer simulated experiments), contracts for learning, departmentalized elementary school, individualized instruction, mastery learning, media-based instruction (also reported separately as film instruction and television instruction), personalized system of instruction, programmed learning (branched and linear programmed learning), self-directed study, use of original source papers in teaching of science, and team teaching. 341 effect sizes from 130 studies are included in the meta-analysis. They state that the mean effect size for all instructional systems is 0.10, indicating that, on the average, an innovative instructional system produce one-tenth of a standard deviation better performance than traditional science teaching. Year of publication, form of publication, grade level, and subject matter are also considered as moderator variables in this meta-analysis study. The studies are limited to the ones conducted in the context of K-12 grade, in the United States and published between 1950 and research date. The studies in which no control group is used are also omitted in this meta-analysis. The instructional systems with a mean effect size more than 0.50 are reported as mastery learning (0.64) and personalized system of instruction (0.60).

Wise and Okey (1983) and Willett et al. (1983) provide the literature with comprehensive meta-analyses and exemplify how to conduct a meta-analysis to

investigate the overall effectiveness of different teaching techniques and instructional systems in science education. However, they have some limitations in terms of their scope and classification of strategies. Firstly, their scope is limited to the studies conducted in the United States. In addition, they do not differentiate neither 'medium and method' nor 'method and technique'.

In another study, Wise (1996) reports the results of another meta-analysis examining the effect of alternative teaching strategies on student achievement at the middle and high school levels. The studies are limited to the ones conducted between 1965 and 1985. He places the research studies into eight categories of alternative teaching strategies including questioning, focusing, manipulation, enhanced materials, testing, inquiry, enhanced context, and instructional media. He proposes that the difference between alternative strategies and traditional ones is the use of inquiry-oriented instruction. He indicates that overall mean effect size for alternative teaching strategies is 0.32. The alternative teaching strategy with the highest effect size is reported as questioning (0.58) while the one with the smallest effect size is instructional media strategies (0.18).

In a more recent study, Marzano et al. (2001) also synthesize several research findings and conclude that nine broad teaching strategies have positive effects on student learning. These are; identifying similarities and differences, summarizing and note taking, reinforcing effort and providing recognition, homework and practice, nonlinguistic representations, cooperative learning, 'generating and testing hypotheses', questions, cues, and advance organizers.

Similarly, Haas (2005), in another meta-analysis, explores the overall effect of teaching methods on algebra student achievement. Six teaching method categories are constructed including cooperative learning, communication and study skills, technology-aided instruction, problem-based learning, 'manipulatives, models and multiple representations', and direct instruction. Totally 35 studies conducted between 1980 and 2002 at the secondary level are included in the study. Two teaching methods with the highest mean effect sizes are indicated as direct instruction (0.55) and problem-based learning (0.52).

In one of the most recent meta-analyses comparing the effect of alternative teaching strategies, Schroeder et al. (2007) reveal eight categories of teaching

strategies: questioning strategies, manipulation strategies, enhanced material strategies, assessment strategies, inquiry strategies, enhanced context strategies, instructional technology strategies, and collaborative learning strategies. This classification is constructed by revising the categories stated by Wise (1996). The studies in the meta-analysis are limited to the ones related to science education, conducted with K-12 student in the United States and published between 1980 and 2004. According to these criteria, totally 61 studies are included in the meta-analysis. Publication type, type of study (experimental or quasi experimental), publication year, grade level, and test content area are also scrutinized as moderator variables in this study. Overall mean effect size for all studies included in the meta-analysis is calculated as 0.67. In addition, enhanced context strategies (1.48), collaborative learning strategies (0.95), questioning strategies (0.74), inquiry strategies (0.65), manipulation strategies (0.57) and assessment strategies (0.51) are reported as the teaching strategies with the effect sizes bigger than 0.50. They emphasize:

> If the students are placed in an environment in which they can actively connect the instruction to their interests and present understandings and have an opportunity to experience collaborative scientific inquiry under the guidance of an effective teacher, achievement will be accelerated. (p. 1452)

The meta-analyses conducted to explore relative effectiveness of various teaching methods exemplify how to perform comprehensive meta-analyses, which are generally conducted as extensive group projects funded by different organizations. However, their scopes are restricted to the studies carried out in the United States and generally with K-12 grade level. In addition, the classifications of teaching methods suffer from serious problems due to lack of systematic ways to develop the categories, which are sometimes not comparable at all: some of them are too specific while there exist some others in the same classification, which are too broad to be called as a teaching method. It is also noteworthy that none of the classifications except for the one developed by Haas (2005) cover problem based learning as a teaching method.

2.5 Previous Meta-Analyses Investigating the Effectiveness of a Particular Teaching Method

Besides the meta-analyses comparing the effectiveness of various alternative teaching methods, there exist many review studies investigating the overall effectiveness of a particular teaching method in the literature; like the studies for CBI (Bayraktar, 2000; Burns & Bozeman, 1981; Christmann, 1997; Clark, 1985; Flinn & Gravatt, 1995; C. C. Kulik & Kulik, 1991; J. A. Kulik, 1983, 1985; J. A. Kulik et al., 1983; J. A. Kulik et al., 1980; Liao, 1999; Niemiec & Walberg, 1985), for cooperative learning methods (Igel, 2010; Jonhson et al., 2000; Qin et al., 1995), for concept mapping as an instructional tool (Campbell, 2009; Horton & Hamelin, 1993; Nesbit & Olusola, 2006), for conceptual change strategies (Guzzetti et al., 1993), and for inquiry based learning (Lott, 1983; Minner et al., 2009; D. Smith, 1996; Sweitzer & Anderson, 1983).

Bayraktar (2000, 2002) conducts a meta-analysis in the scope of her dissertation study to explore the overall effectiveness of CAI on student achievement in secondary and college science education when compared to traditional instruction. 42 studies with 108 effect sizes included in the study result into a small overall effect size of 0.273, which moves a student from the 50^{th} percentile to 62^{nd} percentile. The analysis to investigate the moderator variables indicates that student-computer ratio, CAI mode, and duration of treatment are significantly related to the effectiveness of CAI. Bayraktar claims that computers are more effective when used individually and in simulation or tutorial mode.

In another meta-analysis study, Jonhson et al. (2000) classify the cooperative learning methods and examine the overall effect of each method on student achievement. They compare cooperative learning with both competitive and individualistic learning. The cooperative learning methods stated in the study are Learning Together (LT), Academic Controversy (AC), Student-Team-Achievement-Divisions (STAD), Teams-Games-Tournaments (TGT), Group Investigation (GI), Jigsaw, Teams-Assisted-Individualization (TAI), and Cooperative Integrated Reading and Composition (CIRC). The results of the study show that all eight cooperative learning methods have a significant positive effect compared to competitive and individualistic learning while LT promotes the

greatest positive effect on student achievement for both comparisons. More recently, Igel (2010), in his dissertation study, synthesizes 20 primary studies by means of meta-analysis to investigate the overall effect of cooperative instruction on K-12 student learning as well and results in a moderate overall effect size, 0.44 for cooperative interventions. Because of limited number of studies included in meta-analysis, moderator analysis is performed just by means of descriptive statistics; that is, presenting effect size values for subgroups formed by possible moderator variables like treatment length, subject and grade.

In another dissertation study, Campbell (2009) conducts a meta-analysis to explore how effective the concepts mapping is for enhancing achievement among students from different domains like science and mathematics. She includes 46 effect size values from 38 studies and concludes that concept mapping is effective on achievement in all domains including science except for mathematics. The overall effect size value for science domain, which is represented by 28 effect size values, is calculated as 0.84 while it is -0.91 for mathematics domain. However, it is impossible to assume that the number of studies in mathematics domain is enough to be called as a representative sample for the population of the primary studies in this domain since only two studies examining the effectiveness of concept mapping on mathematics achievement are included. More recently, Nesbit and Olusola (2006) synthesize 67 effect size values from extracted 55 experimental or quasi-experimental studies in a meta-analysis study. In terms of overall impact, the results of this meta-analysis support what is claimed by Campbell about the effectiveness of concept mapping on achievement. They also conclude that concept mapping is effective especially on knowledge retention, calculating an overall mean effect size of 0.604 for all subjects.

2.6 What is PBL?

Maudsley takes attention to the faulty use of the "term" PBL stating that "many 'PBL' single-subject courses within traditional curricula do not use PBL at all" (1999, p.178). He also claims that this conceptual uncertainty casts a suspicion on the effectiveness of PBL over direct instruction. Apart from these erroneous uses, different definitions of PBL being derived as a result of extensive use in

different disciplines exist in the literature ranging from a teaching method to a curricular approach. Barrows (1996) posits that application of PBL has become so wide-spread that many variations have already been evolved from the core model, which makes it necessary to define the basic principles of PBL.

Maudsley (1999) differentiates the terms of "PBL" and "problem based curriculum", former of which is described as an isolated teaching method for not the whole but some parts of curricula or for a specific subject while the later identifies the whole curriculum which is developed on the basis of ill defined problems. On the other hand, Savery (2006) defines PBL as "an instructional (and curricular) learner-centered approach that empowers learners to conduct research, integrate theory and practice, and apply knowledge and skills to develop a viable solution to a defined problem" (p.12). Similarly, Hmelo-Silver (2004) describes PBL as "an instructional method in which students learn through facilitated problem solving" (p. 235). She indicates that, in PBL, students focus on complex, ill defined problems having more than a single correct answer. In addition, Prince (2004) describes PBL as stressing one of its key characteristics: "PBL is an instructional method where relevant problems are introduced at the beginning of the instruction cycle and used to provide the context and motivation for the learning that follows" (p. 1).

Consequently, it is obvious in the literature that PBL has a wide range of definitions, some of which does not represent PBL at all (Maudsley, 1999) while others differs in how it is applied although the basic ideas behind the application is almost the same. Thus, identifying basic characteristics of PBL is essential and may be more functional than trying to describing it more specifically. Barrows (1996) explains six principles that shape PBL, the first one of which is that learning is student centered. That is, students become responsible for their own learning, which shifts the role of teachers from being knowledge transmitters to be facilitators or guides. So, in PBL, teachers are no longer supposed to present information to students directly, rather they are expected to guide and scaffold students though whole learning process (Hmelo-Silver, 2004). Next principle of PBL is that learning generally occurs in small groups of students. Another basic principle is the fact that problems, which are tools for developing students' problem solving skills, form the

focus and stimulus for learning. Finally, self directed learning is one of the key concepts that lie behind the idea of PBL.

Dochy et al. (2003) indicate also a seventh characteristic of PBL that students' competencies should be evaluated by a valid assessment system based on real life problems. In addition, Hmelo-Silver (2004) underlines two key issues for learning through PBL. Firstly, she states that learners construct their own knowledge actively in collaborative groups. Secondly, she claims, the teacher should not be considered as to be the main repository of knowledge anymore. She also summarizes the main goals of PBL as to help students construct an extensive and flexible knowledge base, which is beyond the facts of a domain, develop effective problem-solving skills including appropriate meta-cognitive and reasoning strategies, develop self-directed, lifelong learning skills, extend cooperative learning skills, and finally become intrinsically motivated learners.

2.7 Theoretical Background of PBL

Theoretical background of PBL, which has been developed by different researchers (Ausubel, Novak, & Hanesian, 1978; Bruner, 1961; Dewey, 1938; Piaget, 1954; Vygotsky, 1978), has a long history (Dochy et al., 2003; Gijbels et al., 2005; Hmelo-Silver, 2004). Although PBL has many roots on different educational theories (Savin-Baden & Major, 2004), theoretical principles underlying PBL are mainly based on experiential learning (Dewey, 1938; Kilpatric, 1918, 1921), discovery learning (Bruner, 1961) and finally, perhaps most prominently, constructivism (Piaget, 1954; von Glasersfeld, 1989; Vygotsky, 1978).

Savery and Duffy (2001) identify three primary propositions which characterize the philosophical view of constructivism. Firstly, they call the fact that cognition is distributed as the core concept of constructivism, which means cognition is a part of entire context. Secondly, they posit that cognitive conflict is the basic stimulus for learning. Finally, they emphasize the role of social negotiation and the evaluation of the viability of individual learning in the process of constructing knowledge. They also underline eight instructional principles, which are derived from constructivism and consistent with PBL as follows:

- Embed all learning activities into a larger, real life problem or task.

- Scaffold students to feel ownership for the overall problem.
- Develop an authentic task.
- Do not simplify the learning environment for students; rather design the environment in a way that it reflects the complexity of real life.
- Provide students with the freedom of using their own problem solving process.
- Create a learning environment, which supports and challenges students' judgment.
- Encourage students to test alternative views and to check their views in different contexts.
- Give students with the chance of reflecting their ideas about both the content and the process to promote their reflective thinking.

Apart from Savery and Duffy (2001), it is evident from the literature that there are many opponents of the idea that basic principles of PBL is highly compatible with the instructional and epistemological propositions asserted by constructivism (Akçay, 2009; Chin & Chia, 2005; Duffy & Cunningham, 1996; Gijselaers, 1996; Hmelo-Silver, 2004; Taşkesenligil, 2008).

Although theoretical background of PBL has a long history, it was originally designed for medical education in 1960s at McMaster University in Canada. Then, some of medical schools using conventional curricula designed alternative problem based curricula for some of their students by the early 1980s. Another contribution for wider dissemination of PBL was provided by "Report of the Panel on the General Professional Education of the Physician and College Preparation for Medicine (GPEP Report)", which made recommendations for promoting problem solving and self regulation skills (Barrows, 1996). Then, PBL has spread out in variety of levels and settings including science, engineering and economics ranging from elementary to higher education levels (Ateş, 2009; Hmelo-Silver, 2004).

2.8 Advantages and Disadvantages of PBL

Wood (2003) summarizes some of the advantages of PBL as stating that PBL:

- is student centered, which promotes active learning, enhance understanding and improves life-long learning skills.
- develops students' generic skills like critical evaluation and self-directed learning.
- fosters students to be engaged in the learning process by increasing motivation.
- makes students to be involved in interaction with learning materials and scaffolds them to relate concepts to real life activities.
- is based constructivist approach.

In a similar way, Hmelo-Silver (2004) claims that PBL helps students construct a broad and flexible knowledge base and develop effective problem solving, self-directed, and lifelong learning skills. Furthermore, she asserts that PBL promotes students to be successful collaborators and intrinsically motivated to learn as well.

On the other hand, Wood (2003) posits that PBL has also some limitations or disadvantages. She mentions the possible resistance that tutors may show because they may find PBL process difficult and frustrating. Furthermore, she underlines the increasing need for human and other resources since more tutors are needed in PBL comparing to the need in a lecture based instruction and students have to use the library or computer resources at the same time. Finally, she warns that students may be overloaded as a result of self directed study without appropriate guidance. Likewise, Uden and Beaumont (2006) indicate that being more time consuming comparing to lecturing is one of the main limitations of PBL. They also take attention to increasing need for resources like library materials and extra room for cooperative studies in PBL.

2.9 Effectiveness of PBL on Different Outcomes

Although PBL is highly wide-spread in not only medical education, which is still being in the first place in terms of extensiveness of PBL, but also in many other disciplines as a teaching method or curricular innovation, the effectiveness of PBL on different outcomes especially on achievement is still an exceptionally controversial issue in the literature. Unfortunately, primary studies comparing

effectiveness of PBL and direct instruction does not provide coherent results. It is both possible to find studies favoring PBL (Sungur et al., 2006) and the ones favoring direct instruction (Scott, 2005).

Based on the studies favoring direct instruction and theoretical explanations about human cognitive structure, Kirschner, Sweller, and Clark (2006) declare that all instructional approaches advocating minimal guidance during instruction including constructivist, discovery, problem based, experiential and inquiry based teaching do not work as effective as direct instruction, which provides students with enough guidance. This assertion results in a vehement argument between them and proponents of what they call as minimally guided instruction (Hmelo-Silver, Duncan, & Chinn, 2007; Kuhn, 2007; H. Schmidt, Loyens, Van Gog, & Paas, 2007; Sweller, Kirschner, & Richard, 2007).

First of all, H. Schmidt et al. (2007) disagree with the idea stated by Kirschner et al. (2006) that PBL provides minimal guidance during instruction. They underline that PBL is a teaching approach that allows for flexible adaptation of guidance, thus, the principles underlying PBL runs in with human cognitive structures very well. Furthermore, they criticize Kirschner et al. not to mention the studies either explaining the reasons why PBL seems to be ineffective or claiming that PBL is more effective than direct instruction (Dochy et al. 2003).

Similarly, Hmelo-Silver et al. (2007) emphasize two main defects in Kirschner et al's argument. Firstly, akin to H. Schmidt et al. (2007), they underline that it is not acceptable to categorize PBL or inquiry based learning as a minimally guided instruction since they provide students with extensive scaffolding or guidance to promote student learning. Secondly, they claim that Kirschner et al. ignore the empirical evidence that favors PBL or inquiry based learning rather than direct instruction and present some examples of the primary studies (Capon & Kuhn, 2004) and meta-analyses (Dochy et al., 2003; Vernon & Blake, 1993) which support the proposition that PBL is not significantly less effective on declarative knowledge tests while it is significantly more effective than direct instruction on measures of knowledge application.

Additionally, Kuhn (2007) criticizes Kirschner et al. that they ignore any context of "what it is that is being taught by whom or to whom" while comparing

the effectiveness of teaching methods. They disagree with the idea that "how best to teach and learn are universally applicable, irrespective of what is being taught to whom or why" by emphasizing the importance of motivation, which results from the interaction between individual and subject matter and skills of inquiry and argument, which are essential aspects of science curriculum.

Finally, in their reply to all commentaries on Kirschner et al., Sweller et al. (2007) emphasize once again that they do not agree with the suggestion that the presentation of the relevant information should be lessened by putting more emphasis on teaching students the ways to reach information. They also do not accept the proposition made by H. Schmidt et al. (2007) that PBL do not deemphasize the guidance stating that it is in conflict with essential goal of PBL. They highlight that the "raison d'etre" of PBL is to play down direct instructional guidance remembering that PBL emphasizes "self-directed" learning. They repeat their ideas about the aim of learning, which is, they believe, is to increase knowledge in long term memory. They explain their reasoning by asserting that the central component of human cognition is knowledge, not "the ability to devise novel, general problem solving and thinking strategies" as constructivists claim. Furthermore, they stress the importance of theories and findings of science, down-playing of which, they state, would have adverse consequences.

2.9.1 Research Syntheses Focusing on the Effectiveness of PBL

Several syntheses including both narrative reviews and meta-analyses have already been conducted to provide empirical evidence for hot debate about the effectiveness of PBL by integrating the primary studies, which hardly present consistent results.

Firstly, Vernon and Blake (1993) conduct a meta-analysis to synthesize studies, which, in the time scope from 1970 to 1992, compare the effectiveness of PBL and traditional methods of medical education on different outcomes like program evaluation and achievement covering measures of both factual information and clinical performance. Totally 35 studies representing 19 institutions are included in the review but only for 22 studies, for which effect size values could be calculated, are integrated in effect size analyses while for others, vote counting

analyses are performed. In terms of program evaluation, which is mainly based on students' attitudes and opinions about their programs, PBL is found to be more effective than traditional methods with an overall medium effect size weighted by sample size of 0.55. Similarly, the results favor PBL with a weighted effect size of 0.28 for students' clinical performance. On the other hand, there is no significant difference between PBL and traditional methods on the results for outcomes of factual knowledge while average mean of the students from traditional methods is significantly better than the ones of students from PBL with a small negative effect size, -0.18) on the National Board of Medical Examiners (NBME) Part I, which is the knowledge acquisition part of the examination including multiple choice questions. However, with the addition of five studies, which are not included in effect size analysis due to lack of enough information for an effect size value to be calculated, the results of vote counting method show no difference between effectiveness of PBL an traditional methods based on NBME Part I. To sum up, Vernon and Blake conclude that the results of meta-analysis generally support the superiority of PBL over traditional methods.

Similarly, Albanese and Mitchell (1993) conduct a meta-analysis to compare the overall effectiveness of PBL and conventional instruction by covering the research studies from 1972 to 1992. They state several conclusions based on data included in the meta-analysis, first of which is that PBL is more fostering and enjoyable comparing to conventional instruction. Another conclusion based upon clinical examinations and faculty evaluations is that the graduates of PBL perform at least as well or sometimes better than the graduates of conventional medical instruction. However, they also infer that, in a small number of instances, the results of PBL students are worse on basic science examinations and additionally they feel themselves less-prepared in the basic sciences than conventionally trained students. Finally, they warn practitioners and educational administrators that effectiveness of PBL might be slowed down by the costs when class size are larger than 100.

On the contrary, a narrative review conducted by Berkson (1993) in the same year and for the same purpose with Vernon and Blake (1993) and Albanese and Mitchell (1993), presents highly contradictory results about the effectiveness of PBL on problem solving skills, imparting knowledge, motivation, and self-directed

learning. She compares PBL and traditional medical education in terms of students and faculty satisfaction and cost-effectiveness as well. As a result, she concludes that the effectiveness of PBL and traditional medical instruction is not distinguishable while PBL experience may be more stressful on both student and faculty and the cost of implementation of PBL may be unrealistically high. This study generally suffers from the limitations of narrative reviews like researcher bias and unrepresentative sample of studies, which may explain the conflicting results partially. Distlehorst (1994) and Vernon and Blake (1994), on their commentaries, underline some of the limitations of the conventional review by Berkson, which result in, they believed, conflicting results about the effectiveness of PBL. Distlehorst makes some corrections for the information provided by Berkson about the primary studies she involved in her review. Distlehorst also warns Berkson about the variety of PBL implementations in different institutions, which makes it very difficult to reach some overall conclusions about the effectiveness of PBL in medical education. On the other hand, Vernon and Blake put emphasis some methodological issues about conventional narrative reviews, which they believe, generate the differences between the results of this narrative review and two meta-analyses conducted by Vernon and Blake and Albanese and Mitchell in the same year, the findings of which are almost parallel.

In another meta-analysis study, Kalaian et al. (1999) compare the impact of PBL and traditional curricula on the students' achievement measured by NBME Part I and Part II independently. They include 22 studies evaluating the effects of PBL on NBME I scores, which yield a negative mean effect size of -0.15. Further analyses result that study design and PBL experiences have a significant effect on the impact of PBL on NBME I scores. Furthermore, nine studies providing effect sizes for NBME II scores result in a positive but small effect size of 0.16 favoring PBL over traditional curricula. As a result, they conclude that PBL shows higher performance on clinical science outcomes (NBME II) but lower performance on the basic science outcomes (NBME I). They also propose that well designed research studies conducted in settings with more experience with PBL bring about better NBME I scores.

Colliver (2000), in his review, summarizes three previous research syntheses mentioned above and concludes that three reviews show no persuasive evidence for the effectiveness of PBL on either of knowledge acquisition or clinical skills based on the proposition he claims unreasonably that an effective teaching method is expected to show a large effect of 0.80 or even 1.00. His reasoning for this argument results from the conclusion derived by Bloom (1984) about the optimum effect size that an educational intervention may show. Bloom accepts one-to-one tutoring as the gold standard of teaching methods and as a result of comparisons between one-to-one tutoring and standard classroom teaching, he reaches an effect size of 2.0 favoring the tutoring. Thus, Colliver infers that PBL, as an attempt to approximate the ideal teaching method, is expected to be at least half as effective as the Bloom's optimal instructional method. Consequently, he deduces, after examining some additional randomized studies as well, the effectiveness of PBL is not at expected level at all. However, Albanese (2000) criticizes what Colliver claims about the expected effect size from PBL interventions asserting that the effect sizes of larger than 0.8, which requires some students move from bottom quartile to the top half of the class, is not a reasonable expectation. He underlines that many medical interventions commonly used are based upon the studies with effect size below 0.50, which is accepted as average effect size in the literature. Similarly, Norman and Schmidt (2000) clearly disapprove of Colliver's overall conclusion about PBL stating "PBL does provide a more challenging, motivating and enjoyable approach to education. That may be a sufficient raison d'etre, providing the cost of implementation is not too great" (p. 727) although they agree with the idea proposed by Colliver that PBL does not have remarkable impact on cognitive outcomes.

Furthermore, Smits et al. (2002) take attention to the lack of research synthesis investigating the effectiveness of PBL in postgraduate and continuing medical education in the literature and integrate the results of six studies for this aim. Consequently, he infers that there exists limited evidence that supports the proposition that PBL has more positive impacts on either participants' knowledge, performance or patients' health than other educational strategies do in continuing medical education. It should be noted as a cautionary point that the teaching method

used in control groups of the included studies is not constant, which makes the results of studies less comparable. They also note that "studying the effectiveness of education is complex but we should be able to perform studies of higher quality than those reviewed here, especially when comparing educational methods" (p. 155).

In a well designed meta-analysis examining not only the overall effects of PBL on knowledge and skills but the potential moderators which affect the effectiveness of PBL on these outcomes as well, Dochy et al. (2003) integrate 43 studies, which results in vigorous weighted effect size of 0.460 favoring PBL on the application of knowledge, i.e., skills. However, in terms of knowledge base of students, overall weighted effect size of -0.223 favors conventional learning environment rather than PBL, which they think, may result from two outliers since the combined effect size approaches zero when the analysis is performed by excluding these outliers. Vote counting method supports their ideas about non-robustness of the findings for knowledge base by presenting non-significant results. Finally, as a result of moderator analyses, they conclude that PBL students build up somewhat less knowledge but they keep in mind more of the acquired knowledge and the expertise-level of students is related to the effectiveness of the problem based learning for both knowledge and skills outcomes.

R. A. Smith (2003) performs another meta-analysis, in the scope of his dissertation, to investigate overall effectiveness of PBL on cognitive and non-cognitive outcomes in medical education. Integrating 82 studies conducted in the time interval of 1977 to 2002 to compare the effectiveness of PBL and lecture based learning in 45 universities from 12 countries. Table 2.3 illustrates the corresponding number of studies, unweighted and weighted mean effect sizes with confidence intervals for each outcome. According to the results of the meta-analysis, PBL shows positive impact on all outcomes examined in the scope of this meta-analysis except for biomedical achievement, on which the overall effect of PBL can assumed to be identical to the impact of lecture based learning. However, subgroup analysis shows that the results for biomedical achievement are moderated by assessment type indicating small but significant and positive effect size (d= 0.07) for locally constructed tests (mostly criterion-referenced assessments) while

there exists a negative mean effect size (-0.06) for standardized (licensing) tests. More specifically, the mean effect size increases to 0.13 when multi-method local tests are used as assessment instrument rather than tests including only multiple choice questions.

Table 2.3 *Main effects of problem-based versus lecture-based learning (R. A. Smith, 2003)*

Outcome	N	Mean Effect Size		95% CI	
		Unweighted	Weighted	Lower	Upper
Biomedical achievement	33	-0.01	0.02	-0.02	0.06
Clinical achievement	29	0.31	0.32	0.27	0.38
Problem solving	10	0.30	0.19	0.03	0.34
Self-directed learning	19	0.54	0.47	0.39	0.55
Attitude toward learning	30	0.52	0.45	0.39	0.52

In another meta-analysis study, Gijbels et al. (2005) examine the influence of assessment on the effect sizes calculated to estimate the extent to which PBL is effective. They indicate three levels of knowledge structure assessed in PBL studies, the first one of which is "understanding concepts" while second and third levels are "understanding of the principles that link concepts" and "linking of concepts and principles to conditions and procedures for application" correspondingly. Investigating 40 studies met the inclusion criteria; they conclude that PBL has the greatest positive effects on second level of knowledge structure, i.e. understanding principles that link concepts, with a large effect size of 0.795. They also declare that PBL has small positive effect of 0.339 on the application level of knowledge structure while there is no significant effect of PBL over traditional methods on the first level, i.e. understanding concepts (d= 0.068). As a result, they call attention to that significance of assessment must be taken into consideration while investigating the effectiveness of PBL and possibly in all comparative educational research.

In the most recent meta-analysis, Walker and Leary (2009) not only compare the effectiveness of PBL and traditional instruction but examine the

effectiveness of PBL across different disciplines including medical education and science, implementation types and assessment levels as well. They integrate 82 studies, most of which are from medical education (61) while there are small numbers of studies from teacher education (1), social science (3), allied health (5), business (3), science (8) and engineering (3). According to the results of the meta-analysis, PBL shows the largest impact in the discipline of teacher education with a medium effect size of 0.64 while the effect of PBL can be assumed to be identical to lecture-based approaches in engineering with a very small effect size of 0.05 and in science with again a slight effect size of 0.06. However, the existence of very large discrepancies across the number of studies included from different disciplines makes the findings highly non-robust in terms of discipline analysis. For example, teacher education in which PBL shows the largest impact is represented by only one study whilst there exist 133 outcomes from 61 studies in medical education. Similar problem arises again in the analysis of PBL efficiency across implementation types cited by Barrows (1986) since information about the implementation type is provided in only three studies while the rest is labeled as missing. The results for assessment levels seem to be parallel with the findings of Gijbels et al. (2005) indicating PBL is effective in the second (principle) and third (application) level of assessments with medium effect sizes of 0.205 and 0.334 correspondingly while it shows negative but small effect size of -0.043 for the first level of assessment (concept).

Besides the meta-analyses conducted to effectiveness of PBL on different outcomes, there are some (second order) research syntheses which aim to resolve conflicting situation about how effective PBL is by combining the results of different research syntheses in the literature.

Firstly, Prince (2004) examines the effectiveness of active learning by analyzing previous research syntheses conducted to investigate different types of active learning like collaborative learning, cooperative learning and PBL. He takes attention to the difficulties in integrating different studies scrutinizing the efficiency of PBL by stating the large variation in the ways PBL performs. He underlines that "for meta-studies of PBL to show any significant effect compared to traditional programs, the signal from the common elements of PBL would have to be greater

than the noise produced by differences in the implementation of both PBL and traditional curricula" (p.6). He analyzes the efficiency of different components of PBL like self-directed and inductive learning, some of which result in a positive effect size with differences in magnitude while some others present negative or nonsignificant effect sizes, which may explain, he claims, weak impact of PBL on students' achievement as measured by exams. Finally, he indicates the strong evidence suggesting that PBL is significantly effective on different outcomes like students' attitude and knowledge retention.

Similarly, Strobel and Van Barneveld (2009) emphasize the "heated debate" about the effectiveness of PBL and conduct a meta-synthesis, which is defined as "a qualitative methodology that uses both qualitative and quantitative studies as data or unit of analysis" (p.46). They include eight systematic reviews or meta-analyses, each of which constitutes the unit of analysis of this meta-synthesis. They group outcomes based on their level of assessment into four categories, which are "non-performance, non-skill oriented, non-knowledge based assessment", "knowledge assessment", "performance or skill based assessment", and "mixed knowledge and skill-based assessment". For the first category, which covers "student and faculty satisfaction" and "first choice of residency", reported effect sizes consistently favors PBL rather than traditional instruction. Knowledge assessment outcomes generally tend to favor traditional approaches except for the ones focusing on long-term knowledge retention, which consistently favor PBL. For the last two categories, the results again provide evidence that PBL is more effective than traditional approaches. To sum up, based on the results of this meta-synthesis, Strobel and Van Barneveld claim that PBL is superior on the outcomes of long-term retention, skill development and satisfaction of students and teachers, while traditional approaches are more effective on shorter knowledge acquisition as measured by standardized board exams. Finally, they take attention that the meta-analyses covered in the scope of this meta-synthesis mainly based on the primary studies conducted in the discipline of medical education.

Finally, Hattie (2009) synthesizes over 800 meta-analyses investigating the effect of different variables on achievement, one of which is teaching method variable including PBL. He integrates eight meta-analyses investigating the

effectiveness of PBL by covering 546 effect sizes totally, which result in an overall mean effect size of 0.15. However, it should be noted that "achievement" with the meaning in this meta-synthesis is a very broad term including different outcomes like knowledge acquisition, application of knowledge, self directed learning, attitude towards learning etc... Hattie summarizes the results that for surface level of knowledge, PBL seems to be limited even less effective than traditional approaches. On the other hand, PBL has positive effects on students' learning in terms of deeper level of understanding, which is not unexpected, he claims, since PBL puts more emphasis on meaning and understanding rather than recalling or acquisition of surface level knowledge.

To sum up, 10 syntheses including narrative reviews and meta-analyses and three meta-syntheses summarized above provide us with a comprehensive data about to what extent PBL is effective as an educational tool. However, there is still a clear need for further well designed primary research and research syntheses to resolve the contradictory results about PBL as underlined some of the meta-analyses (Smits et al., 2002; Strobel & Van Barneveld, 2009). More importantly, each of the reviews about the effectiveness of PBL in the literature is mainly, if it is not completely, based on the studies conducted in medical education. There is only one meta-analysis covering only eight studies examining the impact of PBL in science education although recently PBL has become widespread in science education as well, which results in a proliferation in research studies waiting to be synthesized.

2.10 Summary of the Findings of the Related Studies

The results of the related literature can be summarized as follows:

- It is strongly evident from the literature that research syntheses, in particularly meta-analyses, are strongly suggested as they have essential functions to produce cumulative knowledge and to explain the contradictory results about the same research questions (J. Bennett, 2005; Davies, 2000; Dieckmann et al., 2009; Glass, 1976; Rosenthal & DiMatteo, 2001; Torgerson, 2003).

- Although the study of Pearson (1904) can be accepted as the starting point of the quantitative synthesis of the research findings (Lemeshow et al., 2005; Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001), the term of 'meta-analysis' was firstly introduced more recently by Glass (1976), who classifies the data analysis as primary and secondary analysis, and places meta-analysis as an advancement of practice of secondary analysis.
- After the beginning of 1980s, around when the modern era of meta-analysis has began, meta-analysis has gained popularity with an increasing rate in various domains especially in the field of biomedicine and behavioral sciences (Rosenthal & DiMatteo, 2001; Schulze, 2004, 2007; Viechtbauer, 2007).
- Meta analysis, which is "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), has obvious advantages over conventional review, which suffers from serious limitations (Borenstein et al., 2009; Bushman & Wells, 2001; Cooper & Rosenthal, 1980; Glass, 2006; Glass et al., 1981; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001; Oakley, 2002; Petticrew, 2003; Petticrew & Roberts, 2006; Rosenthal & DiMatteo, 2001; Wolf, 1986).
- Besides many advantages of meta-analysis, there exist some criticism of meta-analysis in the literature, which are categorized by Rosenthal and DiMatteo (2001) as 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.
- As the effectiveness of teaching methods is one of the most widely-researched topics, there exist many research syntheses investigating not only the overall effectiveness of a particular alternative teaching method like the studies for CBI (Bayraktar, 2000; Burns & Bozeman, 1981; Christmann, 1997; Clark, 1985; Flinn & Gravatt, 1995; C. C. Kulik & Kulik, 1991; J. A. Kulik, 1983, 1985; J. A. Kulik et al., 1983; J. A. Kulik et al., 1980; Liao,

1999; Niemiec & Walberg, 1985), for cooperative learning methods (Igel, 2010; Jonhson et al., 2000; Qin et al., 1995), for concept mapping as an instructional tool (Campbell, 2009; Horton & Hamelin, 1993; Nesbit & Olusola, 2006), for conceptual change strategies (Guzzetti et al., 1993), and for inquiry based learning (Lott, 1983; Minner et al., 2009; D. Smith, 1996; Sweitzer & Anderson, 1983) but also the relative effectiveness of different alternative teaching methods (D. R. Anderson et al., 1983; Haas, 2005; Marcucci, 1980; Marzano, 1998; Marzano et al., 2001; Schroeder et al., 2007; Wise, 1996; Wise & Okey, 1983).

- The existing meta-analysis studies comparing different teaching strategies have some limitations in the classification of teaching strategies: none of the classifications have been developed in a systematic way; they do not differentiate neither 'medium and method' nor 'method and technique'; and the methods involved in the classifications are not familiar with the ones stated in the literature.
- Theoretical background of PBL, which has been developed by different researchers (Ausubel et al., 1978; Bruner, 1961; Dewey, 1938; Piaget, 1954; Vygotsky, 1978), has a long history (Dochy et al., 2003; Gijbels et al., 2005; Hmelo-Silver, 2004). PBL, as a teaching method or curricular approach, has many roots on different educational theories (Savin-Baden & Major, 2004), the most prominent of which is constructivism (Akçay, 2009; Chin & Chia, 2005; Duffy & Cunningham, 1996; Gijselaers, 1996; Hmelo-Silver, 2004; Savery & Duffy, 2001; Taşkesenligil, 2008).
- It is evident from the literature that the effectiveness of PBL on different outcomes is a controversial issue (Strobel & Van Barneveld, 2009) since it is both possible to find studies favoring PBL (Sungur et al., 2006) and the ones favoring direct instruction (Scott, 2005).
- Several research syntheses have already been conducted to resolve the conflicting results of primary studies examining the effectiveness of PBL (Albanese & Mitchell, 1993; Berkson, 1993; Colliver, 2000; Dochy et al., 2003; Gijbels et al., 2005; Kalaian et al., 1999; R. A. Smith, 2003; Smits et

al., 2002; Vernon & Blake, 1993; Walker & Leary, 2009). There are also three studies, which aim to synthesize the results of existing research syntheses (Hattie, 2009; Prince, 2004; Strobel & Van Barneveld, 2009).

- Although these research syntheses provide us with a comprehensive data about to what extent PBL is effective as an educational tool and some explanations for the contradictory nature of the findings, there is still a clear need for further well designed primary research and research syntheses to make more consistent generalizations about the effectiveness of PBL on different outcomes. More importantly, each of the reviews about the effectiveness of PBL in the literature is mainly, if it is not completely, based on the studies conducted in medical education. There is no research synthesis, which focuses on the impact of PBL on different outcomes in science education.

CHAPTER III

METHODOLOGY

This chapter includes an overview of meta-analysis, comparison of fixed-effect and random-effects model, validity issues including publication bias and quality of primary studies, acquisition of studies covering criteria for inclusion of studies, main steps and results of literature search, coding process comprising development of coding sheet and coding manual, coding of primary studies and coding reliability and further statistical issues covering heterogeneity, moderator and power analyses, effect size indices, unit of analysis and finally software used for statistical analyses.

3.1 An Overview of Meta-Analysis

Meta-analysis is "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). Rosenthal and DiMatteo (2001) highlight what Glass (1976) states about the scope of meta-analysis; it 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 PBL on students' achievement, motivation in science and attitudes towards science.
- Collect and select the 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 o 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.

There are two main statistical models 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 size from a collection of studies (Borenstein et al., 2009; Hedges & Vevea, 1998; Hunter & Schmidt, 2000, 2004; Tweedie, Smelser, & Baltes, 2004).

3.2 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 (Borenstein et al., 2009).

Since all factors that may influence the effect size are assumed to be constant in fixed-effect model, the observed effect (Y_i) for each study is calculated by population mean (θ) and sampling error (ε_i) as

$$Y_i = \theta + \varepsilon_i$$

In Figure 3.1 and Figure 3.2, ▼ symbol is used for combined true effect, while ● represents study true effect, and ■ shows the observed effects for each study. Figure 3.1 presents an example for distribution of sampling error within a fixed-effect model.

In contrast, 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 (ζ_i) resulting from between study variance.

$$Y_i = \mu + \zeta_i + \varepsilon_i$$

Figure 3.2 shows an example of between study and within study variance within a random-effects model.

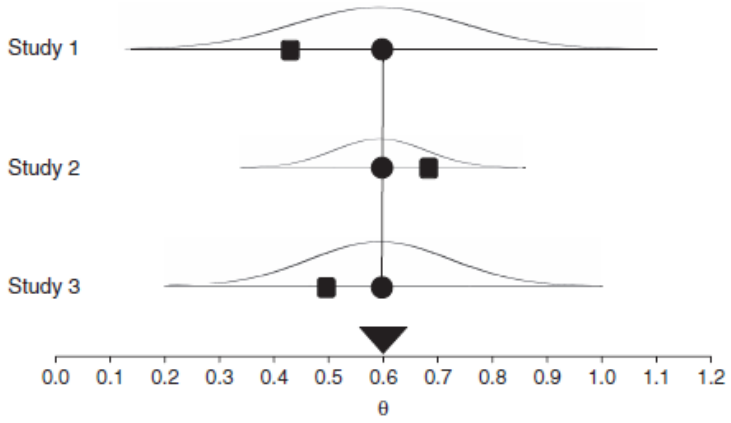

Figure 3.1 Distribution of sampling error in fixed effect model (Borenstein et al., 2009)

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.

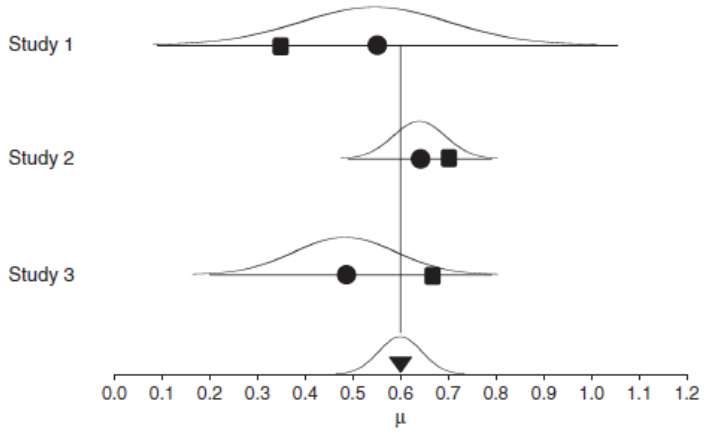

Figure 3.2 Between study and within study variance within a random-effects model (Borenstein et al., 2009).

There exists an important distinction between fixed-effect and random-effects models in terms of estimating the summary effect. Since the main aim is to predict one true effect size, the information from small studies is underestimated, assigning more weight to larger 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, whether small or large, 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.

It is evident from the literature that fixed-effect model have been much more widely used in the meta-analyses conducted until recently (Cooper, 1997; Hunter & Schmidt, 2000; National Research Council, 1992; Overton, 1998; F. L. Schmidt, Oh, & Hayes, 2009) although 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 yields 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; F. L. 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, 2003b; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Overton, 1998; F. L. Schmidt et al., 2009) take attention to the fact that fixed-effect model leads to raised Type I error rates for statistical tests when homogeneity assumption is not met. Hunter and Schmidt (2000) illustrate how Type I error rate is affected by heterogeneity and average sample size of the studies included in the meta-analysis in Table 3.1, which shows 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

Table 3.1 *Type I error rates for the random-effects and the fixed-effect significance test for the mean correlation in meta-analysis (nominal α = .05 and SDρ is the standard deviation of population correlations across the studies included in the meta-analysis) (Hunter & Schmidt, 2000)*

The random-effects model significance test						
Prob (Type I error) = .05 in all cases						
The fixed-effect model significance test						
Study sample sizes	Homogenous case ($SD_\rho = 0$)	Heterogeneous cases ($SD_\rho > 0$)				
		SD_ρ				
		.05	.10	.15	.20	.25
25	.05	.06	.08	.16	.16	.22
100	.05	.08	.28	.38	.38	.46
400	.05	.17	.53	.63	.63	.70
1600	.05	.38	.75	.81	.81	.85
…						
∞	.05	1.00	1.00	1.00	1.00	1.00

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, F. L. 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, 2003b; Hunter & Schmidt, 2000; National Research Council, 1992).

In this meta-analysis, random-effects model is used to calculate the overall effect size for the effectiveness of PBL since not only the findings of the primary studies are highly inconsistent but identification of the generalizable conclusions is the main purpose of this study as well. Furthermore, Borenstein et al. (2009) indicate that there is no cost to using random-effects model since it reduces to fixed-effect model if the between-studies variance is zero.

3.3 Validity Issues in This 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 and how they are controlled in this meta-analysis study.

3.3.1 Publication Bias

It is evident from the literature that publication bias, or "file-drawer problem", is one of the most serious problem 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. They describe publication bias as "the term for what occurs whenever the research that appears in literature is systematically unrepresentative of the

population of completed studies" (p. 1). 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 3.2 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 indicate 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; F. L. 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 3.2 *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)

Publication bias threat is not specific 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. Forest plots, funnel plots, Rosenthal's 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; S. Lewis & Clarke, 2001; Sterne & Egger, 2001; 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 unique strengths and weaknesses, several methods explained in following sections are used in this meta-analysis for diagnosis of publication bias and to estimate the extent to which it affects the results.

3.3.1.1 Forest Plots

Borenstein (2005) asserts that forest plot as the visual representation of the data is a key element in any meta-analysis. Figure 3.3 shows an example of forest plot with Hedge's g for effect size estimates from 16 studies investigating the effect of PBL on critical thinking skills. In Figure 3.3, 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 the 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, analyzing this plot is a logical first step in any analysis (Borenstein, 2005) because a forest plot not only provides the readers with the information of individual studies in the meta-analysis at a glance but also

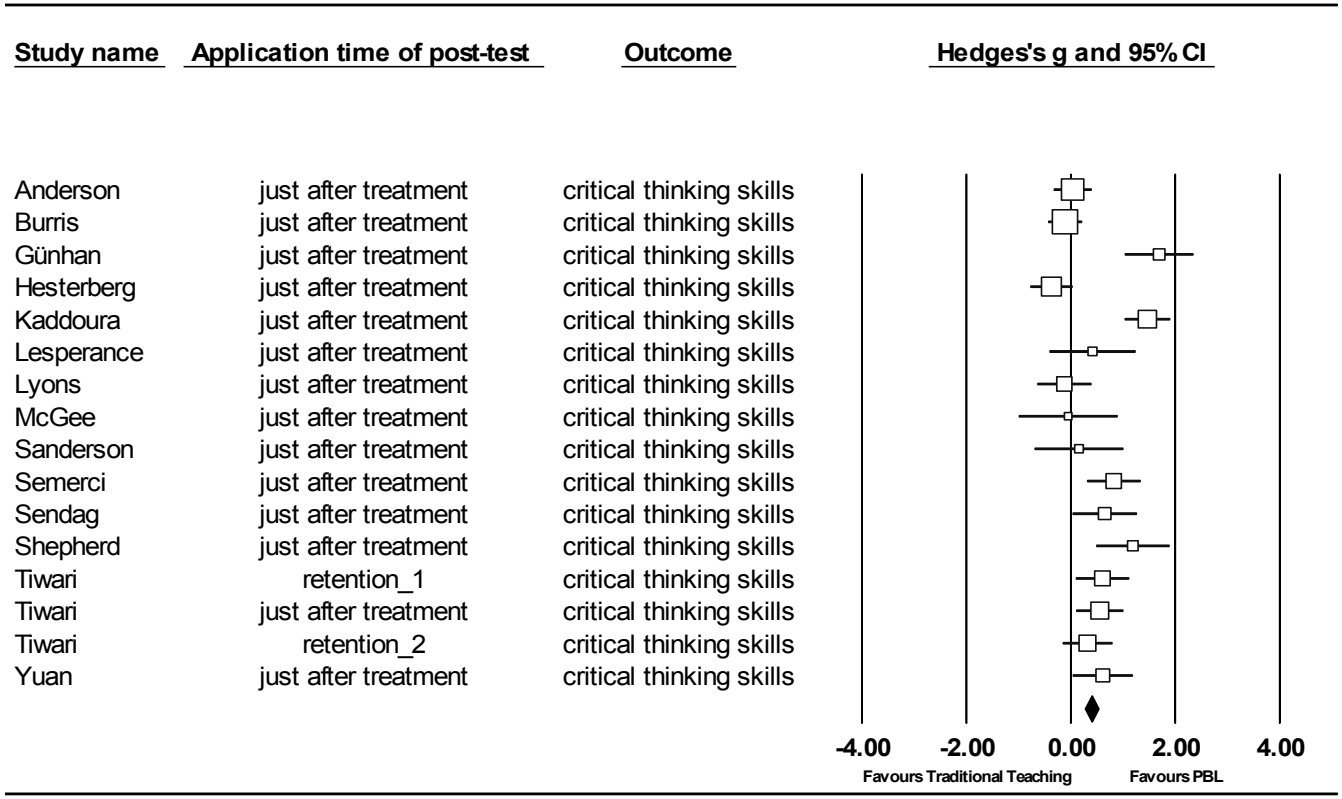

Figure 3. 3 An example of forest plot showing Hedge's g with 95% confidence intervals for 16 studies investigating the effect of PBL on critical thinking skills

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).

3.3.1.2 Funnel Plots

Funnel plots are simple scatterplots of effect sizes estimated from each study against a measure of study 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 studies will scatter widely at the bottom of the plot with smaller spread at the top as a result of larger studies. Thus, in the absence of any bias, the plot is expected to resemble a symmetrical inverted funnel as shown in Figure 3.4.

Conversely, if there is a publication bias, generally a skewed and asymmetrical spread is expected on the funnel plots as shown in Figure 3.5. 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 if there would be no bias.

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; Sterne & Harbord, 2004; Terrin, Schmid, & Lau, 2005). In addition, Tang and Liu (2000) claim that 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.

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.

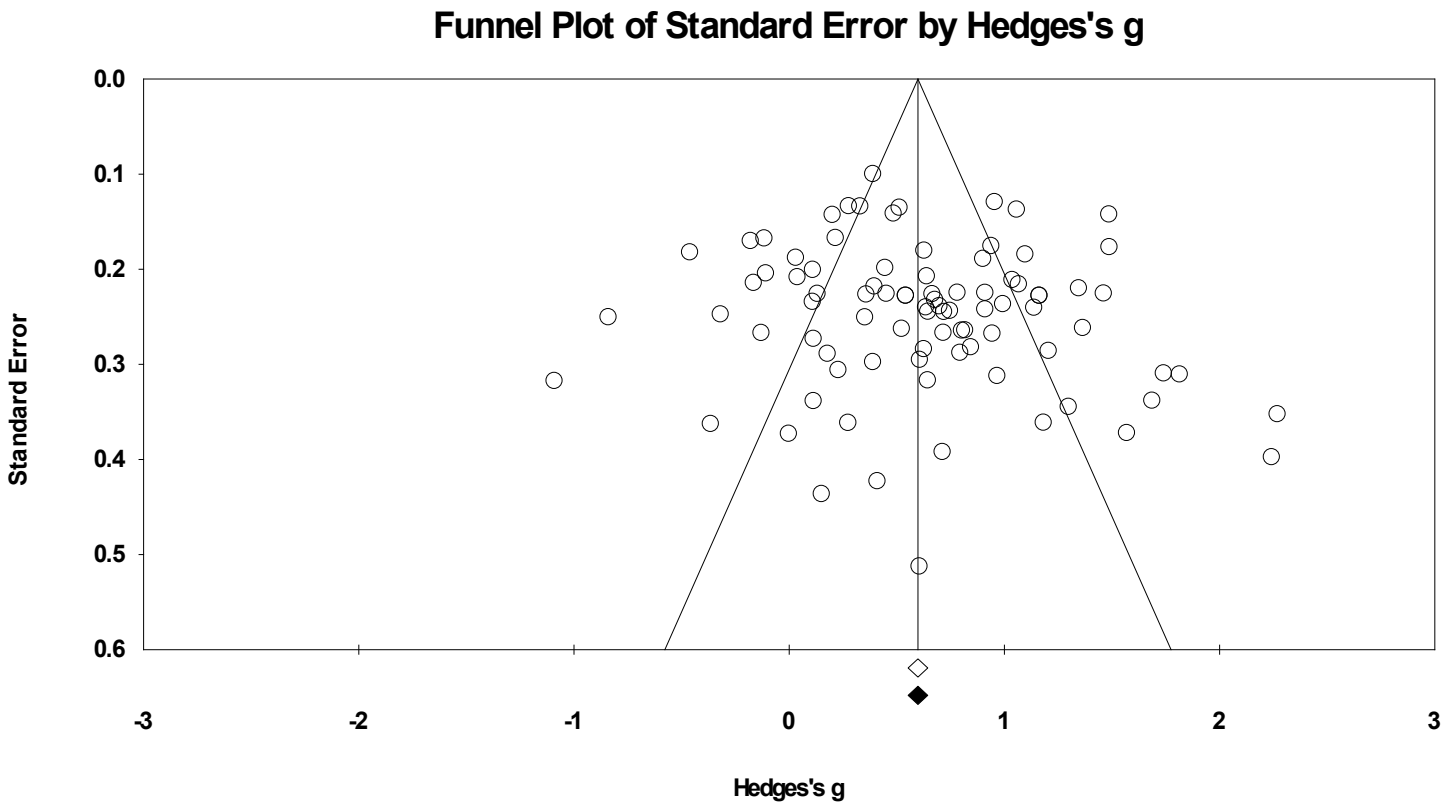

Figure 3.4 A symmetrical funnel plot without bias

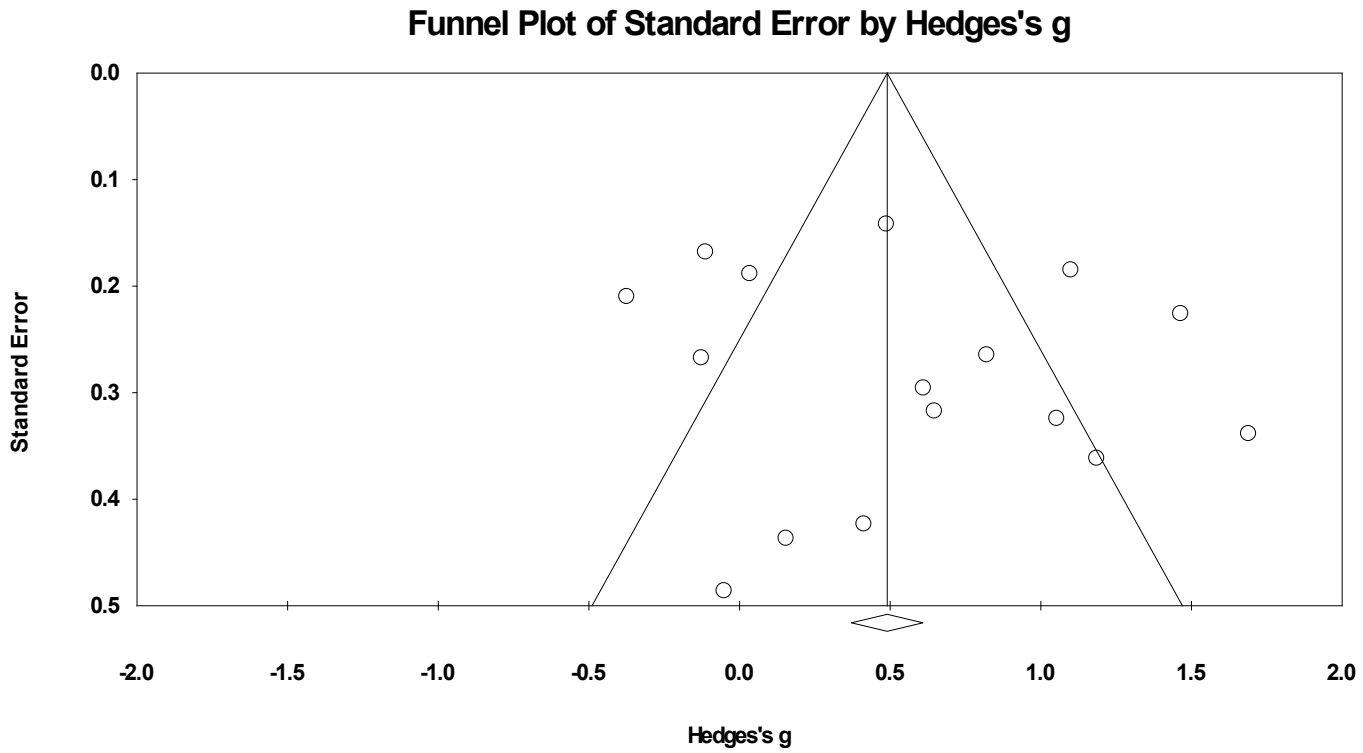

Figure 3.5 An asymmetrical funnel plot with a possible bias

3.3.1.3 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 = standardized mean difference/standard error) is regressed against its precision (prec= 1/standard error) (Sterne & Egger, 2005).

$$E[z_i] = \beta_0 + \beta_1 prec_i$$

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 a 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.

3.3.1.4 Rosenthal's Fail-safe N

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, 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 evidence seems to be sufficiently robust for future studies.

Table 3.3 illustrates an example of for Rosenthal's FSN calculations conducted for six studies investigating the effect of PBL on creativity. 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 3.3 *An example of output for Rosenthal's FSN calculations conducted for six studies investigating the effect of PBL 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

3.3.1.5 Orwin's Fail-safe N

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 3.4 illustrates an example of Orwin's FSN calculations conducted for the same studies in the previous example for Rosenthal's FSN. Results show that 370 additional studies with null effect are needed to bring the overall effect to the effect size value of 0.01, 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 3.4 *An example of output for Orwin's FSN calculations conducted for six studies investigating the effect of PBL on creativity*

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

3.3.1.6 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 3.6 represents an example of funnel plot in which Trim and Fill adjustment is taken into account for a similar data in the previous figure. 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.

3.3.2 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

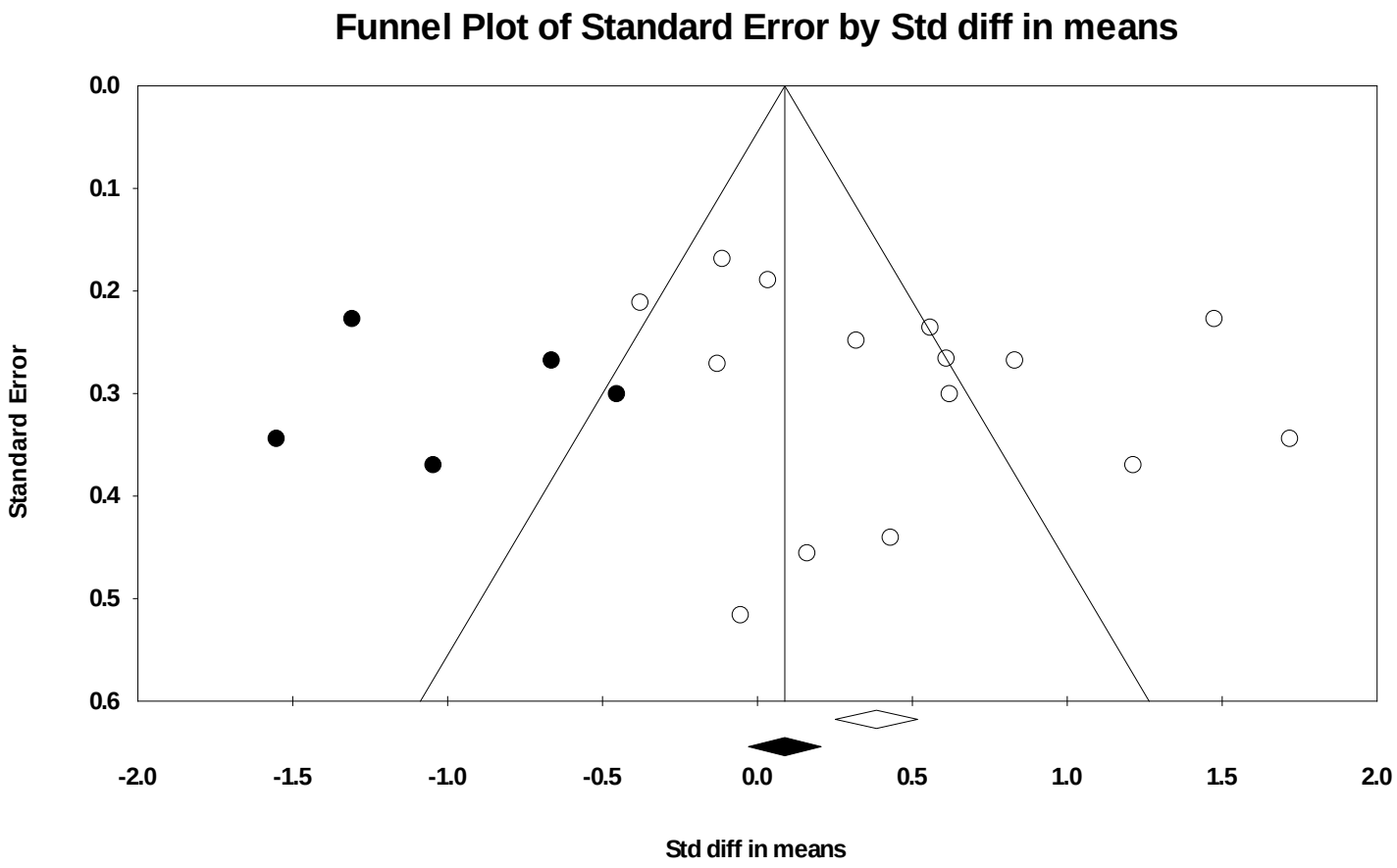

Figure 3. 6 An example of funnel plot with the studies imputed by TFM, resulting in an adjusted effect size

"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.

In this meta-analysis, study quality is not used as one of the exclusion criteria as it is suggested by Glass (2006). Instead, the impact of study quality indicators is investigated by assigning them as moderator variables separately. Internal and external validity issues are used as study quality indicators. Internal validity refers to "the validity of inferences about whether some intervention has caused an observed outcome" (Valentine, 2009, p. 130). In other words, it means that "observed differences on the dependent variable are directly related to the independent variable and not due to some other unintended variable" (Fraenkel & Wallen, 2000, p. 190). On the other hand, external validity refers to "how widely a causal claim can be generalized from the particular realizations in a study to other realizations of interests" (Valentine, 2009, p. 130).

Fraenkel and Wallen (2000) summarize possible threats to internal validity as subject characteristics, mortality; i.e. loss of subjects, location, instrumentation, testing, history, maturation, attitude of subjects, regression and implementation, all of which are included in an explicit item in the coding sheet to score internal validity in the scope of this meta-analysis. However, there was very limited information provided in primary studies about whether these threats had been controlled, which does not necessarily mean that the researchers had not done anything for these threats. As explained in the first paragraph, study quality and reporting quality interfere for these situations and it is impossible to distinguish which one results into providing no information about internal validity. Thus, this item was not used as an indicator of study quality. Rather, some other properties of

primary studies like research design and model, teacher effect and researcher effect were used as indicators of internal validity of the studies and each indicator was included in moderator analysis separately.

Research design item includes options of true experiment, quasi-experiment with randomly assigned clusters and quasi-experiment without randomly assigned clusters. Random assignment, which is a way of controlling threats to internal validity, is essential for true experimental studies; however, it is not very convenient in educational studies. What is more convenient and common in educational settings is to assign the clusters (i.e. classes) randomly to any of experimental or control group conditions, which is still better to control threats to internal validity comparing to the lack of any random assignment.

In research model item the coder is expected to select appropriate research model for the primary study under examination. Post-test only control group design, pre-test post-test control group design, Solomon four group design and factorial design comprise some options for this item. Fraenkel and Wallen (2000) claim that different research models have different effectiveness in controlling the threats to internal validity. For example, if there is no random assignment, application of pre-test provides some control over subject characteristics threat while counterbalanced and factorial design are much more effective for maturation and regression threats to internal validity.

For the item examining teacher effect, the coder is asked to indicate whether both control and experimental groups have been instructed by the same teacher. Similarly, in another item researcher bias is investigated by asking whether researcher has been involved in the study as a teacher. Furthermore, whether the length of treatment is same for both experimental and control groups has also been coded to check its effect on internal validity. However, no study has been reported as the length of treatment was different, therefore, this item could not be involved in the moderator analysis.

On the other hand, sampling method is coded as an indicator of population generalizability, which is an essential part of external validity. In one of the items on the coding sheet, the coder is asked to choose the most appropriate sampling method for the study coded. If sampling procedure consists of more than one stage

and different sampling methods are used in different stages, the most nonrandom one in any of these stages is selected. Many of the studies provided information about sampling procedure, so this item could be coded for the majority of the studies. However, the studies were mostly loaded on the sample method of convenience sampling.

3.4 Acquisition of Studies Included in This Meta-Analysis

3.4.1 Criteria for Inclusion of Studies

For the selection of the studies, some rigorous criteria were employed. Only the studies having the following characteristics were included in the meta-analysis:

- True experimental or quasi experimental studies
- The studies in which PBL is implemented in experimental group while traditional teaching method is used in control group
- The studies published in the time interval of January 1, 1990 and June 1, 2012
- The studies conducted in the school level of elementary, secondary, college, and higher education.
- The studies in the subject area of science, in which student achievement, motivation in science (physics, chemistry, biology, or general science) or attitudes towards science are assigned as dependent variables
- Independently from subject area, the studies in which skills or creativity are assigned as dependent variables
- The studies in which effect size, or the statistics necessary to calculate effect size (means, standard deviations etc…) are reported

Furthermore, no study was included more than once. If a study revealed more than one effect sizes for the same dependent variable, the average effect size was calculated and only one effect size was placed for each dependent variable in each study.

3.4.2 Main Steps of the Literature Search

A broad search was conducted to locate all related journal articles, dissertations, and theses investigating the effectiveness of PBL comparing to

traditional teaching methods in the literature. Firstly, a stepwise search was conducted by means of comprehensive electronic databases, the first one of which is EBSCOHOST, which covers a collective list of databases related to variety of subjects. 34 optional databases listed in the scope of EBSCOHOST were analyzed in terms of their coverage and 12 databases were chosen to be included in the search. These databases were Education Research Complete, Academic Search Complete, ERIC, PsycINFO, Professional Development Collection, Psychology and Behavioral Sciences Collection, MasterFILE Complete, SocINDEX with Full Text, Humanities International Complete, PsycArticles, Middle East Technical University's Catalog, and ULAKBIM Turkish National Databases, which also covers four separate databases: Turkish Engineering and Basic Sciences Database, Turkish Life Sciences Database, Turkish Medical Database, and Turkish Social Sciences Database.

The second comprehensive database searched was Web of Science, which covers a set of databases as well including 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).

In addition to these comprehensive databases, ProQuest Dissertations and Theses (PQDT), which covers dissertation and theses from all around the world and provides full texts for many of them, was searched to be able to reach unpublished studies related to PBL as well. Furthermore, another database, National Thesis Center (NTC), which is provided by National Higher Education Council and covers the dissertations and theses completed in Turkey, was also used to reach the studies which were not covered by PQDT.

Then, literature search was extended by means of different sources not to miss out any studies especially from grey literature. Firstly, an electronic search was conducted by Google Scholar for unpublished data. Then, another comprehensive search was conducted to reach the meta-analysis, which was already conducted to investigate the effectiveness of PBL by means of the electronic databases previously mentioned to employ "snowball method", which means to review the references of selected articles to reach additional studies (Dochy et al.,

2003). In addition, The Interdisciplinary Journal of Problem-based Learning, which is an open access journal specific to the all aspects of implementing PBL in K-12 and post-secondary settings, was searched by hand; that is, issue by issue. Furthermore, in case there may be some studies, which could not be reached during searching process, I sent an e-mail to the e-mail group of Science Education in Turkey to ask whether any of the members have studies not indexed in common databases.

Finally, to check how effective the literature search conducted up to that time was, I conducted series of searches with more specific key words and checked if there was any missing study among the resulted list of studies. Beside the searches by databases, I examined randomly selected issues of two national journals, which were Hacettepe University Journal of Education and Journal of Turkish Science Education, and two international journals; Journal of Research in Science Teaching and International Journal of Science Education for the same purpose.

3.4.3 Results of the Literature Search

As explained in the previous section, literature search for acquisition of the primary studies was started with an electronic search by means of comprehensive databases of EBSCOHOST and Web of Science including sets of databases mentioned previously. Figure 3.7 summarizes the basic steps of study acquisition and results obtained in each of these steps. Rather than running a group of narrow searches by using combinations of specific key words, a wider inquiry was conducted by using following phrases including general key words, Boolean operator, wildcards and truncation symbols:

“problem * learning” OR “case based learning” OR “problem# * öğrenme”

Truncation, which is represented by an asterisk (*), makes it possible to reach all forms of phrases including any word instead of truncation symbol like “problem based learning” or “problem solving learning” etc… On the other hand,

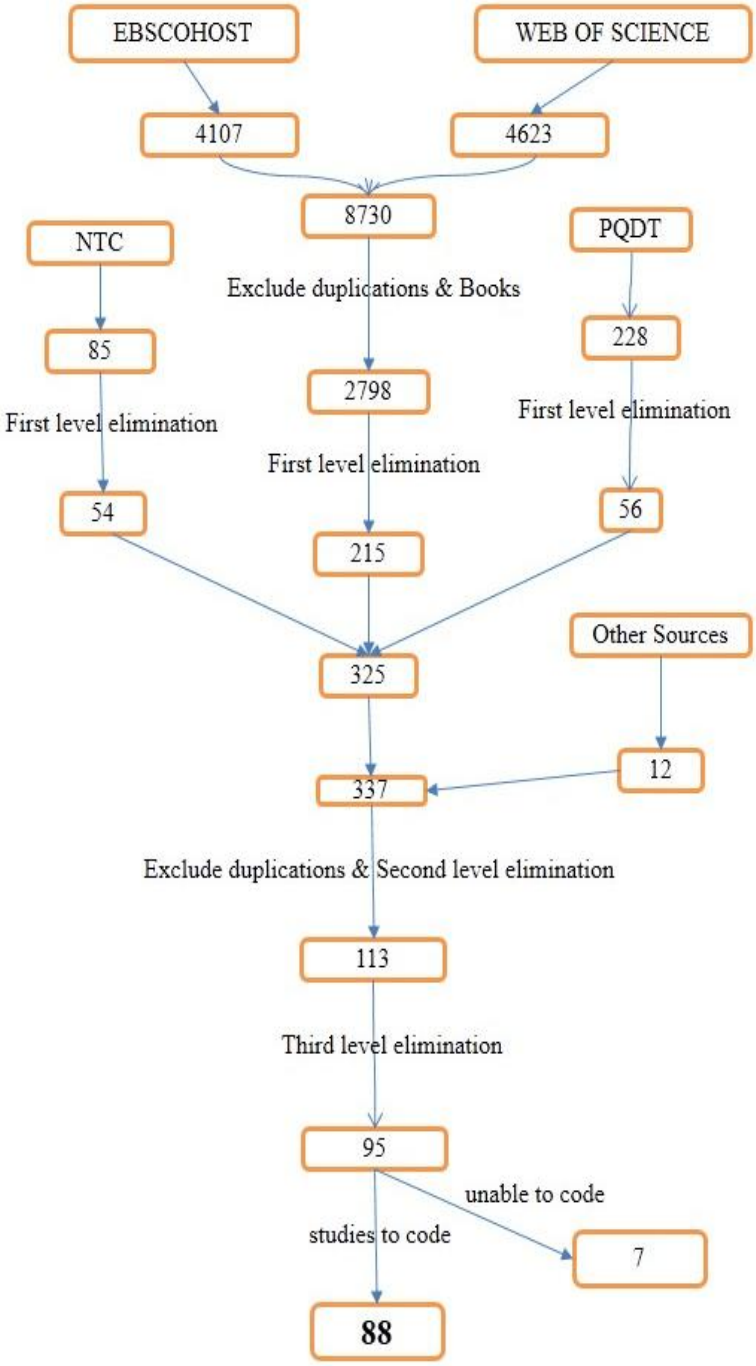

Figure 3.7 Study acquisition process

wildcard, which is represented by (#), is used when alternative spelling may (or may not) contain en extra character. Unlike another wildcard symbol of (?), this symbol allows search engine to return all results with or without any character instead of (#), like “problem merkezli öğrenme”, “probleme dayalı öğrenme” etc… Finally, Boolean operator of (OR) combines search phrases so that each search result contains at least one of the phrases. Since the inquiry results in too many unrelated results without quotation marks, which makes very difficult to distinguish related ones, the search was limited by exact phrases rather than each of the terms in each phrases by using quotation marks. That is, the search would not result in many studies including only one of the terms in keywords like “learning” or “based” etc…

I performed exclusion of studies at three levels, the first one of which was based on the research design and independent variables of the study. That is, the first level elimination was performed to exclude studies which were not either experimental or quantitative, or did not have appropriate independent variables. The studies left after first level elimination were experimental ones with independent variables of PBL, which constitutes experimental group, and traditional teaching method as implemented in control group. At the second level of elimination, the dependent variables of studies were checked. The studies having dependent variables of achievement and motivation in science or attitudes towards science were accepted while the studies focused on achievement, motivation or attitude outputs in the scope of other subjects were excluded. On the other hand, for other outputs rather than achievement, motivation and attitude like skills and creativity, subject area was not set as one of the exclusion criteria. The studies left after second level of elimination were very close to be included in the meta-analysis as one of the primary studies. However, some studies had potential to create lumpiness in the data since some journal articles were based on the dissertation or theses that were included in the selected group as well. Thus, third level of elimination was conducted for the studies which were based on the same sample. Generally, dissertations and theses were chosen to be included in the meta-analysis since it was possible to reach much more detailed information in dissertations or theses comparing to journal articles. However, for some cases, journal articles were more

informative about what was coded in the coding sheet especially for the items to calculate effect size of the study. For these cases, only journal articles were included in the meta-analysis while corresponding dissertation or theses were excluded.

As illustrated in Figure 3.7, the search by means of EBSCOHOST and Web of Science resulted in 8730 studies totally, which, however, included many duplications because of common databases to be covered. After excluding these repetitions and books, 2798 studies were left for first level elimination. After first level of elimination was conducted, there were only 215 studies to be included for further steps of elimination.

On the other hand, the inquiry performed by NTC and PQDT resulted in 85 and 228 studies respectively. 110 studies, which were left totally for these databases after first level of elimination, were added to 215 studies from EBSCOHOST and Web of Science, resulting in 325 studies totally.

When five studies resulted from Google Scholar, six studies obtained by means of snowball method, and one study sent by a member of Science Education in Turkey e-mail group were added, the number of studies came up to 337. It should be noted that the numbers for other sources represent the number of studies, which had not been reached up to that time. For example, four studies were sent by e-mail group members, but three of them had already been included in the sample for meta-analysis.

After exclusion of duplications resulting from studies, which came from three different sources (NTC, PQDT, and comprehensive databases) and conducting second level elimination, only 113 studies were left to be involved in the meta-analysis. Then, third level elimination was performed not to allow the studies with the same samples to be included in the meta-analysis more than once as dissertation or theses and journal article in order to prevent any lumpiness in the data.

As a result, 95 primary studies were coded in this meta-analysis; however, seven studies could not be involved in the analyses since the appropriate information to calculate related effect sizes were not reported in these studies.

While conducting literature search for acquisition of studies to be included in the meta-analysis, it is not easy to decide the correct time to stop doing further

search. To be able to check whether further literature search is necessary or not, I conducted narrow searches by using a combination of more specific sets of keywords. The first set included similar keywords used previously: “problem * learning”, “case based learning”, and “problem# * öğrenme”. The second set was composed of keywords related to the subject areas: “science”, “general science”, “physics”, “chemistry”, and “biology”. The third set was comprised of the terms related to dependent variables of the study: “achievement”, “academic achievement”, “motivation”, “attitude towards science”, “attitude towards physics”, “attitude towards chemistry”, “attitude towards biology”, “critical thinking skills”, “problem solving skills”, “science process skills” etc..

Randomly selected pages for results of the search conducted these keywords and Boolean operators of (OR) and (AND) were examined whether there were any further studies which were not included in the sample of this meta-analysis although they obeyed exclusion criteria. In addition, as previously mentioned, random issues of four journals were examined for the same purpose. As a result, I decided to stop conducting literature search since no further study appeared to be included in meta-analysis as a result of this process. Consequently, 147 effect sizes from 88 studies constituted the sample of this meta-analysis.

3.5 Coding Process

3.5.1 Development of Coding Sheet and Coding Manual

A coding sheet, which is the fundamental instrument in a meta-analysis, includes several items about the primary studies to be coded to gather information about critical characteristics of the studies, which are related to either the calculation of effect size or moderator variables having potential influences on effect size. On the other hand, coding manual includes specific instructions about how to code each item in the coding sheet.

The coding sheet and coding manual for this meta-analysis were developed throughout the following steps:

- Several coding sheets developed in other meta-analyses were examined.
- Research syntheses previously conducted to investigate the effectiveness of PBL were examined to identify possible moderator variables.

- First draft of the coding sheet consisting of 33 items was developed.
- Four experts in the field of educational studies provided feedback about the first draft of the coding sheet.
- Second draft of the coding sheet was developed based on the feedback provided about the first draft.
- Pilot coding of 33 studies was conducted by using second draft of coding sheet.
- Opinions of the members of thesis monitoring committee were taken about the coding sheet.
- Revisions were made to develop third draft of coding sheet.
- First draft of coding manual was developed based on the third version of the coding sheet.
- Thesis supervisor and two other experts provided feedback about third draft of coding sheet and first draft of coding manual.
- Final versions of the coding sheet and coding manual were developed on the basis of feedback provided about the previous version.

The first step to develop the coding sheet was to examine coding sheets developed in other meta-analysis studies. Seven coding sheets from different meta-analyses (Bayraktar, 2000; Campbell, 2009; Igel, 2010; Onuoha, 2007; M. C. Şahin, 2005; D. Smith, 1996; Tinoca, 2004) were checked item by item to decide which items are appropriate to be included in the coding sheet which would be developed in the scope of this meta-analysis. Then, research syntheses previously conducted to investigate the effectiveness of PBL (Albanese & Mitchell, 1993; Berkson, 1993; Colliver, 2000; Dochy et al., 2003; Gijbels et al., 2005; Kalaian et al., 1999; R. A. Smith, 2003; Smits et al., 2002; Vernon & Blake, 1993; Walker & Leary, 2009) were examined to identify the study characteristics that may affect the effectiveness of PBL; i.e. possible moderator variables and the first draft of coding sheet (Appendix A) was developed.

Next, four experts; two professors in the field of educational studies and one senior researcher having PhD in mathematics education with one PhD candidate in medical education, who has meta-analysis experience, investigated the first draft of

coding sheet and provided feedback with some suggestions for new items. Based on the discussions and feedbacks, the second draft of the coding sheet (Appendix B) was developed by adding five new items, which were Item 8 (topic), Item 16 (socio economic status), Item 20 (application time of posttest), Item 24 (the use of group work) and Item 25 (group size). In addition, based on the recommendation by thesis supervisor, Item 30 was reorganized in a way that it became much more convenient to code.

Then, a pilot coding was conducted with a sample of 33 studies from the ones to be selected for the meta-analysis, which was highly effective to check which items on the coding sheet worked and which ones did not work well. Based on the experience gained from pilot coding, some revisions were made for some of the items on the coding sheets and new items were added. Firstly, third item (research design) on the second draft of coding sheet was decided to coded in three separate items, which were Item 3 (research design), Item 4 (research model) and Item 5 (sampling method) on the final version of coding sheet (Appendix C). By this way, much more information was provided in terms of internal validity and population generalizability. Furthermore, during pilot coding, I realized that hybrid models of PBL were used in some studies, which was the reason why Item 27 (any method integrated to PBL) was added to the final version. Next, I recognized that PBL is used as a curriculum model rather than specific teaching method especially in medical education, therefore Item 26 was added to the final version to investigate whether how PBL was used affected its effectiveness. In addition, one of the studies (Serin, 2009) coded during pilot coding had used problem situations that students preferred most. Thus, I wanted to check whether there were other studies in which problem statements were aligned with students' interest by adding Item 30. Furthermore, Item 44 was added to check the extent to which the assumptions of effect size estimation have been met in the primary studies and Item 15 was added to record mean age of the participants in the primary studies, which would provide, if it would be coded properly, a continuous variable besides the categorical variable of grade level. Finally, Item 47 was added to compare the results of male and females in terms of related dependent variables. On the other hand, Item 31 (treatment fidelity) was omitted after pilot coding since there were limited

information provided in the primary studies to assess treatment fidelity and it was impossible to distinguish reporting deficiencies and lack of fidelity. Furthermore, different implementation types of PBL made it more difficult to evaluate the treatment fidelity of the studies.

After that, opinions of the members of thesis monitoring committee were taken about the coding sheet. Based on the pilot coding experience and feedbacks provided by the members of committee, third draft of coding sheet was developed. It was not presented in appendix since it was not very different from final version. In the third draft of the coding sheet, besides the changes mentioned in the previous paragraph, some other changes were made. Firstly, the term of "standardized test" in Item 36 was changed as "pre-existing test" based on the idea that not all pre-existing tests could be called as standardized tests. In addition, two items in the second draft of the coding sheet, Items 28 and 29 related to type of outcome were re-arranged as Items 22 and 23 in the final version. In the later version, any types of skills as an outcome was coded separately from achievement referring to content knowledge, because during pilot coding, I realized that achievement was assessed as content knowledge rather than skills, which were evaluated by separate instruments in the primary studies. Then, the option of "both" in Item 19, which was to code type of the assessment instrument, was changed as "adapted" since the later was decided to represent what it was meant by "both" much better. Finally, some changes were made in the format of the coding sheet and in the order of the items to make it easier to use. Underlined blanks were added before each choice in the items of coding sheet to be able to code on the electronic copy more easily and what is more, related items, which were realized during pilot coding to be presented in similar sections in the primary studies, were arranged in a way that they were close to each other in the third draft of coding sheet.

In addition, the first draft of coding manual, which provided explanations about how to code each item in detail, was developed on the basis of the third draft of coding sheet not only to inform other coders about how to use the coding sheet but to set rules for coding each item definitely and explicitly for myself as well. In the coding manual, the instructions started with a clear explanation about what the

coder was expected to do for each item and then some important points were highlighted on the "be aware of that" part when it was necessary.

Next, the third draft of coding sheet with the first draft of coding manual were sent to three experts including thesis supervisor to get further feedback before constructing their final versions. Based on the feedback provided, three more items were added to the coding sheet, one of which, Item 21 on the final version was added to code whether length of treatment was same for control and experimental conditions. In addition, Item 24 was also added to note how dependent variables were measured; that is, they were measured whether by paper-pencil test (if so, which type of questions were used, objective type, open-ended or both) or process and product assessment. The final newly added item was Item 34 on the final version of coding sheet, in which background information about teachers involved in control and experimental groups (teaching experience and whether they had a master or PhD degree) was asked to code. Besides these additions, some corrections were made in some of the items. For example, "reliability" was specified as "internal reliability" in Item 38, "average difficulty level of the instruments" was corrected as "average item difficulty for each instrument" in Item 39 on the final version, and "average distinctiveness of the instruments" was changed as "average item discrimination for each instrument" in Item 40 on the final version of coding sheet.

After these revisions, one more item was added, as Item 44, to the coding sheet to ask whether the assumptions of effect size estimation, which included normality, homogeneity of variances and independence of observations, were checked on the primary study to be coded. Finally, corresponding revision were made in the first draft of coding manual on the basis of final version of coding sheet.

As a result, the final version of the coding sheet (Appendix C), which included 49 items to be coded, and coding manual (Appendix D) covering corresponding explanations for each item were constructed.

3.5.2 Coding of the Primary Studies Included in the Meta-Analysis

After the final version of coding sheet and coding manual were developed, all of 95 studies selected to be included in the meta-analysis were coded by the researcher. During coding process, all primary studies were printed out and coding was performed on the printed form of coding sheets. All studies were read in detail and the related parts of the studies including necessary information to code the items in coding sheet were highlighted, taking small notes on these parts to make it easier to verify coding, when necessary. As explained before, seven studies were excluded from the analyses since they did not provide enough information to calculate the corresponding effect size, which yielded in 88 studies to be included in further analysis. Then, coding sheets for each of these 88 studies was checked by the researcher once more before sending some of them to be coded by other coders to calculate inter-rater reliability coefficient.

3.5.3 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).

To be able to establish high consistency between studies coded by the researcher; that is to establish high coder reliability, a coding manual, which included very detailed directions about how to code each item, was developed as explained in the previous section. In addition, the coding sheet was piloted to check whether there was any item that did not work as expected before the final version of coding sheet was developed based on the experience gained during pilot coding and the feedback provided by experts. Furthermore, each primary study was coded twice by the researcher, the second one of which was to check whether there was any missing point during the first one. The fact that the sections providing related information on the primary studies were highlighted by the researcher during first

coding made second-coding; i.e. checking procedure, easier. Finally, a subsample of the coded studies was coded by the researcher again four weeks after coding of all primary studies had been completed. To construct a subsample of ten studies, the primary studies were ordered according to their coding date and one of each ten studies was selected randomly; that is one study among the first ten studies and another one from the second ten studies etc… not to remember the original coding. Then, 'agreement rate' (AR) was calculated for ten pairs of studies to reach an average AR. The AR simply was calculated by the following formula (Orwin & Vevea, 2009):

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

An AR of .85 or greater is to be considered as sufficient (Bayraktar, 2002).

The researcher was the only coder in this meta-analysis. However, Lipsey and Wilson (2001) suggest that even for the meta-analysis including a single coder, both dimensions of reliability should be established to verify that another coder can easily reproduce the results. Thus, another subsample of 14 studies was selected to be coded by other researchers. Two studies were assigned for each of seven researchers, who accepted to be a parallel-coder for this meta-analysis. Five of seven researchers had a PhD degree, three of them in the field of physics education while the rest were in chemistry education. Two of the researchers, on the other hand, were PhD candidates, who had already completed their course load in the doctorate program including the ones related to educational research and statistics and had already passed PhD qualification exam. Coding sheet and coding manual are explained to the researchers at the beginning of the coding procedure but no external trainee program was held. In addition, they asked questions related to the difficulties they faced while coding primary studies. One journal article beside one thesis or dissertation were assigned to each of the researchers since journal article and thesis or dissertations were quite different publications in terms of the information provided to code items in the coding sheet. Then, the researchers were given a week to complete coding the studies and again AR was calculated for each pair of coding sheets, one of which had been coded by the researcher and the other

one had been completed by one of the seven researchers. An average AR was calculated, then, by averaging 14 ARs yielding from 14 pairs of coding sheets, which represented the coefficient of inter-coder reliability.

3.6 Further Statistical Issues in This Meta-Analysis

3.6.1 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.

3.6.1.1 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.

3.6.1.2 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 (τ), 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.

3.6.1.3 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) summarizes 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.

In this meta-analysis, not only Q statistic with corresponding statistical significance test but also the measures of T^2 with corresponding T value and I^2 are calculated and presented to check and quantify an important issue in meta-analysis, which is heterogeneity.

3.6.2 Moderator Analysis

One of the major aims of conducting a meta-analysis 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 on 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 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 this study (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.

In this study, either random-effects or mixed-effect model is used to conduct moderator analysis based on the properties of subgroups created in the scope of related moderator variable. For example, mixed-effect model is used for the variable of publication type while random-effects model is used to check whether effectiveness of PBL changes across different countries.

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 1 error rates.

3.6.2.1 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 is generally have low 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 provided by CMA. 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.

3.6.3 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. Figure 3.8 illustrates how power of a meta-analysis using fixed-effect model closes to 1.0 for number of studies larger than 25 even if the effect size as small as 0.20.

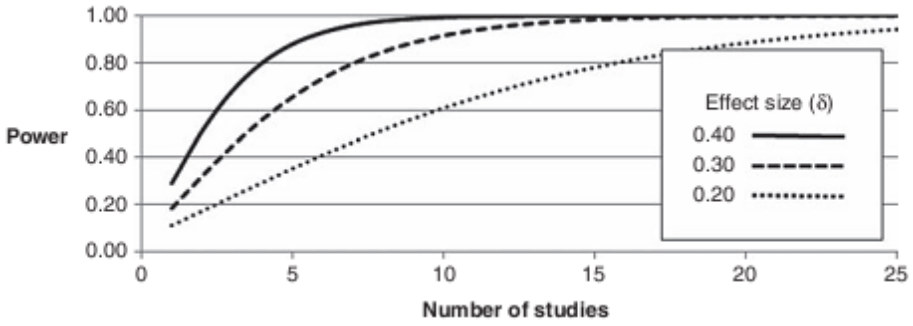

Figure 3.8 Power for a meta-analysis as a function of number of studies and effect size in a fixed-effect model (Borenstein et al., 2009)

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. Figure 3.9 shows how power of a meta-analysis using random-effects model decreases to undesired level with high heterogeneity and small number of studies.

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. Φ (x) can be calculated in EXCEL by using NORMSDIST function (Pigott, 2012).

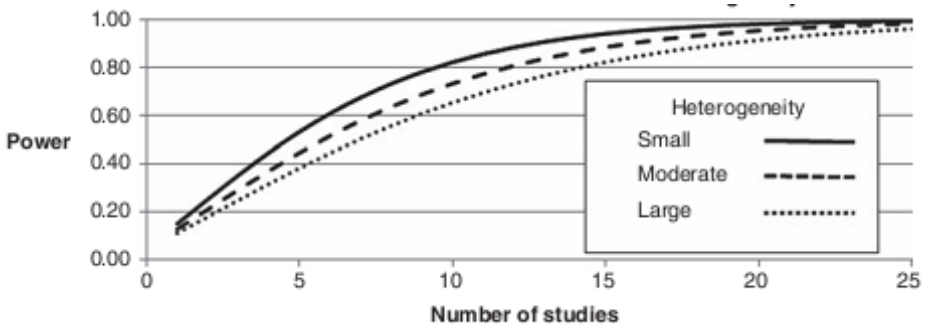

Figure 3.9 Power for a meta-analysis as a function of number of studies and heterogeneity in a random-effects model (Borenstein et al., 2009)

3.6.4 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 3.5 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.

Since all outcomes in the scope of this meta-analysis are measured as continuous variables, one of the effect size indices for groups compared on continuous outcomes would be appropriate to be selected as effect size index of the meta-analysis. As illustrated in Table 3.5, there are four options for this group of indices, one of which, response ratios, would not be appropriate since it is only suitable when the outcome is measured in ratio scale, which is not common in educational studies. Cohen's d is the most common one of the effect sizes representing groups compared on continuous outcomes, which is an uncorrected standardize 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;

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

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

Table 3.5 *Some of the common effect size indices*

d family	Groups compared on dichotomous outcomes	Risk difference (RD) Relative risk (RR) Odds ratio (OR)
	Groups compared on continuous outcomes	Cohen's d Glass delta (Δ) Hedge's g Response ratios I
r family	Correlation indices	Pearson correlation I Kendall's tau (τ) Phi coefficient (φ) Kruskal's lambda (λ)
	Proportion of variance indices	Coefficient of determination (r^2) R squared (R^2) Cohen's f Eta squared (η^2) Epsilon squared (ε^2) Omega squared (ω^2)

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

where Sc is the standard deviation of control group.

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} (Borenstein, 2009). 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).

Since it is an unbiased estimate of effect size and there are many primary studies in the meta-analysis with a wide range of sample size, Hedge's g is used as the effect size index in this meta-analysis, which is also compared to Cohen's d for each of analyses in this study to see how sensitive the results are in terms of the effect of sample size on the measure of effect size.

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 a 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).

3.6.5 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, no studies share 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.

In the scope of this meta-analysis, each primary study is used as unit of analysis across all analyses conducted for different research questions to prevent from lumpiness in the data by making all effect size values included in the same analysis independent from each other. If any of primary studies provides more than one effect size either they are averaged into one single effect size value or only one of them is accepted to be included in the analysis. In other words, no study is allowed to emerge more than one effect size for the analyses. For example, to

calculate an overall effect size for the effectiveness of PBL irrespective of types of outcome, each study is accepted as unit of analysis and one and only one effect size for each primary study is included in the analysis. In addition, since some studies provide two or three effect sizes for achievement outcome including retention assessments, only the effect size based on the measurement just after the treatment is included in the analysis to make them comparable to the effect sizes emerging from other studies.

How the studies with the same samples were excluded to be represented by only a single study in the meta-analysis was already explained previously in Section 2.4.3. Furthermore, the details of unit of analysis are explained at the beginning of each main effect analysis in Chapter 4.

3.6.6 Software for Statistical Analyses

Statistical packages designed for general purposes such as SPSS, SAS 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 (analysis of variance) and meta-regression, they produce incorrect p-values because of different rules for assigning degrees of freedom in meta-analysis (Borenstein et al., 2009). There are some statistical programs developed specifically for meta-analysis like Comprehensive Meta-Analysis (CMA), RevMan, Metawin and MIX. Bax, Yu, Ikeda, and Moons (2007) compares six statistical programs dedicated to meta-analysis and conclude that CMA is the most versatile software and one of the most usable programs designed for meta-analysis.

Thus, CMA version 2.2.064 was purchased to conduct the statistical analyses for this meta-analysis. It 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.

3.7 Summary of the Procedure Followed in This Meta-analysis

The aim of this final section of the chapter is to summarize the whole procedure followed during this meta-analysis study, main steps of which have already been explained in detail in different sections of this chapter. As illustrated in Figure 3.10, firstly, I have investigated how to conduct meta-analysis and how to benefit from different types of software programs, which has been an ongoing learning process throughout the whole study. For this purpose, I have reached many articles (Chan & Arvey, 2012; Glass, 1976, 1982, 2000; Hedges & Vevea, 1998; Rosenthal & DiMatteo, 2001; Shelby & Vaske, 2008) and books (Borenstein et al., 2009; Cooper, Hedges, & Valentine, 2009; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001; Littell et al., 2008; Pigott, 2012) written to explain meta-analysis in different levels and from different perspectives. Besides, I have examined different software packages for general purposes and specifically designed for meta-analysis as explained in the previous section. Then, I have decided to use CMA and learnt the properties of this software in detail.

In addition, I have conducted a comprehensive literature search, the details of which have been explained in Section 3.4, to reach all primary studies to be involved in this meta-analysis. Meanwhile, I have also developed the first version of coding sheet, about which further explanations are available in Section 3.5.1. After pilot coding, some revisions have been made on the earlier versions as a result of an iterative process to develop the final versions of coding sheets and coding manual. Then, all of the primary studies have been coded by the researcher while a sample of randomly selected primary studies has been coded by seven researchers as well to construct inter-coder reliability. Finally, I have conducted main effect and moderator analyses, as explained in Section 3.6, by using CMA before preparing final report to present and discuss the results of the study.

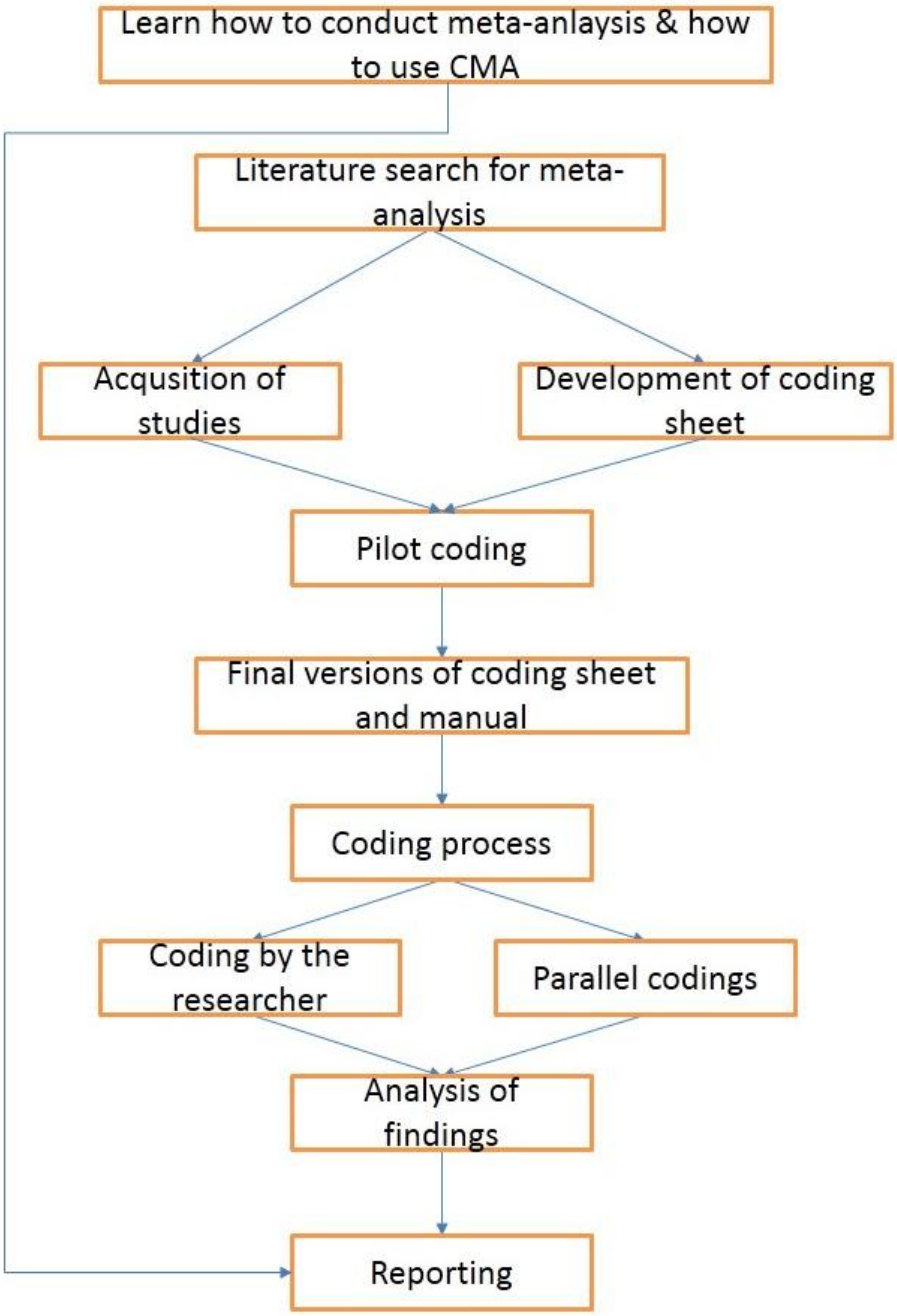

Figure 3.10 Main steps of the procedure followed in this meta-analysis study

CHAPTER IV

RESULTS

This chapter starts with the presentation of some descriptive information about the studies included in the meta-analysis. Then, findings related to each research question constitute other sections, which also cover publication bias analysis if necessary for the corresponding research question. The reason for conducting publication bias separately rather than making an overall evaluation at the very beginning of the chapter is that the sample of studies included for the specific research questions is not same for all cases. For example, the first research question envelops all of the studies coded in the meta-analysis since it seeks an overall generalization about the effectiveness of PBL while the second question only deals with the dependent variable of "achievement", which makes inevitably necessary to conduct publication bias analysis separately for these research questions. However, some research questions share a common sample of primary studies, for which publication bias is checked only for the first research question and not presented for the second one again.

As explained in the previous chapter, 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 et al., 2005, p. 1), mainly results from that fact that statistically significant studies are more likely to be published rather than non-significant ones. Constructing forest and funnel plots, conducting Egger's linear regression test and calculating Rosenthal's and Orwin's FSN are objective and functional ways of assessing publication bias. On the other hand, TFM provides us with not only estimation but also adjustment of the impact of publication bias, which is the most important

concern of meta-analysis in terms of validity. The results of all these methods are presented in the publication bias section for each of the related research questions to get an idea about the extent to which the results of the study are affected by the publication bias.

Although there are different methods to assess this important phenomenon, the most fundamental way of eliminating the effect of publication bias is to include studies from grey, i.e. unpublished, literature. Therefore, the proportions of different publication types including journal articles, doctoral dissertations and master theses are presented in the publication bias sections as well.

Then, the results of corresponding statistical tests are provided for each research question in two separate sections, which are main effect and moderator analyses.

4.1 Descriptive Statistics

147 effect size values revealed from 88 studies were included in this meta-analysis. Figure 4.1 shows histogram of all effect sizes with normal curve.

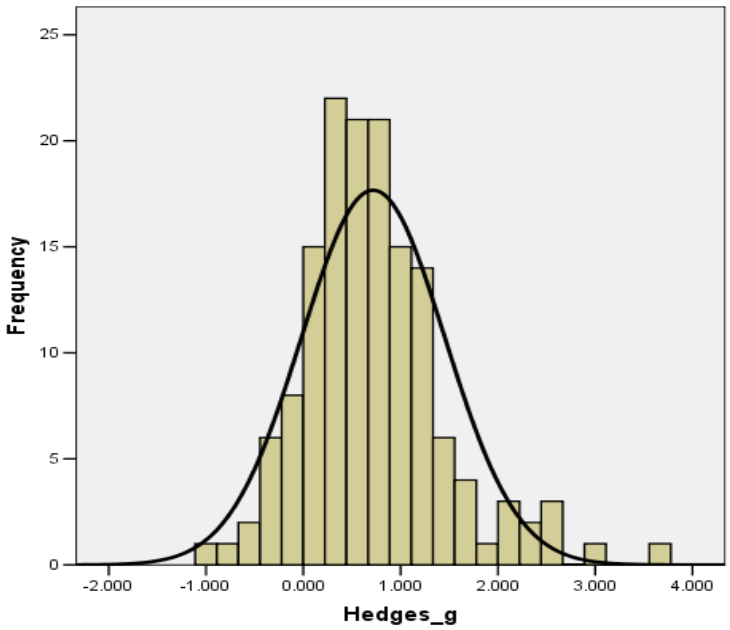

Figure 4.1 Histogram for 147 effect size values included in the meta-analysis

Arithmetic average of all effect sizes is 0.718 with a range from -1.084 to 3.672. Figure 4.2 illustrates a stem and leaf plot including 147 effect sizes. As illustrated on the stem and leaf plot, 18 of 147 effect sizes included in the meta-analysis are negative while 129 effect sizes yield a positive value, 55 of which are in the region of large effect size since they equal or larger than 0.8.

```
Frequency    Stem &  Leaf

    1.00 Extremes    (=<-1.1)
    1.00       -0 .  8
     .00       -0 .
    3.00       -0 .  444
    5.00       -0 .  22233
    8.00       -0 .  01111111
   14.00        0 .  00000001111111
   19.00        0 .  2222222333333333333
   16.00        0 .  4444444445555555
   25.00        0 .  6666666666667777777777777
   12.00        0 .  888888999999
   12.00        1 .  000000011111
   12.00        1 .  222222233333
    5.00        1 .  44455
    3.00        1 .  667
    1.00        1 .  8
    3.00        2 .  011
    7.00 Extremes    (>=2.3)

Stem width:     1.000
Each leaf:       1 case(s)
```

Figure 4.2 Stem and leaf plot for all effect sizes included in the meta-analysis

As explained in Section 3.5, final version of coding sheet includes 49 items to be coded in order to get as detailed information as possible about the primary studies. However, assigning all these items as moderator variables in this meta-analysis is neither feasible nor desirable. It is not feasible because as illustrated in Appendix E, which presents the distribution of primary studies for each item in the coding sheet, for some of the items, there are many primary studies coded as "unspecified" since they do not provide enough information to code corresponding items. For example, for Item 19, which is about "socio economic status" of participants involved in the primary studies, 89% of the studies coded in this meta-analysis do not provide related information, so they are labeled as "unspecified".

Furthermore, for some of the items, the primary studies are loaded in one of the choices, which makes impossible to compare them according to the variable represented by this item. For example, for Item 21, there is no study reporting that the length of treatments have not been same for control and experimental conditions, which means either length of treatment is kept same for both conditions in all studies included in this meta-analysis or there are some reporting deficiencies. In fact, assigning a vast number of variables in moderator analysis is not desirable either since conducting too many statistical tests for moderator analysis inflates Type 1 error rates (Lipsey & Wilson, 2001) as explained in Section 3.6.2.

Therefore, 12 variables represented by different items in the coding sheet, for which the distribution of primary studies was appropriate to compare, were selected to be involved in moderator analysis. Table 4.1 presents the descriptive summary of the primary studies for each subgroup under each independent variable analyzed as a moderator in this meta-analysis.

4.2 Main Effect Analysis

4.2.1 The Results for Research Question One

To what extent does PBL is effective on different outcomes when compared to traditional teaching methods?

4.2.1.1 Unit of Analysis

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the first research question.

Table 4.1 *Descriptive summary of the primary studies for subgroups under each independent variable in moderator analysis*

Variable		Number of Studies	% of Studies	Number of Effect Sizes	Hedge's g	95% Confidence Interval	
						Lower Limit	Upper Limit
Publication Type							
	Doctoral Dissertation	35	40	58	0.356	0.173	0.539
	Master Thesis	18	20	41	0.753	0.508	0.997
	Journal Article	35	40	48	0.830	0.671	0.988
Research Design							
	True Experimental	20	23	38	0.596	0.321	0.872
	QE with RAC	48	54	81	0.646	0.497	0.794
	QE without RAC	20	23	28	0.634	0.379	0.889
Teacher Effect							
	Different Teachers	24	27	35	0.588	0.357	0.820
	Same Teacher	41	47	73	0.594	0.431	0.756
	Unspecified	23	26	39	0.751	0.517	0.986
Researcher Effect							
	Not any of teachers	36	41	56	0.643	0.466	0.820
	One of teachers	10	11	16	0.587	0.376	0.799
	The only teacher	19	22	37	0.547	0.274	0.820
	Unspecified	23	26	38	0.699	0.433	0.966

Table 4.1 (continued)

Variable		Number of Studies	% of Studies	Number of Effect Sizes	Hedge's g	95% Confidence Interval	
						Lower Limit	Upper Limit
Country							
	Turkey	54	61	102	0.812	0.673	0.952
	USA	23	26	30	0.207	-0.014	0.428
	Others	11	13	15	0.632	0.331	0.932
Subject Area							
	Biology	14	16	27	0.457	0.166	0.748
	Chemistry	18	20	35	0.952	0.693	1.212
	Physics	26	30	43	0.618	0.404	0.832
	General Science	4	5	5	0.712	0.104	1.321
	Others	26	29	37	0.532	0.323	0.740
School level							
	Primary	26	30	48	0.834	0.569	1.100
	Secondary	17	19	29	0.606	0.289	0.924
	Higher	45	51	70	0.559	0.435	0.682
PBL Mode							
	Curriculum model	18	20	26	0.495	0.237	0.752
	Teaching method	70	80	121	0.678	0.545	0.811

Table 4.1 (continued)

Variable		Number of Studies	% of Studies	Number of Effect Sizes	Hedge's g	95% Confidence Interval	
						Lower Limit	Upper Limit
Length of treatment							
	0-5 weeks	32	36	49	0.613	0.419	0.807
	6-10 weeks	26	29	54	0.682	0.467	0.897
	Over 10 weeks	19	22	29	0.480	0.232	0.729
	Unspecified	11	13	15	0.898	0.572	1.224
Group size							
	0-6	44	50	76	0.716	0.550	0.882
	Over 6	20	23	31	0.525	0.285	0.765
	Unspecified	24	27	40	0.603	0.376	0.830
Type of questions							
	Only objective	63	72	106	0.583	0.441	0.724
	Only open-ended	3	3	3	0.826	0.277	1.375
	Both	21	24	36	0.734	0.528	0.940
	Unspecified	1	1	2	1.065	0.793	1.337
Type of ass. Instrument							
	Adapted	5	6	6	1.037	0.497	1.578
	Pre-existing	42	48	89	0.488	0.300	0.676
	Researcher-developed	41	46	52	0.856	0.668	1.044

4.2.1.2 Publication Bias

Table 4.2 shows the number of coded studies in different publication types including journal articles, doctoral dissertations and master theses.

Table 4.2 *The number of studies and effect sizes in different publication types and corresponding point estimate for research question one.*

Publication Type	Number of Study (Percentage)	Number of Effect Size	Point Estimate (Hedge's g)
Doctoral Dissertation	35 (40%)	58 (40%)	0.356
Master Thesis	18 (20%)	41 (28%)	0.753
Journal Article	35 (40%)	48 (32%)	0.830
Total	88 (100%)	147 (100%)	0.651

As Table 4.2 illustrates, 53 of 88; i.e. 60% of all studies coded in the meta-analysis are either doctoral dissertations or master theses, from which 99 effect size values emerge yielding smaller mean effect sizes in terms of Hedge's g than journal articles do as predicted. However, covering dissertation and theses rather than including only journal articles increases representativeness of studies coded in meta-analysis while decreasing the possibility of publication bias. It is important to note that the calculation of Hedge's g values is based on random-effects model and it is the calculated value when each study is accepted as unit of analysis.

Figure 4.3, 4.4 and 4.5 show the forest plots of the studies with high, moderate and low precision, respectively. The groups are based on the categorization of primary studies according to their precision, which is inversely related to standard error of the studies. The first 30 studies with highest precision are assigned as the ones with high precision while the next 30 and the last 28 studies are grouped as moderate precision and low precision studies, respectively. As illustrated on the figures, the observed effect sizes revealed from the studies with high precision spread across a narrow range of values while especially the

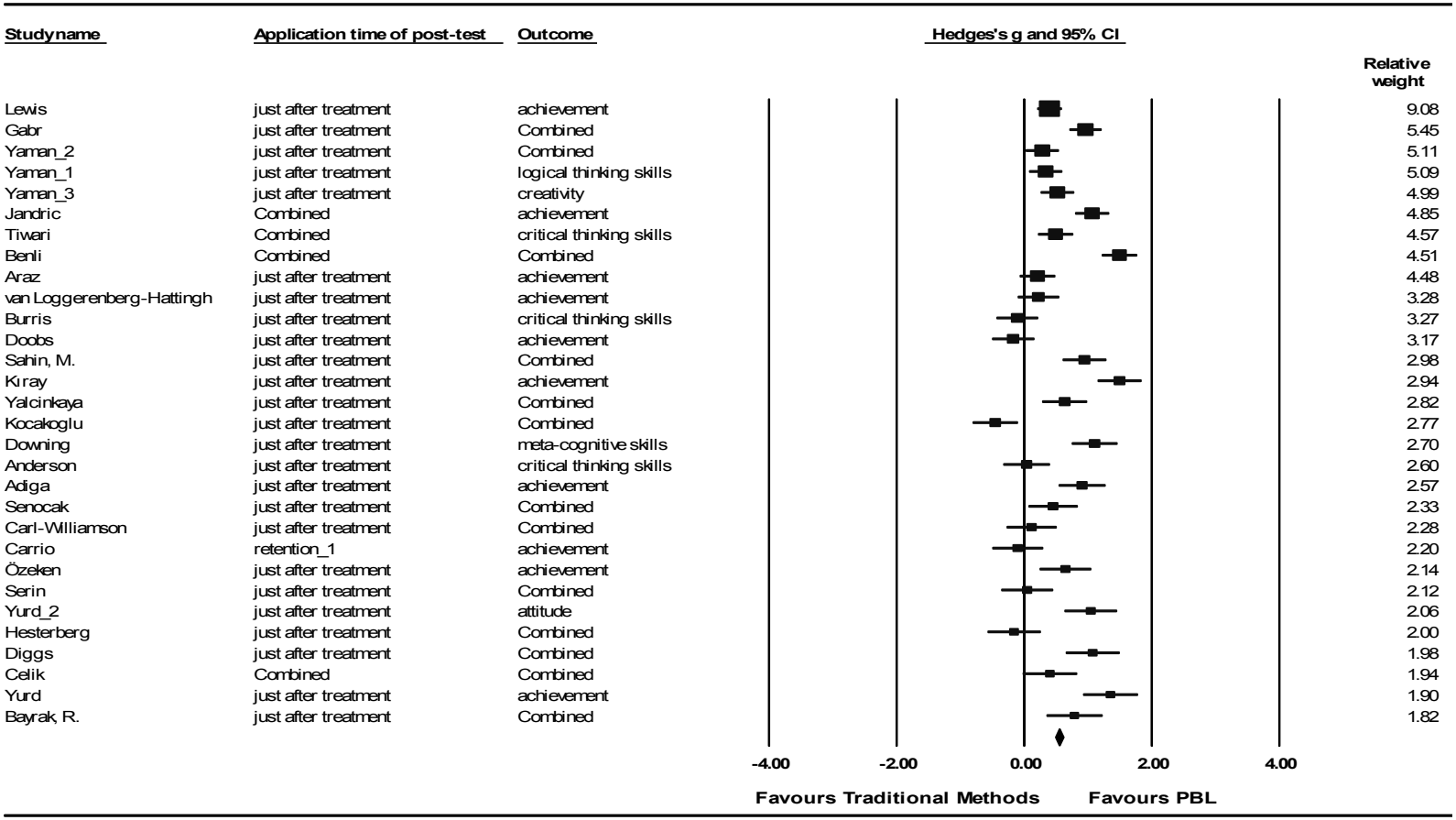

Figure 4.3 Forest plot for the first 30 studies when all studies included in the sample of the first research question are ranked based on their precisions

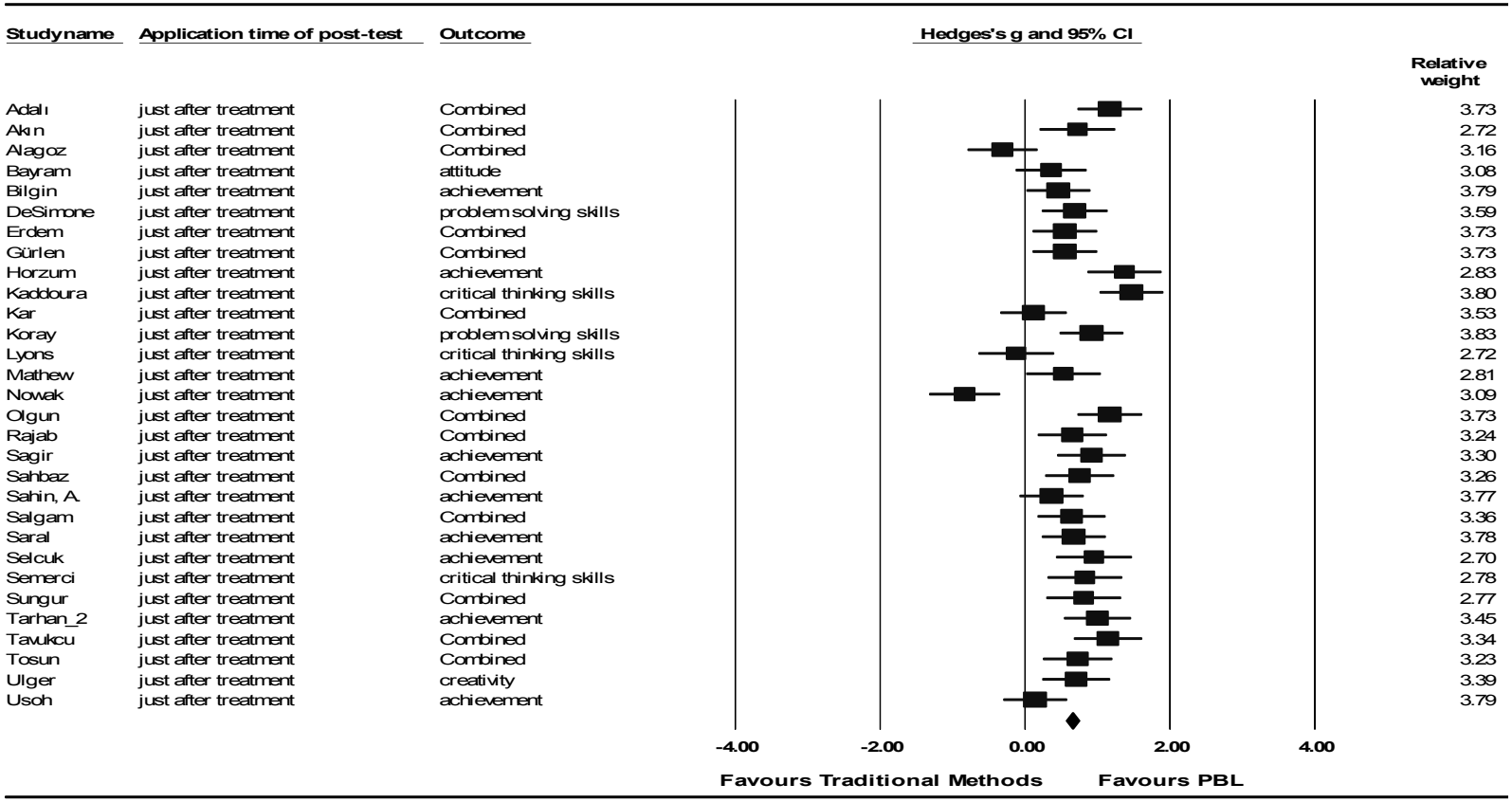

Figure 4.4 Forest plot for the second 30 studies when all studies in the sample of the first research question are ranked based on their precisions

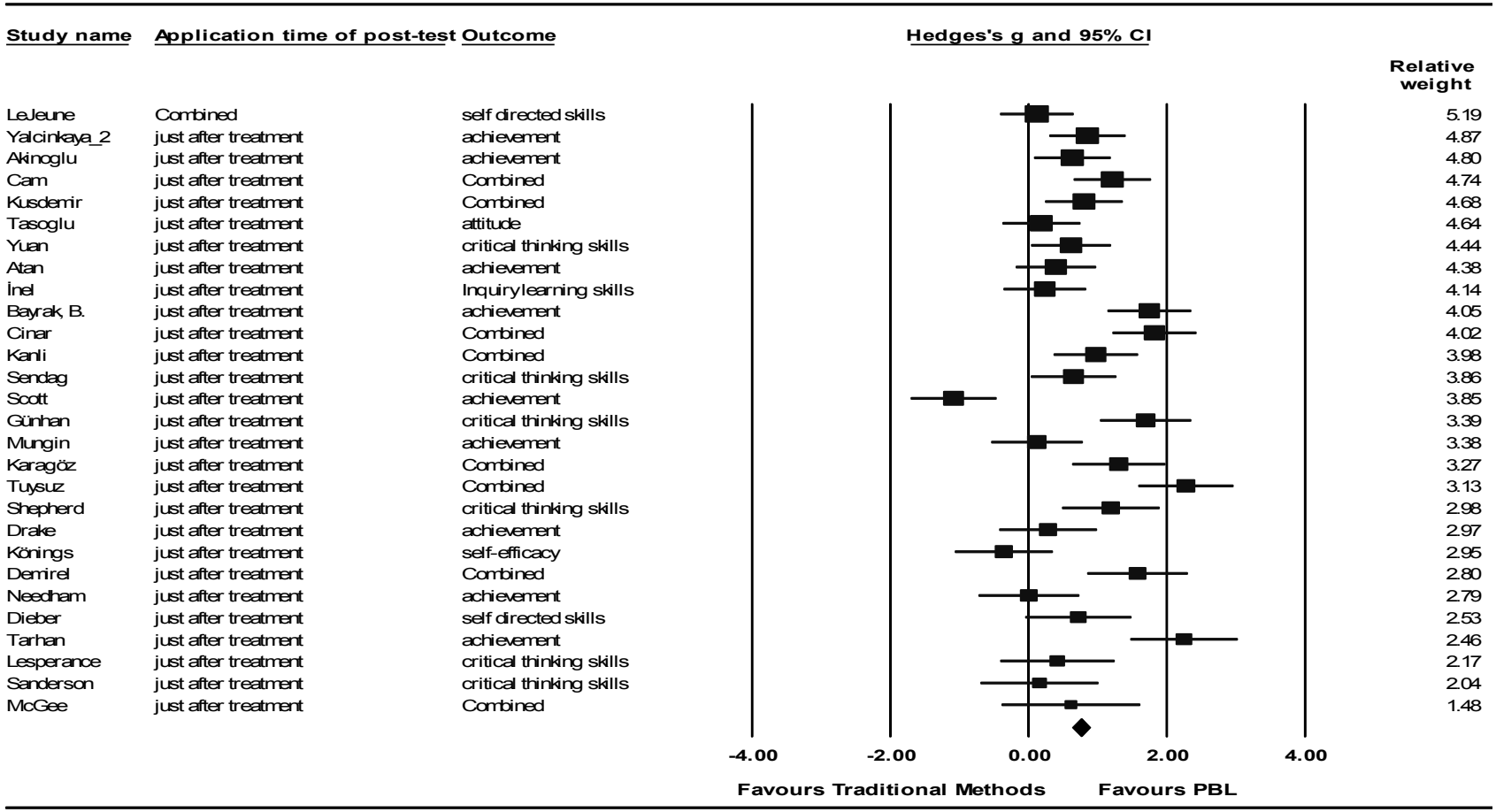

Figure 4.5 Forest plot for the last 28 studies when all studies in the sample of the first research question are ranked based on their precisions

ones from least precise studies show more variation spreading across a wider range of values, which is what is expected when there is no significant bias. However, as summarized in Table 4.3, the average effect size values calculated for the primary studies increase with decreasing precision, which may result from existence of missing studies.

Thus, the results of other methods to detect publication bias should be interpreted cautiously although the effect sizes may be really larger in small studies; i.e. less precise studies.

Table 4.3 *Mean effect size values for the studies with high, moderate and low precision studies in the sample of the first research question*

Precision	Number of Study	Point Estimate (Hedge's g)
High	30	0.535
Moderate	30	0.639
Low	28	0.764
Overall	88	0.633

Figure 4.6 shows the funnel plot constructed upon random effect model by considering each study as unit of analysis. The "funnel" shape of the plot seems to fit the one resulting from unbiased data since it gets wider as the standard error increases at the bottom of the shape. The empty diamond-shaped indicator on the horizontal axis shows the mean effect size computed by random effect model while filled one indicates the adjusted mean effect size, which is the unbiased value computed by TFM by adding some effect sizes emerging from fictitious studies, if necessary. As it can be seen in the figure, the adjusted value is same with the computed mean effect size; therefore there is no imputed study on the plot when we are looking for missing studies either to the left or to the right of the mean effect size based on random-effects model. This result is also supported by the results of Egger's Regression Test as summarized in Table 4.4 since the null hypothesis that *"there is no funnel plot asymmetry ($\beta_0 = 0$)"* cannot be rejected ($p > 0.05$).

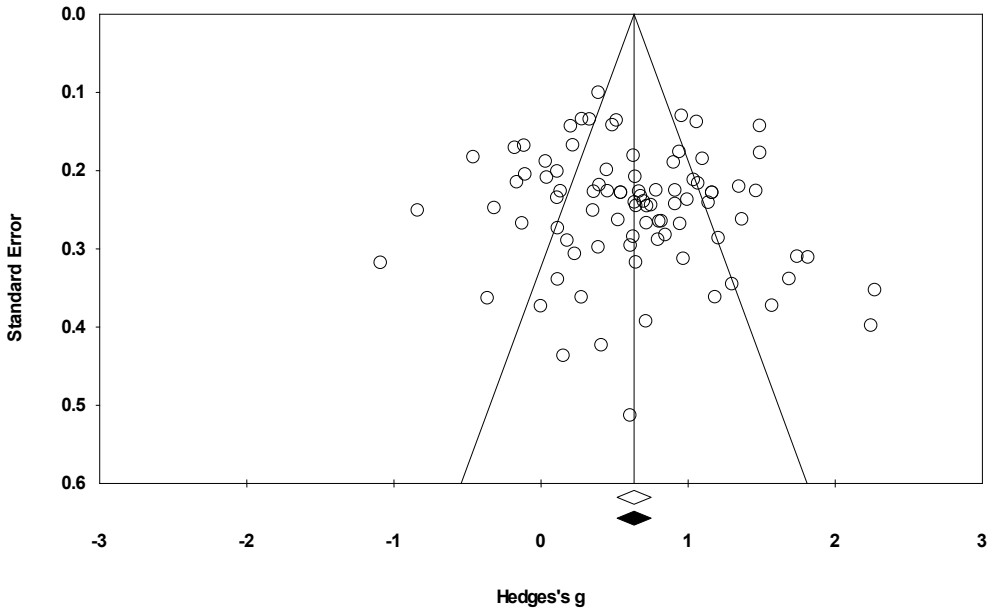

Figure 4.6 Funnel plot of all studies included in the meta-analysis based on random effect model

Table 4.4 *The results of Egger's Regression Test for all studies included in the meta-analysis*

Intercept	0.95298
Standard error	0.84842
95% lower limit (2-tailed)	-0.73363
95% lower limit (2-tailed)	2.63959
t value	1.12323
df	86
p value (2-tailed)	0.26446

In addition, Table 4.5 presents the results of Rosenthal's FSN calculations, which reveal a FSN of 4509 meaning that 4509 additional studies with non-significant results are necessary to nullify the effect. Consequently, the results of the meta-analysis can be accepted as robust to publication bias since computed FSN is much larger than cut-off point stated by Mullen et al. (2001). They propose that N/(5k+10) should exceed 1, meaning that N is expected to be larger than 5k+10, where k represents the total number of studies included in the study. The ratio for

this meta-analysis is 10.02, so the results of this meta-analysis seem to be highly robust the publication bias.

Table 4.5 *Rosenthal's FSN for all studies included in meta-analysis*

Z-value for observed studies	25.24229
p-value for observed studies	0.00000
Alpha	0.05
Tails	2
Z for alpha	1.95996
Number of observed studies	88
Fail safe N	4509

Another indicator to check the impact of publication bias is Orwin's FSN, which is based on practical significance rather than p value. Table 4.6 illustrates the results of the calculations for this meta-analysis, which means that additional 439 studies with effect sizes of 0.00000 are necessary to bring the mean effect of this meta-analysis under 0.1, which can be called as trivial in many contexts.

Table 4.6 *Orwin's FSN for all studies included in meta-analysis*

Hedge's g in observed studies	0.598
Criterion for a 'trivial' Hedge's g	0.100
Mean Hedge's g in missing studies	0.000
Fail safe N	439

Thus, we can clearly conclude that the impact of bias on the results related to first research question is trivial although there are some differences between the mean effect sizes of high, moderate and low precise studies because all other indicators and tests show that the results are very robust to publication bias with no imputed studies as a result of TFM and unrealistic FSN numbers produced by Rosenthal's and Orwin's procedures.

4.2.1.3 Overall Mean Effect Size and Corresponding Statistical Test

Null Hypothesis: H_o: $\delta_1 = 0$

Mean of all true effect sizes for populations represented by the studies investigating the effect of PBL on all types of outcomes chosen in the study equals to zero.

The random effect model conducted to estimate the effectiveness of PBL on the whole when compared to traditional teaching methods reveals an overall effect size of 0.633 with the 95% confidence interval of 0.517 and 0.749, which is a medium to large effect. Furthermore, the null hypothesis related to research question one is rejected at the alpha level of 0.05 (p= 0.000), indicating that the mean of all true effect sizes is significantly different from zero. The details of the results for research question one is presented in Table 4.7.

Table 4.7 *Overall effect size details and corresponding statistical test for research question one.*

Model	Effect Size and 95% confidence interval						Statistical test	
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value
Fixed	88	0.598	0.023	0.001	0.553	0.644	25.641	0.000
Random	88	0.633	0.059	0.003	0.517	0.749	10.709	0.000

4.2.1.4 Power Analysis

The variance for the point estimate of 0.633 is 0.003 based on random-effects model, so the parameter λ is:

$$\lambda = \frac{\delta}{\sqrt{V_\delta}} = \frac{0.633}{\sqrt{0.003}} = 10.729$$

Then, power is calculated with an alpha level of 0.05 as:

$$\text{Power} = 1 - \Phi (c_\alpha - \lambda) + \Phi (-c_\alpha - \lambda) = 1.000$$

As expected, the power of the statistical test is very high in consequence of a large number of primary studies included in the study and medium mean effect size revealed from these studies. If the standard error of the statistical test was 0.2

instead of 0.059, the power analysis would yield still appropriate but smaller value, 0.89. If the mean effect size was small, 0.1 with the same standard error of 0.059, then the power of statistical test would decrease an unacceptable value of 0.40. The power of 1.000 means that the probability of Type II error; i.e. fail to detect a real treatment effect, is almost zero.

$$B = 1 - \text{Power} = 0$$

4.2.1.5 Heterogeneity Analysis

Table 4.8 summarizes different values for heterogeneity analysis. Firstly, according to chi-squared significance test, the null hypothesis of "all studies share a common effect size" is rejected ($p < 0.05$), which means that the distribution of effect sizes shows heterogeneity indicating the possibility of moderator variables. In addition, I^2 statistic quantifies the heterogeneity on the data as indicating that 83% of total variance results from between study variance, which is labeled as high heterogeneity. High values related to tau-squared also supports this conclusion. T^2 quantifies the between study variance in the same metric with Hedge's g, which results in a high value of true variance as well. Corresponding tau value causes a 95% prediction interval of -0.353 to 1.619, meaning that 95% of cases the true effect size in a new study would fall inside in this prediction interval (Borenstein et al., 2009).

Table 4.8 *Heterogeneity test for research question one.*

Heterogeneity				Tau-Squared			
Q-value	df(Q)	p-value	I^2	Tau-squared	Standard Error	Variance	Tau
525.156	87	0.000	83.434	0.243	0.051	0.003	0.493

4.2.2 The Results for Research Question Two

What is the effectiveness of PBL on science achievement when compared to traditional teaching methods?

4.2.2.1 Unit of Analysis

Each study investigating the effect of PBL on achievement in this meta-analysis is accepted as the unit of analysis for the second research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. 52 primary studies are included to examine the second research question.

4.2.2.2 Publication Bias

As illustrated on Table 4.9, 29 of 52 studies covered by the sample of second research question are either doctoral dissertations or master theses. Since there is only one outcome, which is achievement, for research question two, the number of total effect sizes (57) is not very different from the total number of studies (52). An unexpected finding is that the mean effect size of master theses is larger than the one for journal articles, which may result from an outlier in master theses having very large effect size (3.057). Similar to research question one, covering dissertation and theses rather than including only journal articles for research question two increases representativeness of studies coded in meta-analysis while decreasing the possibility of publication bias.

Table 4.9 *The number of studies and effect sizes in different publication types and corresponding point estimate for research question two.*

Publication Type	Number of Study (Percentage)	Number of Effect Size	Point Estimate (Hedge's g)
Doctoral Dissertation	16 (31%)	16 (28%)	0.362
Master Thesis	13 (25%)	17 (30%)	1.168
Journal Article	23 (44%)	24 (42%)	0.947
Total	52 (100%)	57 (100%)	0.820

Figure 4.7, 4.8 and 4.9 show the forest plots of the studies with high, moderate and low precision, respectively. The groups are based on the categorization of primary studies according to their precision, which is inversely

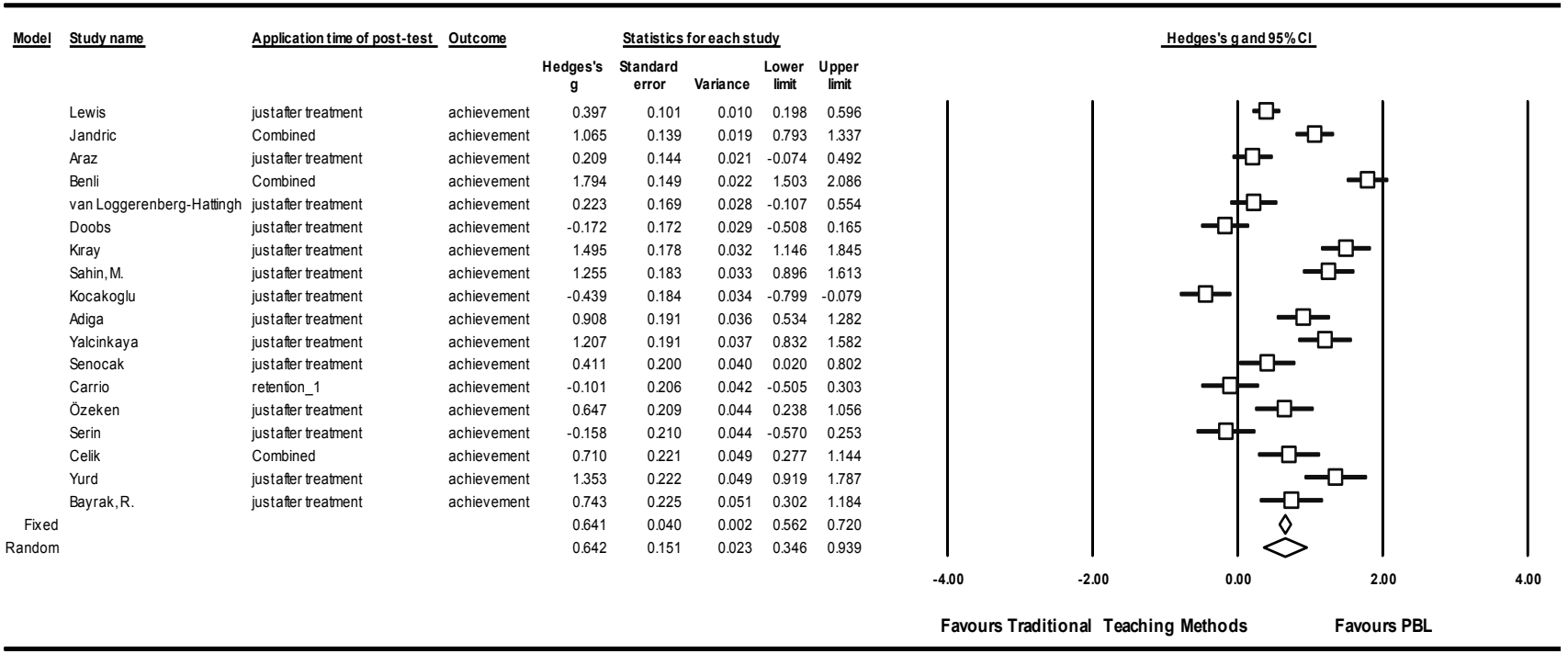

Figure 4.7 Forest plot for the first 18 studies when all studies included in the sample of second research question are ranked based on their precisions

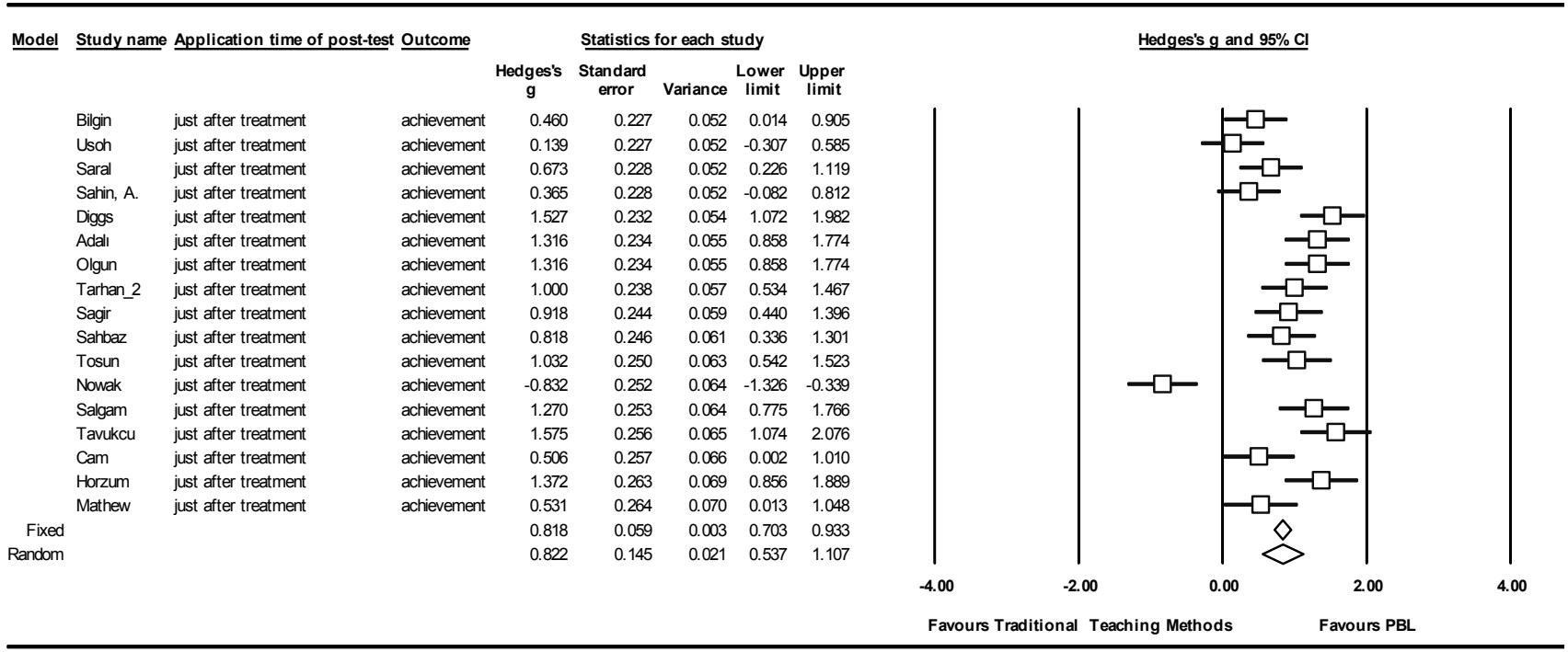

Figure 4.8 Forest plot for the second 17 studies when all studies included in the sample of second research question are ranked based on their precisions

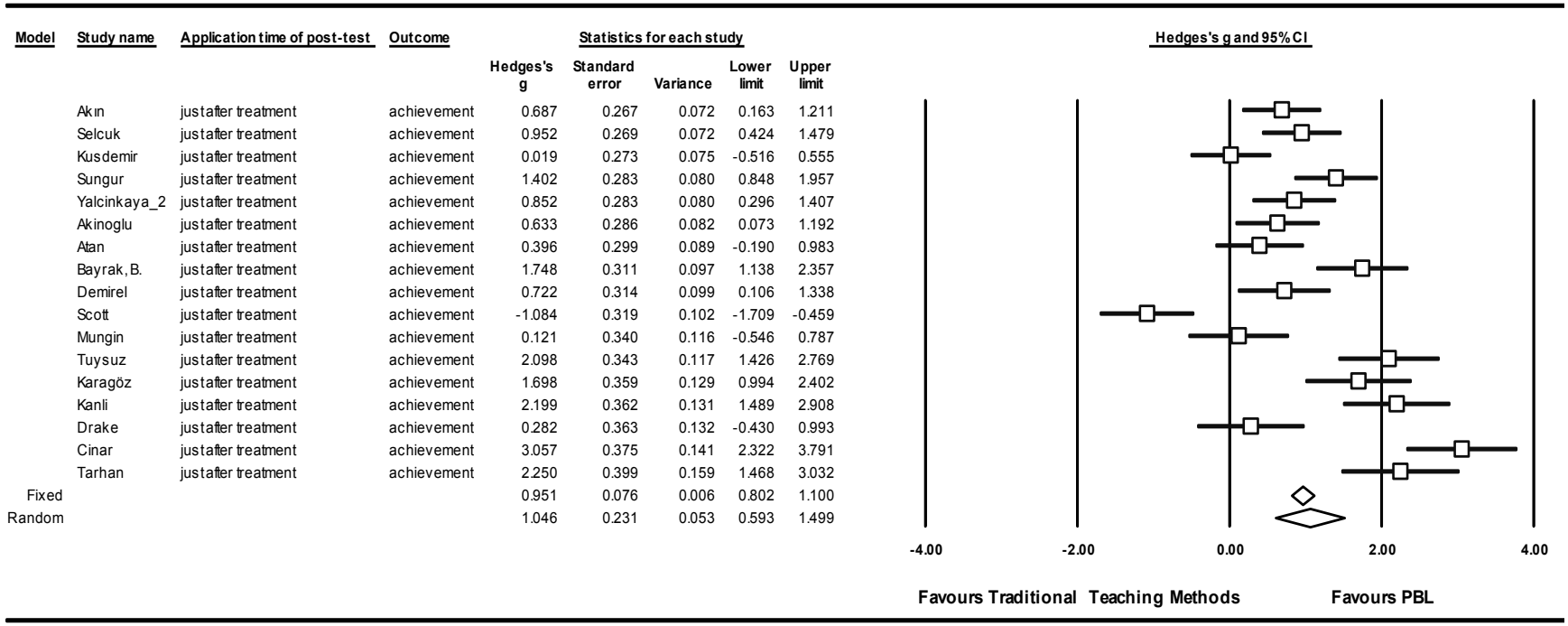

Figure 4.9 Forest plot for the last 17 studies when all studies included in the sample of second research question are ranked based on their precisions

related to standard error of the studies. The first 18 studies with highest precision are assigned as the ones with high precision while the next 17 and the last 17 studies are grouped as moderate precision and low precision studies, respectively. As illustrated in the figures, the observed effect sizes revealed from least precise studies show more variation spreading across a wider range of values, which is what is expected when there is no significant bias. However, as summarized in Table 4.10, the mean effect size values calculated for the primary studies increase with decreasing precision, which may result from existence of missing studies.

Table 4.10 *Mean effect size values for the studies with high, moderate and low precision studies in the sample of the second research question*

Precision	Number of Study	Point Estimate (Hedge's g)
High	18	0.642
Moderate	17	0.822
Low	17	1.046
Overall	88	0.820

The effect sizes may be really larger in less precise studies or there may be some other moderator variables affecting the effect of PBL differently in low and high precise studies. Nevertheless, the results of other methods to detect publication bias should be interpreted carefully.

Figure 4.10 shows the funnel plot constructed upon random effect model by considering each study in the sample of studies as unit of analysis. The "funnel" shape of the plot seems to fit the one resulting from unbiased data since it gets wider as the standard error increases at the bottom of the shape. Furthermore, TFM imputes no additional study for either right or left of the mean resulting in an adjusted effect size equal to the computed one.

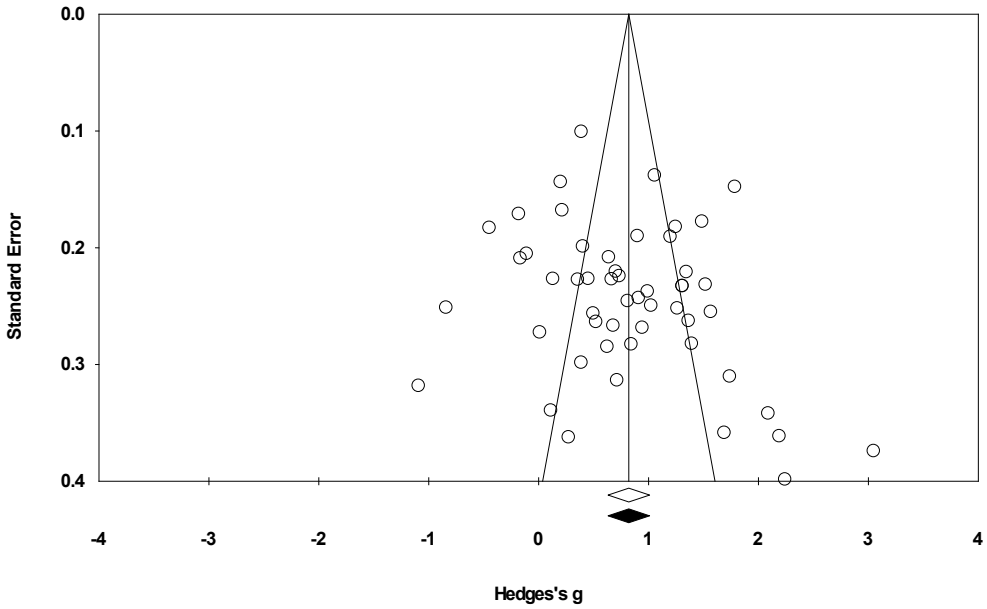

Figure 4.10 Funnel plot of the studies included in the sample of second research question based on random effect model

This result is also supported by the results of Egger's Regression Test as summarized in Table 4.11 since the null hypothesis that *"there is no funnel plot asymmetry ($\beta_0 = 0$)"* cannot be rejected ($p > 0.05$).

Table 4.11 *The results of Egger's Regression Test for all studies included in the sample of the second research question*

Intercept	2.34213
Standard error	1.46504
95% lower limit (2-tailed)	-0.60049
95% lower limit (2-tailed)	5.28475
t value	1.59868
df	50
p value (2-tailed)	0.11619

Rosenthal's FSN computed for the studies included in the meta-analysis for research question two, the details of which are presented in Table 4.12, reveals a ratio of 30.174, which confirms the robustness of the data to publication bias since it is far away from the cutoff point of one.

Table 4.12 *Rosenthal's FSN for all studies included in the sample of the second research question*

Z-value for observed studies	24.60939
p-value for observed studies	0.00000
Alpha	0.05
Tails	2
Z for alpha	1.95996
Number of observed studies	52
Fail safe N	8147

In addition, Orwin's FSN, the details of which are illustrated in Table 4.13, indicates that 296 additional studies with effect sizes of 0.00000 are necessary to bring the mean effect of this meta-analysis under 0.1. Again, this number is much larger than the number of existing studies in the meta-analysis (52).

Table 4.13 *Orwin's FSN for all studies included in the sample of the second research question*

Hedge's g in observed studies	0.739
Criterion for a 'trivial' Hedge's g	0.100
Mean Hedge's g in missing studies	0.000
Fail safe N	333

Therefore, although there are some differences between the mean effect sizes of high, moderate and low precise studies, we can still conclude that the impact of publication bias on the results related to second research question is trivial because all other indicators and tests show that the results are very robust to publication bias with no imputed studies as a result of TFM and unrealistic FSN numbers produced by Rosenthal's and Orwin's procedures.

4.2.2.3 Overall Mean Effect Size and Corresponding Statistical Test

Null Hypothesis: H_o: $\delta_2 = 0$

Mean of all true effect sizes for populations represented by the studies investigating the effect of PBL on science achievement equals to zero.

The random effect model conducted to estimate the effectiveness of PBL on science achievement when compared to traditional teaching methods results in an overall effect size of 0.820 with the 95% confidence interval of 0.631 and 1.010, which is a large effect. Furthermore, the null hypothesis related to research question two is rejected at the alpha level of 0.05 (p= 0.000), indicating that the mean of all true effect sizes is significantly different from zero. The details of the results for research question two is presented in Table 4.14.

Table 4.14 *Overall effect size details and corresponding statistical test for research question two.*

Model	Effect Size and 95% confidence interval						Statistical test	
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value
Fixed	52	0.739	0.030	0.001	0.679	0.799	24.232	0.000
Random	52	0.820	0.097	0.009	0.631	1.010	8.496	0.000

4.2.2.4 Power Analysis

The variance for the point estimate of 0.820 is 0.009 based on random-effects model, so the parameter λ is:

$$\lambda = \frac{\delta}{\sqrt{V_\delta}} = \frac{0.820}{\sqrt{0.009}} = 8.454$$

Then, power is calculated with an alpha level of 0.05 as:

$$\text{Power} = 1 - \Phi(c_\alpha - \lambda) + \Phi(-c_\alpha - \lambda) = 1.000$$

Although the number of studies is smaller than the one for the previous research question, the power of the statistical test is still very high in consequence of large mean effect size revealed from these studies, the number of which is still quite high. If the standard error of the statistical test was 0.2 instead of 0.097, the power analysis would yield still appropriate but smaller value, 0.984. If the mean effect size was small, 0.1 with the same standard error of 0.097, then the power of statistical test would decrease an unacceptable value of 0.178.

The power of 1.000 means that the probability of Type II error; i.e. fail to detect a real treatment effect, is almost zero.

$$B = 1- Power = 0$$

4.2.2.5 Heterogeneity Analysis

Different values for heterogeneity analysis are illustrated in Table 4.15. Firstly, according to chi-squared significance test, the null hypothesis of "all studies share a common effect size" is rejected ($p < 0.05$), which means that the distribution of effect sizes shows heterogeneity indicating the possibility of moderator variables. In addition, I^2 statistic quantifies the heterogeneity on the data as indicating that 89.6% of total variance results from between study variance, which is labeled as high heterogeneity. High values related to tau-squared also supports this conclusion. T^2 quantifies the between study variance in the same metric with Hedge's g, which results in a high value of true variance as well. Corresponding tau value causes a 95% prediction interval of -0.498 to 2.138, meaning that 95% of cases the true effect size in a new study would fall inside in this prediction interval.

Table 4.15 *Heterogeneity test for research question two*

Heterogeneity				Tau-Squared			
Q-value	df(Q)	p-value	I^2	Tau-squared	Standard Error	Variance	Tau
491.528	51	0.000	89.624	0.422	0.109	0.012	0.649

4.2.3 The Results for Research Question Three

What is the effectiveness of PBL on students' attitudes toward science when compared to traditional teaching methods?

4.2.3.1 Unit of Analysis

Each study investigating the effect of PBL on students' attitudes toward science in this meta-analysis is accepted as the unit of analysis for the third research question, meaning that each primary study provides one and only one effect size for

the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. 23 primary studies are included to examine the third research question.

4.2.3.2 Publication Bias

As illustrated on Table 4.16, 18 of 23 studies, i.e. 78% of the studies covered by the sample of third research question are either doctoral dissertations or master theses. Since there is only one outcome, which is attitude toward science, for research question three, and there is no study applying retention assessment for this construct, the number of total effect sizes is equal to the total number of studies (23). As predicted, the mean effect size of journal articles is larger than the ones for master theses and doctoral dissertations. However, covering dissertation and theses rather than including only journal articles for research question three increases representativeness of studies coded in meta-analysis while decreasing the possibility of publication bias.

Table 4.16 *The number of studies and effect sizes in different publication types and corresponding point estimate for research question three.*

Publication Type	Number of Study (Percentage)	Number of Effect Size	Point Estimate (Hedge's g)
Doctoral Dissertation	8 (35%)	8 (35%)	0.399
Master Thesis	10 (43%)	10 (43%)	1.033
Journal Article	5 (22%)	5 (22%)	0.482
Total	23 (100%)	23 (100%)	0.566

Figure 4.11 and 4.12 show the forest plots of the studies with high and low precision, respectively. The groups are based on the categorization of primary studies according to their precision, which is inversely related to standard error of the studies. The first 12 studies with highest precision are assigned as the ones with high precision while the next 11 studies are grouped as low precision studies, respectively. As illustrated on the figures, the observed effect sizes revealed from less precise studies show a bit more variation spreading across a wider range of

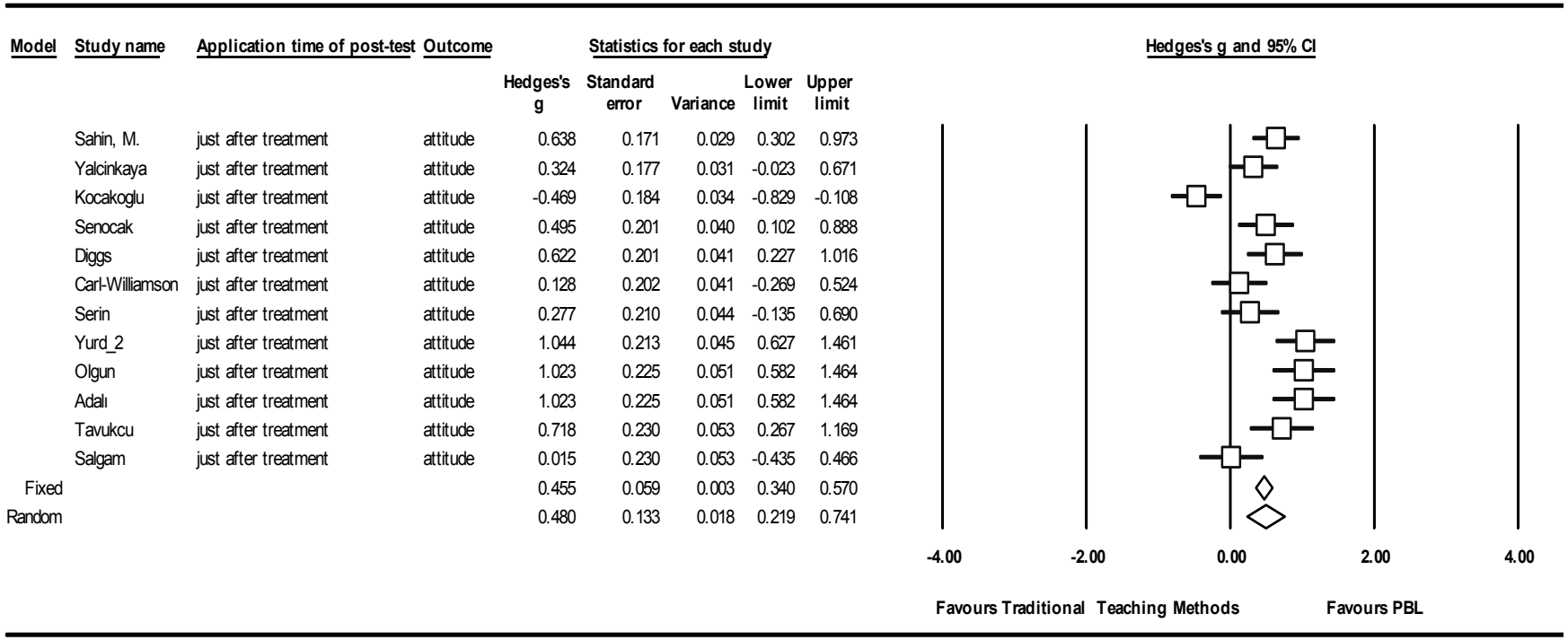

Figure 4.11 Forest plot for the first 12 studies when all studies included in the sample of third research question are ranked based on their precisions

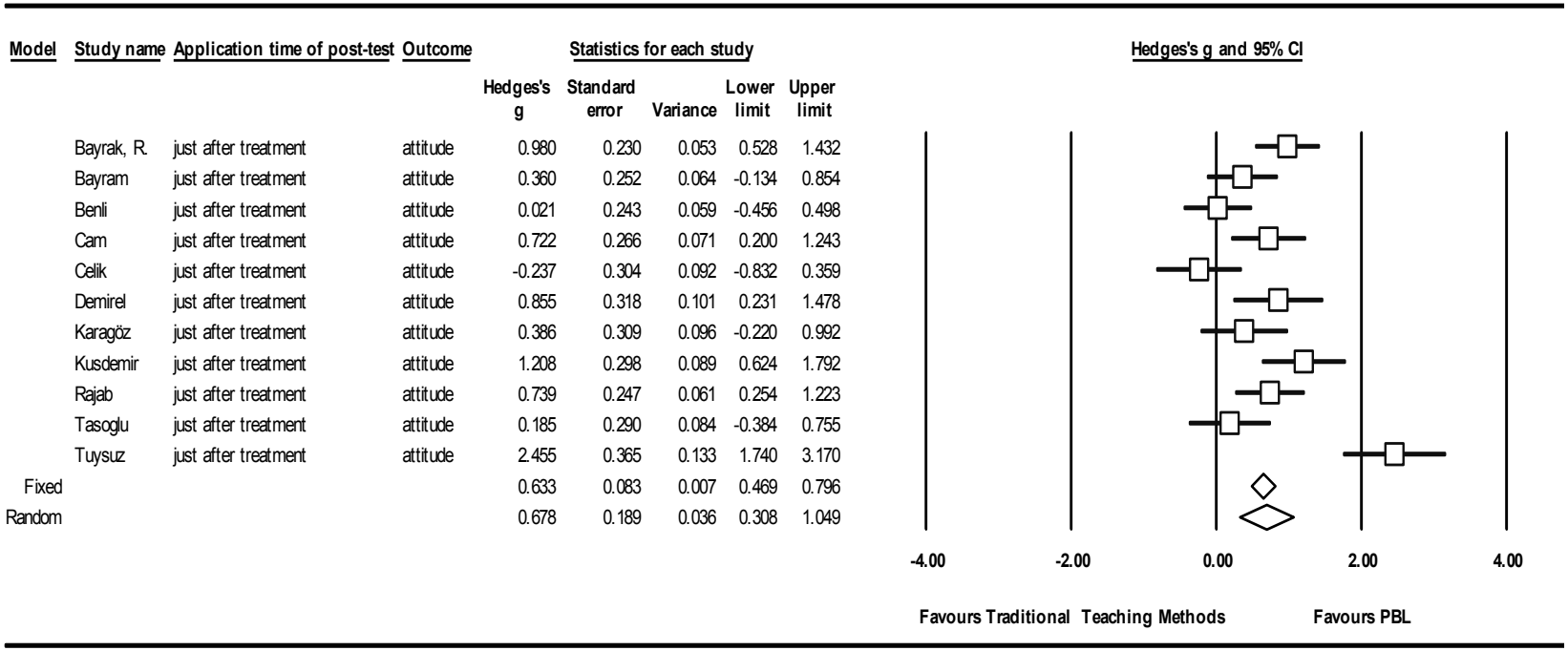

Figure 4.12 Forest plot for the last 11 studies when all studies included in the sample of third research question are ranked based on their precisions

values, which is what is expected when there is no significant bias. However, as summarized in Table 4.17, the mean effect size values calculated for the primary studies increase with decreasing precision, which may result from existence of missing studies.

Table 4.17 *Mean effect size values for the studies with high and low precision studies in the sample of the third research question*

Precision	Number of Study	Point Estimate (Hedge's g)
High	12	0.480
Low	11	0.678
Overall	88	0.566

Figure 4.13 shows the funnel plot constructed upon random effect model by considering each study in the sample of studies as unit of analysis. The "funnel" shape of the plot seems to fit the one resulting from unbiased data since it gets wider as the standard error increases at the bottom of the shape. Furthermore, TFM imputes no additional study for either right or left of the mean resulting in an adjusted effect size equal to the computed one.

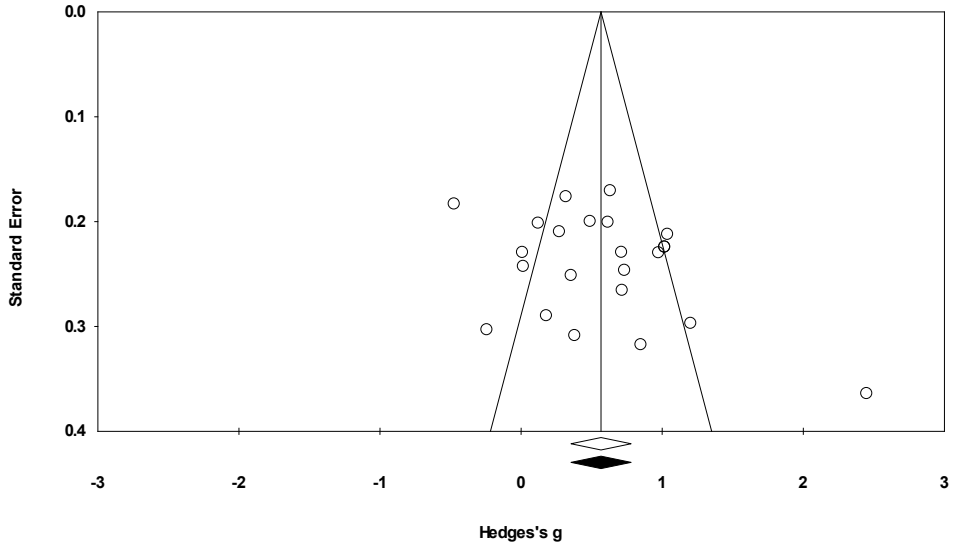

Figure 4.13 Funnel plot of the studies included in the sample of third research question based on random effect model

This result is also supported by the results of Egger's Regression Test as summarized by Table 4.18 since the null hypothesis that *"there is no funnel plot asymmetry ($\beta_0 = 0$)"* cannot be rejected ($p > 0.05$).

Table 4.18 *The results of Egger's Regression Test for all studies included in the sample of the third research question*

Intercept	4.05856
Standard error	2.38642
95% lower limit (2-tailed)	-0.90426
95% lower limit (2-tailed)	9.02139
t value	1.70069
Df	21
p value (2-tailed)	0.10376

Rosenthal's FSN computed for the studies included in the meta-analysis for research question three, the details of which are presented in Table 4.19, reveals a ratio of 5.84, which confirms the robustness of the data to publication bias since it is much larger than the cutoff point of one.

Table 4.19 *Rosenthal's FSN for all studies included in the sample of the third research question*

Z-value for observed studies	11.20777
p-value for observed studies	0.00000
Alpha	0.05
Tails	2
Z for alpha	1.95996
Number of observed studies	23
Fail safe N	730

In addition, Orwin's FSN, the details of which are illustrated in Table 4.20, indicates that 96 additional studies with effect sizes of 0.00000 are necessary to bring the mean effect of this meta-analysis under 0.1. Again, this number is much larger than the number of existing studies in the meta-analysis (23).

Table 4.20 *Orwin's FSN for all studies included in the sample of the third research question*

Hedge's g in observed studies	0.51355
Criterion for a 'trivial' Hedge's g	0.100
Mean Hedge's g in missing studies	0.000
Fail safe N	96

Therefore, although there is a difference between the mean effect sizes of high and low precise studies, we can safely conclude that the impact of publication bias on the results related to the third research question is trivial because all other indicators and tests show that the results are very robust to publication bias with no imputed studies as a result of TFM and unrealistic FSN numbers produced by Rosenthal's and Orwin's procedures.

4.2.3.3 Overall Mean Effect Size and Corresponding Statistical Test

Null Hypothesis: H_o: $\delta_3 = 0$

Mean of all true effect sizes for populations represented by the studies investigating the effect of PBL on students' attitudes toward science equals to zero.

The random effect model conducted to estimate the effectiveness of PBL on science achievement when compared to traditional teaching methods results in an overall effect size of 0.566 with the 95% confidence interval of 0.353 and 0.779, which is a medium effect. Furthermore, the null hypothesis related to research question three is rejected at the alpha level of 0.05 (p= 0.000), indicating that the mean of all true effect sizes is significantly different from zero. The details of the results for research question three is presented in Table 4.21.

4.2.3.4 Power Analysis

The variance for the point estimate of 0.566 is 0.012 based on random-effects model, so the parameter λ is:

$$\lambda = \frac{\delta}{\sqrt{V_\delta}} = \frac{0.566}{\sqrt{0.012}} = 5.193$$

Then, power is calculated with an alpha level of 0.05 as:

$$\text{Power} = 1 - \Phi\,(c_\alpha - \lambda) + \Phi\,(-c_\alpha - \lambda) = 0.999$$

Although the number of studies is smaller than the ones for the first and second research questions, the power of the statistical test is still very high in consequence of medium mean effect size revealed from these studies, the number of which is not small. If the standard error of the statistical test was 0.2 instead of 0.109, the power analysis would yield still appropriate but smaller value, 0.808. If the mean effect size was small, 0.1 with the same standard error of 0.109, then the power of statistical test would decrease an unacceptable value of 0.150.

The power of 1.000 means that the probability of Type II error; i.e. fail to detect a real treatment effect, is almost zero.

$$B = 1- Power = 0.001$$

Table 4.21 *Overall effect size details and corresponding statistical test for research question three.*

Model	Effect Size and 95% confidence interval						Statistical test	
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value
Fixed	23	0.514	0.048	0.002	0.420	0.607	10.714	0.000
Random	23	0.566	0.109	0.012	0.353	0.779	5.211	0.000

4.2.3.5 Heterogeneity Analysis

Different values for heterogeneity analysis are summarized in Table 4.22. Firstly, according to chi-squared significance test, the null hypothesis of "all studies share a common effect size" is rejected ($p < 0.05$), which means that the distribution of effect sizes shows heterogeneity indicating the possibility of moderator variables. In addition, I^2 statistic quantifies the heterogeneity on the data as indicating that 80% of total variance results from between study variance, which is labeled as high heterogeneity. High values related to tau-squared also supports this conclusion. T^2 quantifies the between study variance in the same metric with Hedge's g, which results in a high value of true variance as well. Corresponding tau value causes a 95% prediction interval of -0.416 to 1.548, meaning that 95% of cases the true effect size in a new study would fall inside in this prediction interval.

Table 4.22 *Heterogeneity test for research question three*

Heterogeneity				Tau-Squared			
Q-value	df(Q)	p-value	I^2	Tau-squared	Standard Error	Variance	Tau
109.823	22	0.000	79.968	0.212	0.083	0.007	0.461

4.2.4 The Results for Research Question Four

What is the effectiveness of PBL on motivational constructs in science when compared to traditional teaching methods?

4.2.4.1 Unit of Analysis

Each study investigating the effect of PBL on motivational constructs in science in this meta-analysis is accepted as the unit of analysis for the fourth research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. Eight primary studies are included to examine the fourth research question.

4.2.4.2 Publication Bias

As illustrated in Table 4.23, six of eight studies covered by the sample of fourth research question are either doctoral dissertations or master theses. The number of total effect sizes (10) is close to the total number of studies (8). Since the number of studies in each publication type is small, the mean values are not robust to the effect of confounding variable nevertheless; the mean effect size of doctoral dissertation is smaller than the one for journal article as predicted. Including dissertation and theses rather than including only journal articles for research question four increases representativeness of studies coded in meta-analysis while decreasing the possibility of publication bias.

Figure 4.14 shows the forest plot of the studies ranked in order of highest to lowest precision, which is inversely related to standard error of the studies. As

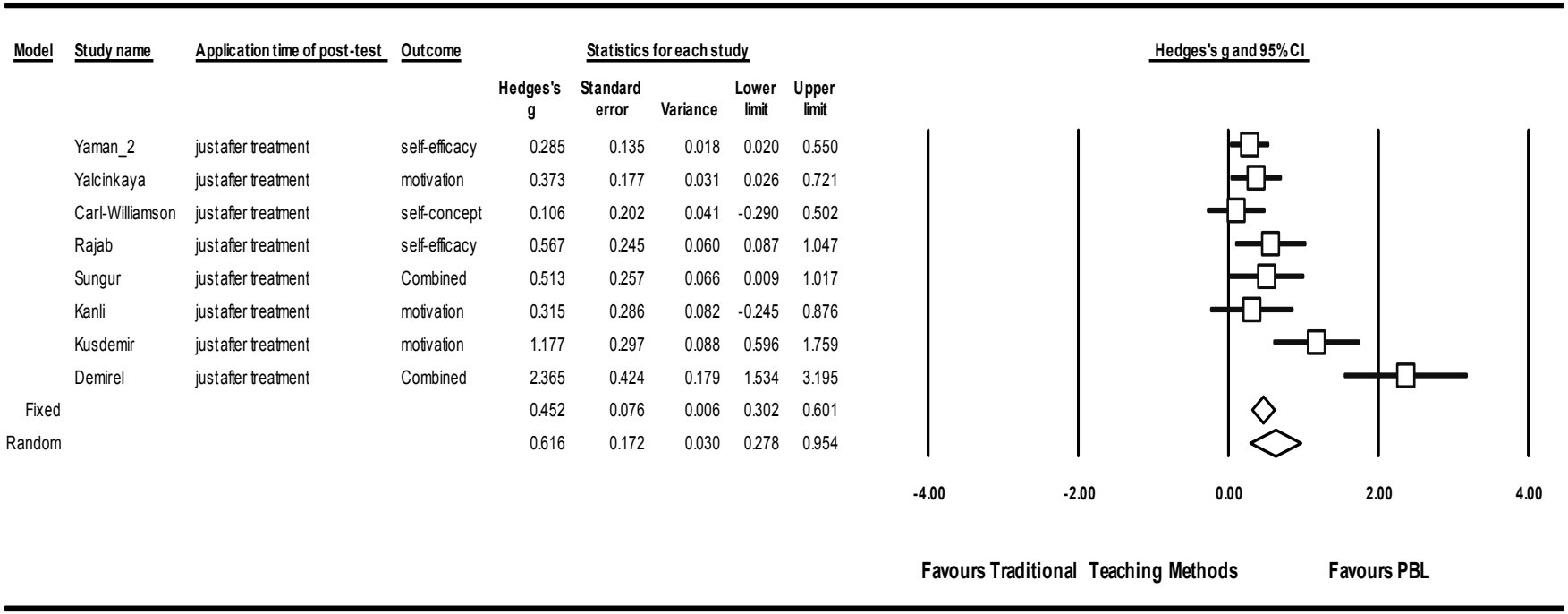

Figure 4.14 Forest plot for the studies in the sample of fourth research question, which are ranked based on their precisions in the order of highest to lowest precision.

Table 4.23 *The number of studies and effect sizes in different publication types and corresponding point estimate for research question four.*

Publication Type	Number of Study (Percentage)	Number of Effect Size	Point Estimate (Hedge's g)
Doctoral Dissertation	4 (50%)	5 (50%)	0.360
Master Thesis	2 (25%)	2 (20%)	0.731
Journal Article	2 (25%)	3 (30%)	0.477
Total	8 (100%)	10 (100%)	0.616

illustrated on the forest plot, the effect size values tend to increase with decreasing precision, which may result from a biased sample. Thus, we have to check the existence of publication bias by using further methods.

Figure 4.15 shows the funnel plot constructed upon random effect model by considering each study in the sample of studies as unit of analysis. The shape of the plot is close to be "funnel" shaped, which is the one resulting from unbiased data. Furthermore, TFM imputes only one additional study when looking for missing studies to the right the mean as illustrated on the figure while there is no imputed study when looking for the missing studies to the left of the mean. TFM results in an adjusted mean effect size value of 0.712, which is a bit larger than computed one from the sample. Thus, we can conclude that there is no publication bias in the direction indicated by forest plot, which is affected by an outlier at the bottom of the plot.

As summarized on Table 4.24, according to the results of Egger's Regression Test, the null hypothesis that *"there is no funnel plot asymmetry (β_0 = 0)" is* rejected ($p < 0.05$).

On the other hand, Rosenthal's FSN computed for the studies included in the meta-analysis for research question four, the details of which are presented in Table 4.25, reveals a ratio of 1.88, which confirms the robustness of the data to publication bias since it is larger than the cutoff point of one.

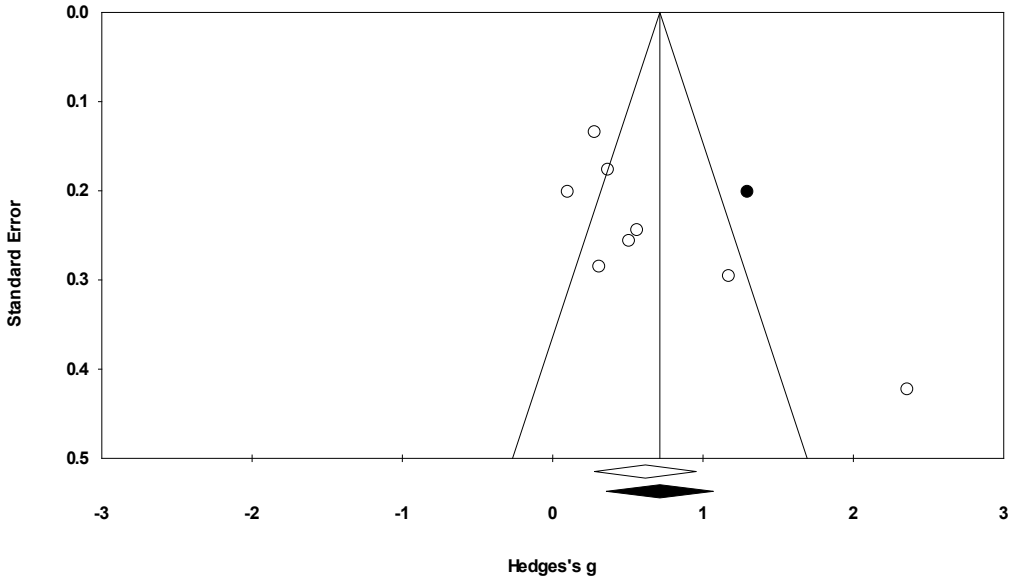

Figure 4.15 Funnel plot of the studies included in the sample of fourth research question based on random effect model

Table 4.24 *The results of Egger's Regression Test for all studies included in the sample of the fourth research question*

Intercept	4.75349
Standard error	1.67213
95% lower limit (2-tailed)	0.66193
95% lower limit (2-tailed)	8.84506
t value	2.84277
df	6
p value (2-tailed)	0.02946

Table 4.25 *Rosenthal's FSN for all studies included in the sample of the fourth research question*

Z-value for observed studies	6.96592
p-value for observed studies	0.00000
Alpha	0.05
Tails	2
Z for alpha	1.95996
Number of observed studies	8
Fail safe N	94

In addition, Orwin's FSN, the details of which are illustrated in Table 4.26, indicates that 29 additional studies with effect sizes of 0.00000 are necessary to bring the mean effect of this meta-analysis under 0.1. Again, this number is larger than the number of existing studies in the meta-analysis (8).

Table 4.26 *Orwin's FSN for all studies included in the sample of the fourth research question*

Hedge's g in observed studies	0.45180
Criterion for a 'trivial' Hedge's g	0.100
Mean Hedge's g in missing studies	0.000
Fail safe N	29

Therefore, although Egger's Regression Test indicates an existence of biased results, we can still conclude that the impact of publication bias on the results related to fourth research question is modest because all other indicators and tests show that the results are robust to publication bias with only one imputed study as a result of TFM, resulting in an adjusted value of 0.712, which is close to the computed one, 0.616. It should also be underlined that the bias is not in the direction of the one expected from publication bias. The adjusted value is bigger than the computed one, meaning that the mean effect size is underestimated in the sample, not overestimated as expected from the sample affect by publication bias. Thus, it can be concluded that the modest bias results from small number of studies, not from high proportion of published studies in the sample, which is supported by FSN values calculated by Rosenthal's and Orwin's procedures.

4.2.4.3 Overall Mean Effect Size and Corresponding Statistical Test

Null Hypothesis: H_o: $\delta_4 = 0$

Mean of all true effect sizes for populations represented by the studies investigating the effect of PBL on motivational constructs in science equals to zero.

The random effect model conducted to estimate the effectiveness of PBL on motivational constructs in science when compared to traditional teaching methods results in an overall effect size of 0.616 with the 95% confidence interval of 0.278 and 0.954, which is a medium to large effect. Furthermore, the null hypothesis

related to research question four is rejected at the alpha level of 0.05 (p= 0.000), indicating that the mean of all true effect sizes is significantly different from zero. The details of the results for research question four is presented in Table 4.27.

Table 4.27 *Overall effect size details and corresponding statistical test for research question four.*

Model	Effect Size and 95% confidence interval						Statistical test	
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value
Fixed	8	0.452	0.076	0.006	0.302	0.601	5.924	0.000
Random	8	0.616	0.172	0.030	0.278	0.954	3.573	0.000

4.2.4.4 Power Analysis

The variance for the point estimate of 0.616 is 0.030 based on random-effects model, so the parameter λ is:

$$\lambda = \frac{\delta}{\sqrt{V_\delta}} = \frac{0.616}{\sqrt{0.030}} = 3.581$$

Then, power is calculated with an alpha level of 0.05 as:

$$\text{Power} = 1 - \Phi(c_\alpha - \lambda) + \Phi(-c_\alpha - \lambda) = 0.948$$

Although the number of studies is quite small and the analysis is based on random-effects model rather than fixed-effect model, the power of the statistical test is still high in consequence of medium mean effect size revealed from these studies and small within study variance as a result of large total sample size. If the standard error of the statistical test was 0.2 instead of 0.172, the power analysis would yield still appropriate but smaller value, 0.869. If the mean effect size was small, 0.1 with the same standard error of 0.172, then the power of statistical test would decrease an unacceptable value of 0.090.

The power of 0.948 means that the probability of Type II error; i.e. fail to detect a real treatment effect, is only 5%.

$$B = 1 - \text{Power} = 0.052$$

4.2.4.5 Heterogeneity Analysis

Different values for heterogeneity analysis are summarized in Table 4.28. Firstly, according to chi-squared significance test, the null hypothesis of "all studies share a common effect size" is rejected ($p < 0.05$), which means that the distribution of effect sizes shows heterogeneity indicating the possibility of moderator variables. In addition, I^2 statistic quantifies the heterogeneity on the data as indicating that 77.8% of total variance results from between study variance, which is labeled as high heterogeneity. High values related to tau-squared also supports this conclusion. T^2 quantifies the between study variance in the same metric with Hedge's g, which results in a high value of true variance as well. Corresponding tau value results in a 95% prediction interval of -0.455 to 1.687, meaning that 95% of cases the true effect size in a new study would fall inside in this prediction interval.

Table 4.28 *Heterogeneity test for research question four*

Heterogeneity				Tau-Squared			
Q-value	df(Q)	p-value	I^2	Tau-squared	Standard Error	Variance	Tau
31.530	7	0.000	77.799	0.175	0.132	0.018	0.418

4.2.5 The Results for Research Question Five

What is the effectiveness of PBL on different types of skills when compared to traditional teaching methods?

4.2.5.1 Unit of Analysis

Each study investigating the effect of PBL on different types of skills including critical thinking skills, problem solving skills, science process skills, self-directed skills, meta-cognitive skills, inquiry learning skills, logical thinking skills, and self-regulation skills in this meta-analysis is accepted as the unit of analysis for the fifth research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for

each study. Totally 37 primary studies are included to examine the fifth research question.

4.2.5.2 Publication Bias

Table 4.29 shows the number of coded studies in different publication types including journal articles, doctoral dissertations and master theses. Totally 25 of 37 studies covered by the sample of fifth research question are either doctoral dissertations or master theses from which 29 effect size values emerge yielding smaller mean effect sizes in terms of Hedge's g than journal articles do as predicted. However, covering dissertation and theses rather than including only journal articles increases representativeness of studies coded in meta-analysis while decreasing the possibility of publication bias.

Table 4.29 *The number of studies and effect sizes in different publication types and corresponding point estimate for research question five*

Publication Type	Number of Study (Percentage)	Number of Effect Size	Point Estimate (Hedge's g)
Doctoral Dissertation	19 (51%)	23 (52%)	0.364
Master Thesis	6 (16%)	6 (14%)	0.725
Journal Article	12 (33%)	15 (34%)	0.765
Total	37 (100%)	44 (100%)	0.565

Figure 4.16 and 4.17 show the forest plots of the studies with high and low precision respectively. The groups are based on the categorization of primary studies according to their precision, which is inversely related to standard error of the studies. The first 19 studies with highest precision are assigned as the ones with high precision while the next 18 studies are grouped as low precision studies respectively. As summarized in Table 4.30, the mean effect size values calculated for the primary studies increase with decreasing precision, which may result from existence of missing studies.

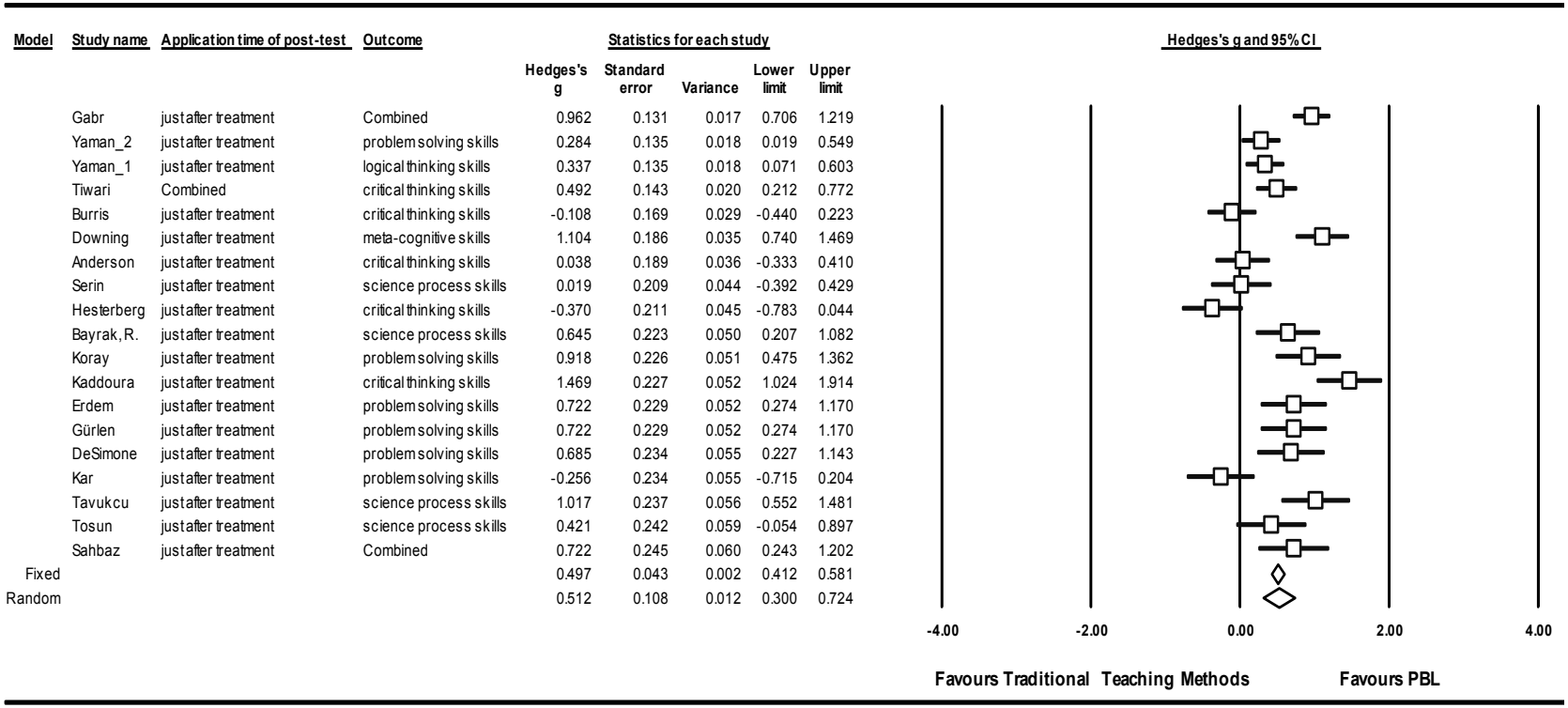

Figure 4.16 Forest plot for the first 19 studies when all studies included in the sample of fifth research question are ranked based on their precisions

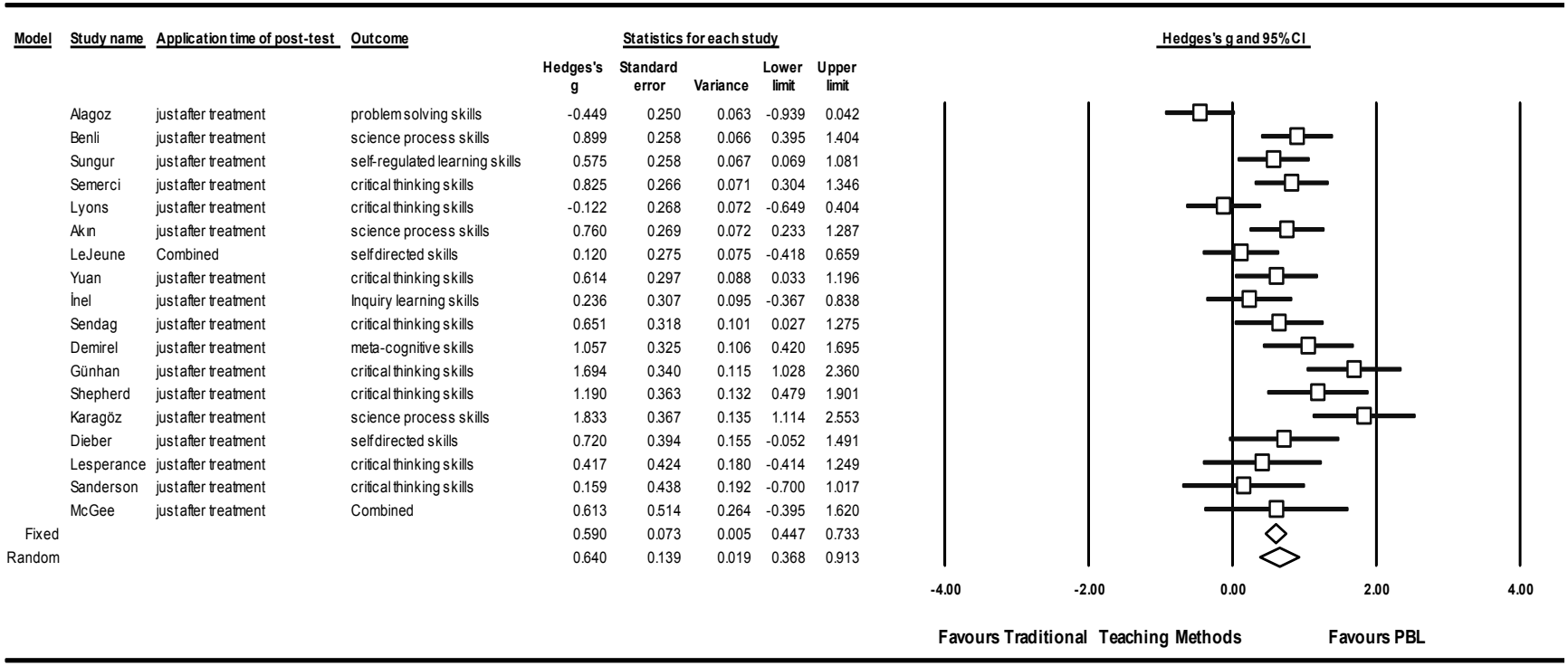

Figure 4.17 Forest plot for the last 18 studies when all studies included in the sample of fifth research question are ranked based on their precisions

Table 4.30 *Mean effect size values for the studies with high and low precision studies in the sample of the fifth research question*

Precision	Number of Study	Point Estimate (Hedge's g)
High	19	0.512
Low	18	0.640
Overall	37	0.565

Figure 4.18 shows the funnel plot constructed upon random effect model by considering each study in the sample of studies as unit of analysis. The shape of the plot seems to be a bit biased without the filled dots, which are imputed by TFM when looking for missing studies to the left the mean as illustrated on the figure while there is no imputed study when looking for the missing studies to the right of the mean. Thus, the mean effect size may be overestimated slightly due to publication bias, which is the reason why TFM offers a smaller adjusted mean effect size value of 0.423.

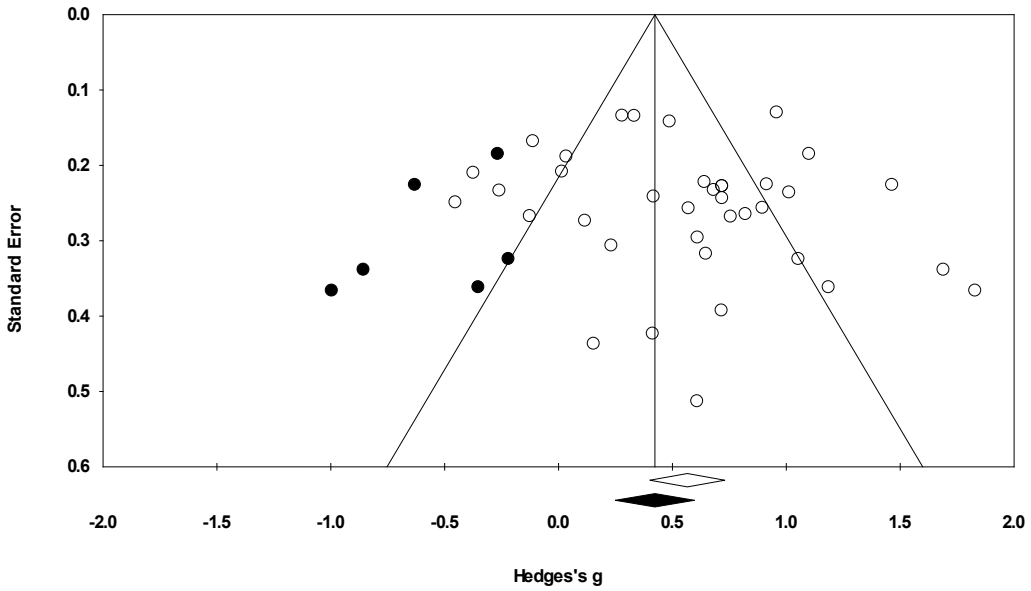

Figure 4.18 Funnel plot of the studies included in the sample of fifth research question based on random effect model

On the other hand, as summarized on Table 4.31, according to the results of Egger's Regression Test, the null hypothesis that *"there is no funnel plot asymmetry ($\beta_0 = 0$)"* cannot be rejected ($p > 0.05$).

Table 4.31 *The results of Egger's Regression Test for all studies included in the sample of the fifth research question*

Intercept	1.18592
Standard error	1.12801
95% lower limit (2-tailed)	-1.10405
95% lower limit (2-tailed)	3.47590
t value	1.05134
df	35
p value (2-tailed)	0.30031

In addition, Rosenthal's FSN computed for the studies included in the meta-analysis for research question five, the details of which are presented in Table 4.32, reveals a ratio of 9.54, which confirms the robustness of the data to publication bias since it is much larger than the cutoff point of one.

Table 4.32 *Rosenthal's FSN for all studies included in the sample of the fifth research question*

Z-value for observed studies	14.03231
p-value for observed studies	0.00000
Alpha	0.05
Tails	2
Z for alpha	1.95996
Number of observed studies	37
Fail safe N	1860

Similarly, Orwin's FSN, the details of which are illustrated in Table 4.33, indicates that 156 additional studies with effect sizes of 0.00000 are necessary to bring the mean effect of this meta-analysis under 0.1. Again, this number is larger than the number of existing studies in the meta-analysis (37).

Table 4.33 *Orwin's FSN for all studies included in the sample of the fifth research question*

Hedge's g in observed studies	0.52081
Criterion for a 'trivial' Hedge's g	0.100
Mean Hedge's g in missing studies	0.000
Fail safe N	156

As a result, although other tests and indicators show that the results seem to be robust to publication bias, the possibility that the results may be slightly biased cannot be eliminated completely since TFM offers a smaller adjusted mean effect size of 0.423, indicating that the mean effect size calculated from the sample of primary studies may be overestimated.

4.2.5.3 Overall Mean Effect Size and Corresponding Statistical Test

Null Hypothesis: H_o: $\delta_5 = 0$

Mean of all true effect sizes for populations represented by the studies investigating the effect of PBL on different types of skills equals to zero.

The random effect model conducted to estimate the effectiveness of PBL on skills when compared to traditional teaching methods results in an overall effect size of 0.565 with the 95% confidence interval of 0.401 and 0.730, which is a medium effect. Furthermore, the null hypothesis related to research question five is rejected at the alpha level of 0.05 (p= 0.000), indicating that the mean of all true effect sizes is significantly different from zero. The details of the results for research question five is presented in Table 4.34.

Table 4.34 *Overall effect size details and corresponding statistical test for research question five.*

Model	Effect Size and 95% confidence interval						Statistical test	
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value
Fixed	37	0.521	0.037	0.001	0.448	0.594	14.031	0.000
Random	37	0.565	0.084	0.007	0.401	0.730	6.741	0.000

4.2.5.4 Power Analysis

The variance for the point estimate of 0.565 is 0.007 based on random-effects model, so the parameter λ is:

$$\lambda = \frac{\delta}{\sqrt{V_\delta}} = \frac{0.565}{\sqrt{0.007}} = 6.726$$

Then, power is calculated with an alpha level of 0.05 as:

$$\text{Power} = 1 - \Phi\,(c_\alpha - \lambda) + \Phi\,(-c_\alpha - \lambda) = 1.000$$

The power of the statistical test is very high in consequence of medium mean effect size revealed from these studies and small within study variance as a result of large total sample size. If the standard error of the statistical test was 0.2 instead of 0.084, the power analysis would yield still appropriate but smaller value, 0.806. If the mean effect size was small, 0.1 with the same standard error of 0.084, then the power of statistical test would decrease an unacceptable value of 0.222.

The power of 1.000 means that the probability of Type II error; i.e. fail to detect a real treatment effect, is almost zero:

$$\beta = 1 - \text{Power} = 0$$

4.2.5.5 Heterogeneity Analysis

Different values for heterogeneity analysis are summarized in Table 4.35. Firstly, according to chi-squared significance test, the null hypothesis of "all studies share a common effect size" is rejected ($p < 0.05$), which means that the distribution of effect sizes shows heterogeneity indicating the possibility of moderator variables. In addition, I^2 statistic quantifies the heterogeneity on the data as indicating that 78.7% of total variance results from between study variance, which is labeled as high heterogeneity. High values related to tau-squared also supports this conclusion. T^2 quantifies the between study variance in the same metric with Hedge's g, which results in a high value of true variance as well. Corresponding tau value results in a 95% prediction interval of -0.337 to 1.467, meaning that 95% of cases the true effect size in a new study would fall inside in this prediction interval.

Table 4.35 *Heterogeneity test for research question five*

Heterogeneity				Tau-Squared			
Q-value	df(Q)	p-value	I^2	Tau-squared	Standard Error	Variance	Tau
169.283	36	0.000	78.734	0.191	0.065	0.004	0.437

4.3 Moderator Analysis

4.3.1 The Results for Research Question Six

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by publication type (doctoral dissertations, master theses and journal articles)?

4.3.1.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the sixth research question.

Mixed-effect analysis is conducted for this moderator analysis since subgroups are assumed to be fixed while it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are not pooled to calculate a common variance since there are more than five studies within each subgroup (Borenstein et al., 2009).

4.3.1.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of doctoral dissertations, master theses and journal articles are equal to each other.

Table 4.36 illustrates the results of heterogeneity analysis within subgroups and Table 4.37 summarizes the results of analog ANOVA conducted for the

moderator variable of publication types. As expected, the mean effect size for journal articles is much higher than the one for doctoral dissertations, which are, differently from journal articles, not necessarily published documents although often they are accepted to be published as well. As a result, the null hypothesis is rejected indicating that the mean effect sizes for publication types significantly differ from each other at the level of 0.05 (p=0.000).

Table 4.36 *The results of heterogeneity analysis within subgroups for publication types.*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Doc. Dissertation	179.253	34.000	0.000	81.032
Master Thesis	173.945	34.000	0.000	80.454
Journal Article	85.466	17.000	0.000	80.109
Total within	438.664	85.000	0.000	

Given Q_{total} is 438.664, df is 85, C_{total} is 1805.041 and T^2_{total} is 0.243, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.196$$

Then, the adjusted R^2 index can be calculated as:

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.194$$

In addition, overall I^2 is given as 83.434, so we can conclude that 19.4% of between study variance, which constitutes 83.4% of total variance, can be explained by the moderator variable of publication bias. However, it should be noted that both the results of chi-squared test and I^2 statistic indicates high degree of heterogeneity within each subgroups as illustrated in Table 4.36.

Table 4.37 *The results of mixed effect moderator analysis for publication type*

Publication Type	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Doc. Dissertation	35	0.356	0.093	0.009	0.173	0.539	3.808	0.000			
Master Thesis	18	0.753	0.125	0.016	0.508	0.997	6.038	0.000			
Journal Article	35	0.830	0.081	0.007	0.671	0.988	10.242	0.000			
Total Between									15.480	2	0.000
Overall	88	0.651	0.055	0.003	0.543	0.759	11.848	0.000			

4.3.2 The Results for Research Question Seven

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of research design (true experimental, quasi experimental with randomly assigned clusters and quasi experimental without randomly assigned clusters)?

4.3.2.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the seventh research question.

Mixed-effect analysis is conducted for this moderator analysis since subgroups are assumed to be fixed while it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are not pooled to calculate a common variance since there are more than five studies within each subgroup.

4.3.2.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of true experimental, quasi-experimental with randomly assigned clusters and quasi experimental without randomly assigned clusters are equal to each other.

Table 4.38 illustrates the results of heterogeneity analysis within subgroups and Table 4.39 summarizes the results of analog ANOVA conducted for the moderator variable of research designs. The mean effect size for true experimental studies is a bit smaller than quasi experimental studies with and without randomly assigned clusters, which have effect size values very close to each other. However, the null hypothesis cannot be rejected indicating that the mean effect sizes for

different research designs do not significantly differ from each other at the level of 0.05 (p=0.953).

Table 4.38 *The results of heterogeneity analysis within subgroups for research design.*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
True Experimental	251.972	47.000	0.000	81.347
QE with RAC	119.829	19.000	0.000	84.144
QE without RAC	151.207	19.000	0.000	87.434
Total within	523.009	85.000	0.000	

Furthermore, given Q_{total} is 523.009, df is 85, C_{total} is 1805.041 and T^2_{total} is 0.243, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.2427$$

Then, the adjusted R^2 index can be calculated as;

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.001$$

So, study design is responsible for only 0.1% of the between study variance, which shows that research design does not moderate the results of the study significantly. Furthermore, both the results of chi-squared test and I^2 statistic indicates high degree of heterogeneity within each sub-groups as illustrated in Table 4.38.

4.3.3 The Results for Research Question Eight

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of 'teacher effect' (same teacher or different teachers for control and experimental conditions)?

4.3.3.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one

Table 4.39 *The results of mixed effect moderator analysis for research design.*

Research Design	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
True Experimental	20	0.596	0.141	0.020	0.321	0.872	4.242	0.000			
QE with RAC	48	0.646	0.076	0.006	0.497	0.794	8.520	0.000			
QE without RAC	20	0.634	0.130	0.017	0.379	0.889	4.872	0.000			
Total Between									0.096	2	0.953
Overall	88	0.634	0.059	0.004	0.518	0.751	10.688	0.000			

effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the eighth research question. However, for the studies that do not provide information about teacher effect, the item is called as 'unspecified'.

Mixed-effect analysis is conducted for this moderator analysis since subgroups are assumed to be fixed while it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are not pooled to calculate a common variance since there are more than five studies within each subgroup.

4.3.3.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of studies in which same teacher or different teachers are assigned in experimental and control conditions are equal to each other.

Table 4.40 illustrates the results of heterogeneity analysis within subgroups and Table 4.41 summarizes the results of analog ANOVA conducted for the moderator variable of teacher effect. The mean effect sizes for the studies in which same teacher or different teachers are assigned in experimental and control conditions are almost equal while it is larger for the studies coded as 'unspecified'. However, the null hypothesis cannot be rejected indicating that the mean effect sizes for the subgroups of 'teacher effect' do not significantly differ from each other at the level of 0.05 (p=0.511).

Table 4.40 *The results of heterogeneity analysis within subgroups for teacher effect*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Different Teachers	148.4912	23	0.000	84.51087
Same Teacher	223.8142	40	0.000	82.12803
Unspecified	136.5377	22	0.000	83.88723
Total within	508.843	85	0.000	

Table 4.41 *The results of mixed effect moderator analysis for teacher effect.*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Different Teachers	24	0.588	0.118	0.014	0.357	0.820	4.978	0.000			
Same Teacher	41	0.594	0.083	0.007	0.431	0.756	7.164	0.000			
Unspecified	23	0.751	0.120	0.014	0.517	0.986	6.273	0.000			
Total Between									1.344	2.000	0.511
Overall	88	0.631	0.059	0.003	0.515	0.746	10.682	0.000			

Furthermore, given Q_{total} is 508.843, df is 85, C_{total} is 1805.041 and T^2_{total} is 0.243, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.235$$

Then, the adjusted R^2 index can be calculated as;

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.034$$

So, teacher effect is responsible for 3.4% of the between study variance, which seems to be a nontrivial proportion. However, the results are highly affected by unspecified data rather than the studies in which teacher effect is reported. The difference between the studies in which teachers in control and experimental conditions are same or different is only 0.01 in terms of Hedge's g, which is trivial. Furthermore, both the results of chi-squared test and I^2 statistic indicates high degree of heterogeneity within each sub-groups as illustrated on Table 4.40.

4.3.4 The Results for Research Question Nine

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of 'researcher effect' (whether researcher is any of teachers in experimental or control conditions)?

4.3.4.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the ninth research question. However, for the studies that do not provide information about researcher effect, the item is coded as 'unspecified'.

Mixed-effect analysis is conducted for this moderator analysis since subgroups are assumed to be fixed while it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group

estimates of between-study variance (tau-squared) are not pooled to calculate a common variance since there are more than five studies within each subgroup.

4.3.4.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of 'researcher effect' are equal to each other.

Table 4.42 illustrates the results of heterogeneity analysis within subgroups and Table 4.43 summarizes the results of analog ANOVA conducted for the moderator variable of researcher effect. The mean effect sizes for the studies in which researcher is one of the teachers or he/she is the only teacher in experimental and control conditions are almost equal while it is a bit larger for the studies in which researcher is not any of teachers and for the ones coded as 'unspecified'. However, the null hypothesis cannot be rejected indicating that the mean effect sizes for the subgroups of 'researcher effect' do not significantly differ from each other at the level of 0.05 (p=0.856).

Table 4.42 *The results of heterogeneity analysis within subgroups for researcher effect.*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Not any of teachers	201.579	35.000	0.000	82.637
One of teachers	25.394	9.000	0.003	64.559
The only teacher	117.958	18.000	0.000	84.740
Unspecified	168.549	22.000	0.000	86.947
Total within	513.480	84.000	0.000	

Furthermore, given Q_{total} is 513.480, df is 85, C_{total} is 1805.041 and T^2_{total} is 0.243, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.237$$

Table 4.43 *The results of mixed effect moderator analysis for researcher effect.*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Not any of teachers	36	0.643	0.090	0.008	0.466	0.820	7.130	0.000			
One of teachers	10	0.587	0.108	0.012	0.376	0.799	5.452	0.000			
The only teacher	19	0.547	0.139	0.019	0.274	0.820	3.926	0.000			
Unspecified	23	0.699	0.136	0.018	0.433	0.966	5.150	0.000			
Total Between									0.772	3.000	0.856
Overall	88	0.622	0.056	0.003	0.511	0.732	11.033	0.000			

Then, the adjusted R^2 index can be calculated as:

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.023$$

So, researcher effect is responsible for only 2.3% of the between study variance, which also shows that researcher effect does not moderate the results of the study significantly. Furthermore, both the results of chi-squared test and I^2 statistic indicates high degree of heterogeneity within each sub-groups as illustrated on Table 4.42.

4.3.5 The Results for Research Question Ten

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by the countries where the studies are conducted (Turkey, USA and others)?

4.3.5.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the tenth research question. However, since the numbers of studies conducted in the countries other than Turkey and the USA are very small in the sample of this meta-analysis, they are combined as 'others' representing different countries.

Fully random-effects analysis is conducted for this moderator analysis since subgroups are not assumed to be fixed and it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are pooled to calculate a common variance since it is the only option for fully random-effects analysis.

4.3.5.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of the studies conducted in different countries are equal to each other.

Table 4.44 illustrates the results of heterogeneity analysis within subgroups and Table 4.45 summarizes the results of analog ANOVA conducted for the moderator variable of country. The mean effect size for the studies conducted in Turkey is larger than the mean effect sizes for the ones in the USA and other countries. Especially, the difference between the mean effect sizes for the studies in Turkey and the USA is quite large. As a result of these mean differences, the null hypothesis is rejected indicating that the mean effect sizes for the subgroups of 'country' variable significantly differ from each other at the level of 0.05 (p=0.000).

Table 4.44 *The results of heterogeneity analysis within subgroups for country variable*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Turkey	61.395	10.000	0.000	83.712
USA	284.685	53.000	0.000	81.383
Others	95.233	22.000	0.000	76.899
Total within	441.313	85.000	0.000	

Furthermore, given Q_{total} is 441.313, df is 85, C_{total} is 1770.827 and T^2_{total} is 0.247, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.201$$

Then, the adjusted R^2 index can be calculated as:

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.172$$

So, country variable is responsible for 17.2% of the between study variance, which also shows that country variable moderates the results of the study significantly.

Table 4.45 *The results of fully random-effects moderator analysis for country variable.*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Turkey	54	0.812	0.071	0.005	0.673	0.952	11.439	0.000			
USA	23	0.207	0.113	0.013	-0.014	0.428	1.838	0.066			
Others	11	0.632	0.154	0.024	0.331	0.932	4.113	0.000			
Total Between									20.633	2	0.000
Overall	88	0.554	0.207	0.043	0.149	0.959	2.682	0.007			

On the other hand, both the results of chi-squared test and I^2 statistic indicates high degree of heterogeneity within each subgroup as illustrated on Table 4.44, which shows the existence of other moderator variables.

4.3.6 The Results for Research Question Eleven

Does the effectiveness of PBL in science when compared to traditional teaching methods differ by subject matter (physics, chemistry or biology)?

4.3.6.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. 58 primary studies which investigate the effectiveness of PBL in science are included to examine the eleventh research question. However, the subgroup of general science is excluded from the analysis since the number of studies covered by this subgroup is very small.

Fully random-effects analysis is conducted for this moderator analysis since subgroups are not assumed to be fixed and it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are pooled to calculate a common variance since it is the only option for fully random-effects analysis.

4.3.6.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of ‘subject areas’ are equal to each other.

Table 4.46 illustrates the results of heterogeneity analysis within subgroups and Table 4.47 summarizes the results of analog ANOVA conducted for the moderator variable of subject areas.

The mean effect size for the studies in Chemistry (0.952) is estimated as larger than the mean effect sizes for Physics (0.618) and Biology (0.457). As a

result of these mean differences, the null hypothesis is rejected indicating that the mean effect sizes for different subject areas in Science significantly differ from each other at the level of 0.05 (p=0.033).

Table 4.46 *The results of heterogeneity analysis within subgroups for subject areas.*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Biology	99.946	13.000	0.000	86.993
Chemistry	105.064	17.000	0.000	83.819
Physics	146.964	25.000	0.000	82.989
Total within	351.974	55.000	0.000	

Furthermore, given Q_{total} is 351.974, df is 55, C_{total} is 1248.303 and T^2_{total} is 0.267, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.238$$

Then, the adjusted R^2 index can be calculated as:

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.109$$

So, subject areas variable is responsible for 10.9% of the between study variance, which also shows that subject areas variable moderates the results of the study significantly. On the other hand, both the results of chi-squared test and I^2 statistic indicate high degree of heterogeneity within each subgroup as illustrated on Table 4.46, which shows the existence of other moderator variables.

4.3.7 The Results for Research Question Twelve

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by school level (primary, secondary and higher education)?

Table 4.47 *The results of fully random-effects moderator analysis for subject areas*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Biology	14	0.457	0.149	0.022	0.166	0.748	3.076	0.003			
Chemistry	18	0.952	0.132	0.018	0.693	1.212	7.195	0.000			
Physics	26	0.618	0.109	0.012	0.404	0.832	5.662	0.000			
Total Between									6.809	2	0.033
Overall	58	0.679	0.148	0.022	0.390	0.968	4.600	0.000			

4.3.7.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the twelfth research question. However, the subgroups of college and post-graduate levels are combined with undergraduate level since the numbers of studies covered by these subgroups are very small. This combined group is called as higher education level.

Mixed-effect analysis is conducted for this moderator analysis since subgroups are assumed to be fixed while it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are not pooled to calculate a common variance since there are more than five studies within each subgroup.

4.3.7.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of 'school level' are equal to each other.

Table 4.48 illustrates the results of heterogeneity analysis within subgroups and Table 4.49 summarizes the results of analog ANOVA conducted for the moderator variable of school level. The mean effect size for primary level is much larger than secondary level, which is also larger than higher education level. In other words, mean effect size decreases with increasing school level. However, the null hypothesis cannot be rejected indicating that the mean effect sizes for the subgroups of 'school level' do not significantly differ from each other at the level of 0.05 (p=0.182).

On the other hand, given Q_{total} is 490.742, df is 85, C_{total} is 1770.827 and T^2_{total} is 0.247, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.229$$

Then, the adjusted R^2 index can be calculated as:

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.072$$

So, school level variable is responsible for 7.2 % of the between study variance, which is not a negligible value. Thus, we can conclude the results of the studies are moderated by school level variable in some degree. On the other hand, both the results of chi-squared test and I^2 statistic indicate high degree of heterogeneity within each subgroup as illustrated in Table 4.48, which shows the existence of other moderator variables.

Table 4.48 *The results of heterogeneity analysis within subgroups for school level*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Primary	159.870	44.000	0.000	72.478
Secondary	194.257	25.000	0.000	87.130
Higher	136.615	16.000	0.000	88.288
Total within	490.742	85.000	0.000	

4.3.8 The Results for Research Question Thirteen

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of PBL mode (curriculum model or teaching method)?

4.3.8.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size,

Table 4.49 *The results of mixed effect moderator analysis for school level*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Primary	26	0.834	0.135	0.018	0.569	1.100	6.162	0.000			
Secondary	17	0.606	0.162	0.026	0.289	0.924	3.743	0.000			
Higher	45	0.559	0.063	0.004	0.435	0.682	8.864	0.000			
Total Between									3.412	2.000	0.182
Overall	88	0.607	0.054	0.003	0.502	0.713	11.276	0.000			

average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the thirteenth research question. PBL can be used as a curriculum model or as a teaching method, which constitute subgroups of the moderator of PBL mode.

Mixed-effect analysis is conducted for this moderator analysis since subgroups are assumed to be fixed while it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are not pooled to calculate a common variance since there are more than five studies within each subgroup.

4.3.8.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of 'PBL mode' are equal to each other.

Table 4.50 illustrates the results of heterogeneity analysis within subgroups and Table 4.51 summarizes the results of analog ANOVA conducted for the moderator variable of PBL mode. The mean effect size for the studies in which PBL is used as a teaching method is larger than the mean effect size for the ones which benefits PBL as a curriculum model. However, the null hypothesis cannot be rejected indicating that the mean effect sizes for the subgroups of 'PBL mode' do not significantly differ from each other at the level of 0.05 (p=0.215).

Table 4.50 *The results of heterogeneity analysis within subgroups for PBL mode*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Curriculum model	108.016	17.000	0.000	84.262
Teaching method	415.854	69.000	0.000	83.408
Total within	523.870	86.000	0.000	

On the other hand, given Q_{total} is 523.870, df is 86, C_{total} is 1770.827 and T^2_{total} is 0.247, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

Table 4.51 *The results of mixed effect moderator analysis for PBL mode*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Curriculum model	18	0.495	0.131	0.017	0.237	0.752	3.770	0.000			
Teaching method	70	0.678	0.068	0.005	0.545	0.811	9.981	0.000			
Total Between									1.537	1.000	0.215
Overall	88	0.639	0.060	0.004	0.521	0.757	10.597	0.000			

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.247$$

Then, the adjusted R^2 index can be calculated as;

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.00$$

So, we can safely conclude that the results of the studies are not moderated by PBL mode variable. In other words, the difference between mean effect sizes of the studies in which PBL is used as a curriculum model or teaching method is not significant. In addition, both the results of chi-squared test and I^2 statistic indicate high degree of heterogeneity within each subgroup as illustrated in Table 4.50.

4.3.9 The Results for Research Question Fourteen

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by length of treatment?

4.3.9.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the fourteenth research question. However, due to small number of studies covered by the subgroups of ’11-15 weeks’ and ‘over 15 weeks’, these groups are combined as ‘over 10 weeks’.

Fully random-effects analysis is conducted for this moderator analysis since subgroups are not assumed to be fixed and it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are pooled to calculate a common variance since it is the only option for fully random-effects analysis.

4.3.9.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of 'length of treatment' are equal to each other.

Table 4.52 illustrates the results of heterogeneity analysis within subgroups and Table 4.53 summarizes the results of analog ANOVA conducted for the moderator variable of length of treatment. The mean effect size for the studies in which treatment length is 6 to 10 weeks is larger than the mean effect sizes for the ones conducted during shorter (0-5 weeks) or much longer (over 10 weeks) period. In addition, the studies which do not specify the length of treatment reveals the largest mean effect size. However, the null hypothesis cannot be rejected indicating that the mean effect sizes for the subgroups of 'length of treatment' do not significantly differ from each other at the level of 0.05 (p=0.239).

Table 4.52 *The results of heterogeneity analysis within subgroups for length of treatment*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
0-5 weeks	171.607	31.000	0.000	81.935
6-10 weeks	168.666	25.000	0.000	85.178
Over 10 weeks	90.823	18.000	0.000	80.181
Unspecified	59.383	10.000	0.000	83.160
Total within	490.479	84.000	0.000	

On the other hand, given Q_{total} is 490.479, df is 84, C_{total} is 1770.827 and T^2_{total} is 0.247, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.230$$

Then, the adjusted R^2 index can be calculated as;

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.070$$

Table 4.53 *The results of fully random-effects moderator analysis for length of treatment*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
0-5 weeks	32	0.613	0.099	0.010	0.419	0.807	6.207	0.000			
6-10 weeks	26	0.682	0.110	0.012	0.467	0.897	6.228	0.000			
Over 10 weeks	19	0.480	0.127	0.016	0.232	0.729	3.792	0.000			
Unspecified	11	0.898	0.166	0.028	0.572	1.224	5.400	0.000			
Total Between									4.217	3.000	0.239
Overall	88	0.651	0.088	0.008	0.479	0.823	7.415	0.000			

Thus, length of treatment variable is responsible for 7.0 % of the between study variance, which is not a negligible value. Thus, we can conclude the results of the studies are moderated by length of treatment variable in some degree. On the other hand, both the results of chi-squared test and I^2 statistic indicate high degree of heterogeneity within each subgroup as illustrated in Table 4.52, which shows the existence of other moderator variables.

4.3.10 The Results for Research Question Fifteen

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by group size?

4.3.10.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. All of 88 primary studies are included to examine the fifteenth research question. However, due to small number of studies covered by some subgroups, the groups in the scope of this moderator variable are reorganized into two categories, which are '0-6 students' and 'over 6 students'.

Fully random-effects analysis is conducted for this moderator analysis since subgroups are not assumed to be fixed and it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are pooled to calculate a common variance since it is the only option for fully random-effects analysis.

4.3.10.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of 'group size' are equal to each other.

Table 4.54 illustrates the results of heterogeneity analysis within subgroups and Table 4.55 summarizes the results of analog ANOVA conducted for the

moderator variable of group size. The mean effect size for the studies in which group size is smaller than six is larger than the mean effect sizes for the ones having group size of larger than six students. In addition, the studies which do not specify group size reveal a mean effect size between these values. However, the null hypothesis cannot be rejected indicating that the mean effect sizes for the subgroups of 'group size' do not significantly differ from each other at the level of 0.05 (p=0.409).

Table 4.54 *The results of heterogeneity analysis within subgroups for group size*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
0-6	230.771	43.000	0.000	81.367
Over 6	86.746	19.000	0.000	78.097
Unspecified	189.133	23.000	0.000	87.839
Total within	506.650	85.000	0.000	

On the other hand, given Q_{total} is 506.650, df is 85, C_{total} is 1770.827 and T^2_{total} is 0.247, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.238$$

Then, the adjusted R^2 index can be calculated as;

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.036$$

Thus, group size variable is responsible for 3.6% of the between study variance, which is not a negligible value. Thus, we can conclude the results of the studies are moderated by group size variable in some degree. On the other hand, both the results of chi-squared test and I^2 statistic indicate high degree of heterogeneity within each subgroup as illustrated in Table 4.54, which shows the existence of other moderator variables.

Table 4.55 *The results of fully random-effects moderator analysis for group size*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
0-6	44	0.716	0.085	0.007	0.550	0.882	8.447	0.000			
Over 6	20	0.525	0.122	0.015	0.285	0.765	4.291	0.000			
Unspecified	24	0.603	0.116	0.013	0.376	0.830	5.199	0.000			
Total Between									1.790	2	0.409
Overall	88	0.633	0.072	0.005	0.491	0.774	8.779	0.000			

4.3.11 The Results for Research Question Sixteen

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of questions in the assessment instrument?

4.3.11.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size, average of these effect sizes are calculated to result in only one estimate for each study. 87 primary studies are included to examine the sixteenth research question while one of the studies is excluded from the analysis since it does not provide necessary information to code this variable and a subgroup with only one study changes the results of statistical test dramatically.

Mixed effect analysis is conducted for this moderator analysis since subgroups are assumed to be fixed and it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are not pooled to calculate a common variance although one of the subgroups include only three studies since the estimation of tau-squared for other two groups with adequate number of studies are quite different.

4.3.11.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of ‘type of questions’ variable are equal to each other.

Table 4.56 illustrates the results of heterogeneity analysis within subgroups and Table 4.57 summarizes the results of analog ANOVA conducted for the moderator variable of ‘type of questions’. The mean effect size for the studies in which only open-ended questions are used for assessment is larger than the mean effect size of the ones which uses both open-ended and objective type questions for

assessment while the smallest mean effect size is revealed from the studies which uses only objective type questions in the assessment instruments. However, the null hypothesis cannot be rejected indicating that the mean effect sizes for the subgroups of 'type of questions' do not significantly differ from each other at the level of 0.05 (p=0.389).

Table 4.56 *The results of heterogeneity analysis within subgroups for type of questions variable*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Both	82.653	20.000	0.000	75.803
Only objective	405.065	62.000	0.000	84.694
Only open-ended	6.674	2.000	0.036	70.035
Total within	494.393	84.000	0.000	

On the other hand, given Q_{total} is 494.393, df is 84, C_{total} is 1720.152 and T^2_{total} is 0.249, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.239$$

Then, the adjusted R^2 index can be calculated as;

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.042$$

Thus, this variable is responsible for 4.2% of the between study variance, which is which is not a negligible value. Thus, we can conclude the results of the studies are moderated by 'type of questions' variable in some degree. On the other hand, both the results of chi-squared test and I^2 statistic indicate high degree of heterogeneity within each subgroup as illustrated in Table 4.56, which shows the existence of other moderator variables.

Table 4.57 *The results of mixed effect moderator analysis for 'type of questions' variable*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Both	21	0.744	0.106	0.011	0.535	0.952	6.988	0.000			
Only objective	63	0.590	0.073	0.005	0.447	0.733	8.088	0.000			
Only open-ended	3	0.837	0.283	0.080	0.283	1.390	2.961	0.003			
Total Between									1.889	2	0.389
Overall	87	0.647	0.059	0.003	0.532	0.763	11.006	0.000			

4.3.12 The Results for Research Question Seventeen

Does the effectiveness of PBL on different outcomes when compared to traditional teaching methods differ by types of assessment instrument (pre-existing, researcher developed or adapted)?

4.3.12.1 Unit of Analysis and Model

Each study included in the meta-analysis rather than each effect size provided by these studies is accepted as the unit of analysis for this specific research question, meaning that each primary study provides one and only one effect size for the calculations. For the studies giving out more than one effect size using the same instrument, average of these effect sizes are calculated to result in only one estimate for each study. However, for the studies offering more than one effect size using different types of instrument, only one of the outcomes is included in the meta-analysis, rather than averaging them, to be able to put studies into one of the subgroups of this moderator variable. As a result, 88 effects sizes revealed from 88 primary studies are included into the analysis to examine the seventeenth research question.

Mixed effect analysis is conducted for this moderator analysis since subgroups are assumed to be fixed and it is highly possible that more than a single true effect size exist within the subgroups. In addition, within-group estimates of between-study variance (tau-squared) are pooled to calculate a common variance since one of the subgroups includes only five primary studies.

4.3.12.2 Analog ANOVA

Null Hypothesis:

Means of all true effect sizes within subgroups of ‘type of assessment instrument’ variable are equal to each other.

Table 4.58 illustrates the results of heterogeneity analysis within subgroups and Table 4.59 summarizes the results of analog ANOVA conducted for the moderator variable of ‘type of assessment instrument’. The mean effect size for the studies in which adapted test is used as assessment instrument is larger than the mean effect size of the ones which uses researcher developed tests while the smallest mean effect size is revealed from the studies in which pre-existing test is

used as assessment instrument. The null hypothesis is rejected indicating that the mean effect sizes for the subgroups of 'type of assessment instrument' significantly differ from each other at the level of 0.05 (p=0.011).

Table 4.58 *The results of heterogeneity analysis within subgroups for type of assessment instrument variable*

Subgroups	Heterogeneity			
	Q-value	df (Q)	P-value	I-squared
Adapted	8.260	4.000	0.083	51.573
Pre-existing	256.205	41.000	0.000	83.997
Researcher-developed	357.105	40.000	0.000	88.799
Total within	621.570	85.000	0.000	

On the other hand, given Q_{total} is 621.570, df is 85, C_{total} is 1760.521 and T^2_{total} is 0.326, T^2_{within}, which is the pooled variance across subgroups, is calculated as:

$$T^2_{within} = \frac{Q_{total} - df}{C_{total}} = 0.305$$

Then, the adjusted R^2 index can be calculated as;

$$R^2 = 1 - \frac{T^2_{within}}{T^2_{total}} = 0.065$$

Thus, type of assessment instrument variable is responsible for 6.5% of the between study variance, which also shows that this variable moderates the results of the study significantly. On the other hand, both the results of chi-squared test and I^2 statistic indicate high degree of heterogeneity within each subgroup as illustrated in Table 4.58, which shows the existence of other moderator variables.

Table 4.59 *The results of mixed effect moderator analysis for 'type of assessment instrument' variable*

Subgroups	Effect Size and 95% confidence interval						Statistical test		Heterogeneity		
	Number of Studies	Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	Z-value	p-value	Q-value	df(Q)	P value
Adapted	5	1.037	0.276	0.076	0.497	1.578	3.760	0.000			
Pre-existing	42	0.488	0.096	0.009	0.300	0.676	5.098	0.000			
Researcher-developed	41	0.856	0.096	0.009	0.668	1.044	8.911	0.000			
Total Between									9.030	2.000	0.011
Overall	88	0.692	0.066	0.004	0.563	0.821	10.512	0.000			

CHAPTER V

DISCUSSION, CONCLUSIONS AND IMPLICATIONS

This chapter starts with a summary of the meta-analysis including main steps followed during this study. Then, the results related to each research question are discussed and conclusions are stated explicitly, which is followed by explanations for validity and reliability issues. After explaining limitations of the study, the chapter is ended with presenting the implications and suggestions for further research.

5.1 Summary of the Study

The main goal of this meta-analysis was to investigate the effect of PBL on not only students' achievement and motivation in science, but also attitudes toward science and different types of skills. The effectiveness of PBL on different outcomes was also examined by getting all outcomes together to construct a bigger picture illustrating how much PBL was effective in educational settings. Furthermore, 12 variables were checked whether they moderated the overall effectiveness of PBL.

This meta-analysis was started with constructing a framework for research questions and deciding the inclusion/exclusion criteria for primary studies. Then, an extensive literature search was conducted and the first draft of coding sheet was developed. Constructing the items in coding sheet was an iterative process of trial coding and revising if necessary. After final versions of coding sheet and coding manual were developed and the primary studies to be included in the meta-analysis was selected, all of the primary studies were coded by the researcher and the analysis was conducted by the help of CMA. 147 effect size values revealed from

88 primary studies were included in the meta-analysis and each primary studies was set as unit of analysis in different levels of meta-analysis. Besides examining five main effects, 12 variables were put into moderator analysis using analog ANOVA to investigate to what extent these variables affect the effectiveness of PBL in educational contexts.

5.2 Discussion of the Results

5.2.1 Discussion of the Main Effect Analyses

As summarized in Table 5.1, five main effects were calculated in terms of the effectiveness of PBL comparing to traditional teaching methods. Firstly, the results of the meta-analysis show clearly that PBL is more effective than traditional teaching methods on the whole, i.e. on different outcomes including not only content and skills outcomes but also attitudinal and motivational constructs in science. The overall estimate of the mean effect size was calculated as 0.633, which is a medium effect size, when 147 effect sizes revealed from 88 studies were analyzed under random-effects model. Narrow confidence interval indicates the existence of small variance for the mean effect size while wider prediction interval results from high heterogeneity of the data. Finally, very high statistical power assures that there is almost no probability of doing Type II error. Small variance and high power are caused by large number of studies and medium overall effect size. Despite losing some details, covering different types of outcomes provides a big picture to create generalizations, which is the major aim of any type of research synthesis (Chalmers et al., 2002; Cooper & Hedges, 2009; Davies, 2000; Petticrew & Roberts, 2006). This result does not guarantee that PBL works very well in all types of contexts in all educational settings; however, it indicates that PBL is an effective teaching method at least for some outcomes in some contexts, the details of which should be investigated as well.

When the outcomes were analyzed separately as science achievement, attitudes toward science, motivational constructs and skills, effect size of at least medium-size was estimated for each of the outcomes. Since the number of studies included in the analysis for each outcome decreased, wider confidence intervals were constructed comparing to overall effect size analysis. As illustrated by

Table 5.1 *Summary of the results for main effects*

Dependent Variable	Number of Studies	Publication bias	Mean Effect Size	95% Confidence Interval		95% Prediction Interval		Statistical Power
				Lower Limit	Upper Limit	Lower Limit	Upper Limit	
Overall	88	trivial	0.633 *0.598**	0.517 *0.553*	0.749 *0.644*	-0.353	1.619	1.000
Science achievement	52	trivial	0.820 *0.739*	0.631 *0.679*	1.010 *0.799*	-0.498	2.138	1.000
Attitudes toward science	23	trivial	0.566 *0.514*	0.353 *0.420*	0.779 *0.607*	-0.416	1.548	0.999
Motivational constructs in science	8	modest	0.616 **(0.712)**** *0,452*	0.278 *0.302*	0.954 *0.601*	-0.455	1.687	0.948
Skills	37	modest	0.565 **(0.423)** *0.521*	0.401 *0.448*	0.730 *0.594*	-0.337	1.467	1.000

*Italic items presents the values calculated under fixed effect model

**Bold items are adjusted effect size values for the corresponding calculated ones due to possibility of modest publication bias.

prediction intervals, high heterogeneity existed in each sample for individual outcomes. Furthermore, although it is not uncommon to have relatively low statistical power as a result of the analyses under random-effects model especially the number of studies are small (Borenstein et al., 2009; Pigott, 2012), statistical power revealed from each of analyses for the main effects was very high, ranging from 0.948 to 1.000, even for the one covering only eight primary studies, which was mainly caused by magnitude of the effect size estimated, 0.565 for the smallest one.

Another important point to be underlined about these analyses is that the subgroups of outcomes are not exclusive because of two main reasons. Firstly, what is called by science achievement by primary studies may cover some skills as well based on the scope of their assessment instruments. The second reason is inherent to the nature of the constructs covered by this meta-analysis. For example, meta-cognitive skills as a member of 'skills' construct is sometimes assessed under another construct, self-regulation skills, both of which are labeled as motivational constructs as well. Thus, the aim of this meta-analysis is not to construct an exclusive classification of these constructs; rather it aims to make contributions to cumulative knowledge so that each researcher or practitioner can reach some generalizations about the effectiveness of PBL as a teaching method on the outcomes to be interested in.

The sample of science achievement outcome covers 52 primary studies including 57 effect size values and reveals the largest effect size estimate in this meta-analysis, 0.820, which is a large effect size according to the classification of Cohen (1988). This construct mainly focuses on content knowledge rather than explicit types of skills, therefore it can be used to compare the effectiveness of PBL on science achievement based on content knowledge and on 'skills', which includes the studies investigating the effect of PBL on the development of different types of skills explicitly. It should also be noted that this construct covers all three levels of knowledge structure stated by Gijbels et al. (2005), which are "understanding concepts", "understanding of the principles that link concepts" and " linking of concepts and principles to conditions and procedures for application"; however, it is not possible the effect of PBL on these levels of knowledge structure separately

since neither separate scores nor raw data to calculate these scores was provided by primary studies included in this meta-analysis.

Next, the overall effect size for the effectiveness of PBL on the students' attitudes toward science was estimated by the sample of 23 primary studies and a medium effect size of 0.566 was calculated as a result of random-effects model with a confidence interval, which is again in medium-size region. Despite small number of studies included in the analysis, there was no nontrivial effect of publication bias; however, the results showed high heterogeneity although this sample only included the primary studies from science education, indicating again the existence of moderator variables.

The analysis conducted to obtain an overall mean effect size for motivational constructs suffered from the small numbers of primary studies included, yielding a modestly biased effect size estimate of 0.616, which was corrected by TFM as 0.712. However, it should be underlined that the bias is not in the direction that is caused by publication bias. The effect size is expected to be overestimated with the existence of publication bias, whereas in this analysis the effect size was underestimated and corrected to a higher level by TFM. That is, the bias might result from some missing studies, or even may not exist at all. Motivational constructs in this analysis consist of meta-cognitive skills, self-concept, self-efficacy and self-regulated learning skills. Koballa and Glynn (2007) suggest motivation researcher that they should avoid simple categorizations and "they should adapt broader perspectives that serve to synthesize orientations and constructs" (p. 94). Thus, neither providing a framework for motivational constructs nor estimating a mean effect size for each of outcomes building up motivation is one of the aims of this meta-analysis, rather, one of the major aims of this study is to provide an overall mean effect size that summarizes the effect of PBL on students' motivation in science education. A final point to be emphasized about this construct that the number of primary studies investigating the effectiveness of PBL on motivation in science is limited in the literature, therefore there is a clear need for further primary studies to be able to conduct more comprehensive research synthesis about the relationship between PBL and motivation in science.

The ‘skills’ outcome consists of different types of skills as well, including critical thinking skills, inquiry learning skills, logical thinking skills, meta-cognitive skills, problem solving skills, science process skills, self-directed and self-regulated learning skills. 37 studies examining the effectiveness of PBL on these skills in science was selected as the sample of primary studies for this analysis. The results seemed to be affected by modest publication bias, therefore, the calculated effect size, 0.565 was adjusted as 0.423 by TFM. Whichever effect size was more correct estimation of mean of true effect sizes, the result seemed to be interesting since the mean effect size for skills was smaller than the one revealed from the studies focusing mainly on content rather than skills explicitly, which is not parallel what is stated in the literature about the effectiveness of PBL (Gijbels et al., 2005). Since achievement outcome consisted of only studies of science subject whereas there was no subject area restriction for skills outcome, I conducted another analysis including only 14 studies of science subject area to check whether the results were affected by the variable of subject areas. However, the mean effect size for these studies, 0.576, was very close to the previously calculated one for studies of different subject areas under the same model for analysis.

Actually, the mean effect size for skills outcome is only slightly larger than the ones stated by different studies in the literature (Dochy et al., 2003; R. A. Smith, 2003). On the other hand, the effect size estimated for science achievement is clearly larger than previously estimated mean effect sizes by other researchers (Dochy et al., 2003; Gijbels et al., 2005; Kalaian et al., 1999). Subject areas may be responsible for this case because previously conducted meta-analyses focus on medical education, rarely including studies from other disciplines while in this meta-analysis primary studies from only science disciplines were examined. Thus, we can safely predict that PBL works better in science education comparing to medical education in terms of students’ achievement. This difference may not be surprising because of nature of science education, which is dominated by “principles that link concepts”, and Gijbels et al. (2005) claim that PBL has the greatest positive effect on the second level of knowledge structure, i.e. “underlying principles that link concepts”. They estimate the effect size in this level of

knowledge structure as 0.795, which is close to the one calculated for science achievement in this meta-analysis; 0.820.

The results of this study are based on the random-effects model, however, the results for mean effect size and corresponding confidence intervals under fixed effect model are also presented in Table 5.1 to examine how sensitive the results are to the model the analysis based on. As illustrated in the table, the mean effect sizes estimated by fixed effect model are consistently smaller than the ones calculated under random-effects model, which is not general rule but results from the difference how two models weighted the primary studies. Fixed effect model assigns more weight to the studies with larger samples while it underestimates the information from the studies with small samples to make best prediction for one true effect size. On the other hand, the primary studies are weighted in a more balanced way in random-effects model to estimate the mean of distributions of effect sizes (Borenstein et al., 2009). In this meta-analysis, small sample studies reveal slightly larger effect sizes, which increases mean effect sizes in random-effect model in which they are assigned with more weight comparing to fixed effect model. As expected, fixed effect model, which has the assumption of no between-study variance across primary studies, results in narrower confidence intervals, overestimating the precision of results for main effects. Finally, high prediction intervals confirm once again that fixed effect model is not appropriate for this data set.

Finally, it is important to note that what is estimated as main effects in random-effects model is the mean of a distribution of effects since more than one true effect sizes representing different populations are assumed to exist rather than one true effect size, which is the assumption of fixed-effect model in which, therefore, it is possible to estimate single effect size. Thus, in this meta-analysis all point estimates should be interpreted as a mean of various true effect sizes.

5.2.2 Discussion of the Moderator Analyses

After estimating main effects, 12 variables were examined whether they moderate the results of the meta-analysis for main effects. The results of moderator analyses are summarized in Table 5.2 in the order of decreasing percentage of

variance explained by the variable. Publication type is the moderator variable that explains the largest proportion of total between-study variance while PBL mode explains almost no between-study variance.

Table 5.2 *Summary of the results for moderator analysis*

Moderator Variable	Number of Studies	Model	p value	R-squared	% of variance explained
Publication type	88	mixed-not pooled	0.000*	0.194	19.4
Country	88	random-pooled	0.000*	0.172	17.2
Subject area	58	random-pooled	0.033*	0.109	10.9
School level	88	mixed-not pooled	0.182	0.072	7.2
Length of treatment	88	random-pooled	0.239	0.070	7.0
Assessment instrument	88	mixed- pooled	0.011*	0.065	6.5
Type of questions	87	mixed-not pooled	0.389	0.042	4.2
Group size	88	random-pooled	0.409	0.036	3.6
Teacher effect	88	mixed-not pooled	0.511	0.034	3.4
Researcher effect	88	mixed-not pooled	0.856	0.023	2.3
Research design	88	mixed-not pooled	0.095	0.001	0.1
PBL mode	88	mixed-not pooled	0.215	0.000	0.0

*Significant at the alpha level of 0.05

It is evident from the literature that published studies reveals larger effect size than unpublished ones (Rothstein et al., 2005). The findings of this meta-analysis support this claim, estimating much larger mean effect size for journal articles (0.830) comparing to doctoral dissertations (0.356). However, what is more interesting is that master theses have a large mean effect size (0.753) as well, which is much bigger than the one for doctoral dissertations. This may result from the fact that in dissertations, possible confounding variables are controlled better comparing to master theses. If it is correct, the estimated mean effect sizes can be interpreted as overestimated but it is impossible to reach such a conclusion since the variable of how much the confounding variables are controlled could not be coded because of the reasons previously explained in Chapter 3. Country is another obvious variable that moderates the results of the study. The mean effect size is much larger for the

studies conducted in Turkey (0.812) comparing to the studies conducted in the USA (0.207). This result may be affected by other moderator variables, for example the proportion of studies in primary level to the ones in higher levels is larger for the studies in Turkey than the ones in the USA. However, the effect of country variable is much larger than school level variable, so it is not possible to explain this difference only by confounding variables.

Subject area is another moderator variable, which explains a high proportion of total between-study variance as well. However, it should be underlined that this between-study variance is not identical to the variance revealed from other variables since the sample of this moderator variable is different from the ones for other variables. Nevertheless, it explains high proportion of variance, yielding statistically significant results as well. The results indicate a much larger estimated effect size for chemistry (0.952) comparing to physics (0.618) and biology (0.457). To control the country effect on this variable, I conducted another moderator analysis including only the studies conducted in Turkey. Although the differences between subject areas decreased slightly, the order did not change with the largest effect size estimate for Chemistry (1.015), and smaller effect sizes for Physics (0.739) and Biology (0.702). In addition, estimate of effect size for each subject area increased, as expected, when the country was kept constant as Turkey. However, it is obvious that subject area is another moderator variable that affects the effectiveness of PBL.

School level variable explains a noteworthy amount of between-study variance although the results are not statistically significant, which is highly affected by many variables like the categories constructed for the variable. For example, when higher education is separated into college, undergraduate and postgraduate levels, the results turn into significant, but the analysis was conducted with three levels of education because otherwise the number of studies would be very small for some levels. The effect of school level is quite obvious; the effect size values decreases with increasing school level. For example, for primary school level, the estimated effect size is 0.834, which is a large effect size while it decreases to medium effect sizes for secondary (0.606) and higher education levels (0.559).

Length of treatment is another variable, which explains a remarkable amount of variance but corresponding results are not statistically significant. The results show that the effect of PBL has the largest value when the length of treatment is six to ten weeks (0.682). The effect size values for both shorter (0.613) and longer period (0.480) are smaller comparing to the one for this length of treatment. A smaller effect size value for shorter period is understandable because it may take some time for students to get used to a new teaching method, which is completely different from how they are normally instructed. One of the reasons for decrease for longer periods may be diminishing effect of novelty. Difficulty in controlling all other variables for a very long period may also affect the effectiveness of teaching method.

Another moderator variable for the effectiveness of PBL is the type of assessment instruments, which reveals statistically significant results as well. It is highly obvious that the studies which use adapted or researcher developed tests for assessment reports much larger effect sizes than the studies using pre-existing tests. If we assume that pre-existing tests are more reliable than researcher developed tests, the results might be overestimated because of this moderator effect. On the other hand, researcher–developed test may be more valid since it is developed for this specific context, which affects the results in opposite direction. It is impossible to make empirical decisions about either hypothesis; however, we can clearly conclude that type of assessment instrument highly affects the measured effectiveness of PBL.

Another moderator related to assessment is 'type of questions' variable, which examines whether type of questions in the assessment instruments affect the study results. Although the results of moderator analysis are not statistically significant, the pattern is very obvious: when open-ended questions are used in the assessment instrument, the studies reveal larger effect size values. That is highly reasonable because higher level of knowledge structure are assessed better by means of open-ended questions rather than objective type questions and PBL is more effective on higher level of knowledge structure rather than understanding concepts (Gijbels et al., 2005). This finding is parallel with what is stated by Dochy et al. (2003) about the effect of type of assessment method on the effectiveness of

PBL. Based on more dramatic results, they similarly claim that PBL results in larger effect sizes when the assessment instrument includes modified essay type (0.476) or essay type (0.165) questions rather than multiple choice questions (-0.309).

Group size is another moderator variable, which reveals a reasonable pattern with nontrivial explained variance but corresponding findings are statistically non-significant. The results seem to be highly affected by a large number of "unspecified" data. However, the findings can be interpreted as PBL works better when group size is smaller than six (0.716) comparing to larger groups.

On the other hand, teacher effect, researcher effect, research design and PBL mode are the variables that have slight or trivial effects on the study results. Teacher effect seems to be explaining some portion of variance, which is, however, caused by unspecified data rather than the studies in which teacher effect is reported. The difference between the studies in which teachers in control and experimental conditions are same or different is only 0.01 in terms of Hedge's g, which is trivial. Similar explanations are valid for the results of the analysis for researcher effect variable. Finally, research design and PBL mode explains almost no between-study variance, which eliminates the possibility of their being moderators for the study results.

5.3 Reliability and Validity

5.3.1 Coding Reliability

As explained in the methodology chapter, two different procedures were followed to establish coder reliability and inter-coder reliability. Firstly, a subsample of 10 studies was coded by the researcher again four weeks after coding of all primary studies had been completed and AR was calculated as explained in Section 2.5.3. An average AR of 0.979 was obtained with the range from 0.958 to 1.00. The details of the results are illustrated in Appendix F. Furthermore, another subsample of 14 studies was coded by other researchers to establish inter-coder reliability. Two studies were assigned for each of seven researchers, who accepted to be a parallel-coder for this meta-analysis and an average AR of 0.885 was obtained with a range from 0.792 to 0.938, which is also high enough to feel safe

about reliability issues. The details of calculations are presented in the table in Appendix G.

5.3.2 Internal Validity

Publication bias and quality of primary studies are two major concerns about the validity of meta-analysis results (Lipsey & Wilson, 2001; Rendina-Gobioff, 2006; Rothstein et al., 2005). Thus several methods were used to control the effect of these threats to validity of meta-analysis. Firstly, the effect of publication bias was examined for each sample of studies used in this meta-analysis. Table 5.1 summarizes the decisions about publication bias, which were based on different tests and indicators assessing and quantifying the magnitude of bias.

The results of publication bias analyses show that the magnitude of bias is either trivial or modest for any of the samples included in this meta-analysis, which results from high proportion of unpublished studies including doctoral dissertations and master theses within primary studies covered by meta-analysis. On the other hand, the results indicate another source of bias: small sample studies tend to have larger effect sizes than the studies with larger samples. However, the analysis also shows that it can easily be compensated owing to variety of primary studies included in the meta-analysis.

The concern of quality of primary studies was examined by using some indicators of study quality, which were used as variables in moderator analyses. Firstly, research design was coded to investigate whether the results of true experimental studies differed from quasi-experimental studies with or without randomly assigned clusters and the findings of moderator analysis showed that there was no significant difference between the effect sizes revealed from true and quasi experiments, and research design explained almost zero between-study variance. Secondly, teacher effect was analyzed as a moderator and similar results was found except for the proportion of explained variance, which seems to be larger than trivial one. However, as explained in previous section, this finding resulted from the biased effect of “unspecified” subgroup, which revealed the largest effect size values. Finally, researcher effect was analyzed as another variable that might

affect study quality and it was found that there was no nontrivial effect of this variable on the effect size values emerged from primary studies either.

PBL is used either a curriculum model or a teaching method in the literature. Thus, I also investigated whether the results of different PBL modes were comparable. All primary studies were coded as either "a curriculum model" or "a teaching method" and the moderator analysis showed that there was no considerable difference between either uses of PBL. Although the mean effect size for the studies in which PBL is used as a "teaching method" was slightly larger than the one for the studies in which PBL is used as a "curriculum model", the results were neither statistically significant nor substantial in terms of explained between study variance.

5.3.3 External Validity

In meta-analyses, it is aimed to reach all of the primary studies investigating research questions covered by meta-analyses; however, for many cases, if it is not for all, it is not realistic to believe that all the relevant studies are included in the analyses. Thus, it is widely accepted that the primary studies included in the meta-analysis are accepted as a sample of all relevant studies in the literature and calculated effect sizes are accepted as estimates of population(s), which is the reason for performing statistical tests for main effects in the meta-analyses.

Thus, similar to the primary studies, the external validity of meta-analyses is determined by how representative of the sample of studies included in the meta-analyses are for the population of all relevant studies. In this meta-analysis, a systematic literature search was performed to be able to as many studies as possible and after all primary studies were collected, some new searches were conducted to check whether there was any missing study as described in detail in the methodology chapter. As a result, a large sample of primary studies was used in this study, which shows almost no publication bias.

However, there may be some language bias in this meta-analysis since studies only written in English or in Turkish were accepted in the study. Nevertheless, there was no reason to believe that the bias is more than being modest

since English is the most widely accepted scientific language and the primary studies included in this meta-analysis come from 12 different countries.

Finally, it is essential to emphasize that primary reason to conduct a meta-analysis is making generalizations. Thus, many studies are synthesized in a meta-analysis to construct general conclusions having enormous external validity including both population and ecological validity. So, it should be noted that one of the advantages of meta-analysis similar to other research syntheses is to have substantial external validity comparing to primary studies.

5.4 Limitations of the Study

Firstly, although many of the primary studies included in the meta-analysis conducted in a pre-test post test research design, rarely the pre-test results could be concerned since the correlation between pre-test and post-test was not reported on most of the studies, which was essential to be able to compute an effect size based on both pre and post test results. Many of the studies provided different kinds of evidence to show that control and experimental groups were not different from each other in terms of concerned dependent variables, however, still the results should be evaluated cautiously since they are sensitive to initial differences between groups.

Secondly, it is not possible to construct cause-effect relationships by means of meta-analyses since they are based on data from the primary studies, which is impossible to be manipulated by meta-analysts. Thus, any conclusion referring to a cause-effect relationship should be evaluated as a claim, which needs further investigations by experimental studies (Borenstein et al., 2009).

Next, although judgment is highly necessary during coding procedure, all the information coded in this meta-analysis is based on what is provided by primary studies. Thus, the findings of the meta-analysis are vastly dependent upon the reporting quality of primary studies.

In addition, the number of primary studies included in the analysis performed to evaluate the overall effect of PBL on motivation in science is quite small, therefore, the results of this analysis has relatively limited external validity when compared to other variables in this meta-analysis.

Finally, moderator analyses conducted in this meta-analysis are based on an analog ANOVA, which does not give the chance of controlling some covariate variables. Thus, interaction effects may result in some fictitious relationships.

5.5 Conclusions

Based on the results of the main effect and moderator analyses of the data provided by 88 primary studies investigating the effectiveness of PBL on different outcomes, following conclusions can be drawn in this meta-analysis:

1. PBL is more effective on different outcomes comparing traditional teaching methods.
2. PBL has a large effect on science achievement when compared to traditional teaching methods.
3. PBL has a medium effect on students' attitudes toward science when compared to traditional teaching methods.
4. PBL has a medium to large effect on motivational constructs in science when compared to traditional teaching method.
5. PBL has a medium effect on different types of skills when compared to traditional teaching methods.
6. The effect size values estimated for the effectiveness of PBL on different outcomes are moderated by publication type: journal articles reveal larger effect sizes comparing to master theses, which have larger values when compared to doctoral dissertations.
7. The effect size values estimated for the effectiveness of PBL on different outcomes are also moderated by country: the studies conducted in Turkey indicate much higher effect size estimates than the studies performed in the USA and other countries.
8. PBL is more effective in Chemistry comparing to other science subject matters, Physics and Biology. There is also a noteworthy difference between Physics and Biology in favor of Physics.
9. The effectiveness of PBL on different outcomes differs by school level: it is much more effective in primary level and the effectiveness of PBL decreases with increasing school level.

10. The effectiveness of PBL on different outcomes differs by length of treatment, the mean effect size is larger for treatments of six to ten weeks duration comparing to shorter and longer duration.
11. Type of assessment instrument is a significant moderator variable for the effectiveness of PBL on different outcomes: the studies in which adapted or researcher-developed assessment instruments are used indicate much larger effect sizes when compared to the studies in which assessment is performed by means of pre-existing tests.
12. The results of studies are moderated by types of questions used in the assessment instruments as well: the studies in which assessment instruments include open-ended questions tend to reveal larger effect sizes comparing to the studies in which assessment is based on only objective type items.
13. The effectiveness of PBL on different outcomes seems to be affected by group size: the studies in which small groups of students are formed tend to have larger effect sizes than the studies in which students study in larger groups.
14. The results of the study are not significantly affected by any of teacher or researcher effect, research design and PBL mode, i.e. PBL is used as a curriculum model or a teaching method.

5.6 Implications of the Study

This meta-analysis draws some general and significant conclusions, which have practical implications as well. Some of the implications of the study for teachers, curriculum developers and policy makers can be summarized as follow:

1. PBL is suggested to be implemented in science classes since it has a large effect not only on science achievement but also on other essential constructs like attitudes toward science, motivation in science and different kinds of skills.
2. Since PBL seems to be more effective in primary levels and in chemistry subjects, it should be encouraged to be used especially in these contexts.

3. During implementation of PBL, group size should be kept in small numbers and types of questions in assessment instruments should not be limited to objective test items, including open-ended questions as well.
4. Curriculum developers and policy makers should be aware of the fact that PBL provides an effective way of teaching and learning science when it is implemented and assessed in an appropriate way.

In addition, based on the difficulties experienced by the researcher while performing this meta-analysis, some issues about reporting quality could be emphasized for researchers to make the primary studies serve better in cumulative knowledge. These are as follow:

1. Implementations for both control and experimental groups should be explained in detail.
2. Results of the study should be reported in detail including descriptive data not only for synthesis purposes but also to make the findings much more intuitive, for which presenting only results of the inferential statistics may not be sufficient.
3. The correlation between pre and post-assessment should be reported to be able to calculate an effect size eliminating the effect of possible differences between the groups at the beginning of the experiment.
4. Statistical tests based on inferential statistics should be interpreted as one of the tools that can be used to show how sensitive the findings are to generalizations rather than accepting them as the main purpose for conducting research.

5.7 Recommendations for Further Research

This meta-analysis study has revealed some interesting relationships and new topics to be studied by further research, which can be outlined as follow:

1. The effectiveness of PBL in different countries rather than Turkey and the USA could be compared and further research could be conducted to explain reasons for substantial difference between the effectiveness of PBL in Turkey and the USA.

2. Further research is necessary to explain the reasons why the effect of PBL differs in different disciplines in science.
3. Further research could be conducted to clarify the reasons why the effectiveness of PBL decreases with increasing school level.
4. Further research is necessary to examine the relationship between length of treatment and the effectiveness of PBL in detail to provide explanations about why the effect of PBL differs by length of treatment.
5. Further research could be conducted to provide empirically based explanations for the relationship between various properties of assessment and the effectiveness of PBL.
6. Further research is necessary to clarify the reasons for the interaction between group size and the effectiveness of PBL.
7. There is an obvious poverty of research investigating the effect of PBL on motivational constructs in science. Further experimental research could be conducted to provide the research synthesists with more empirical data to be able to make more valid generalizations.

REFERENCES

The primary studies included in this meta-analysis are indicated by an asterisk (*).

*Adalı, B. (2005). *İlköğretim 5. sınıf fen bilgisi dersinde "virüsler, bakteriler, mantarlar ve protistler" konularının öğreniminde örnek olaya dayalı öğrenme yöntemi kullanılmasının öğrencilerin akademik başarılarına ve fen bilgisi dersine yönelik tutumlarına etkisi*. Unpublished master thesis. Mustafa Kemal Üniversitesi. Hatay.

*Adiga, U., & Adiga, S. (2011). Case based learning in biochemistry. *İnternational Journal of Pharma and Bio Sciences, 2*(2), 332-336.

Akçay, B. (2009). Problem-based learning in science education. *Journal of Turkish Science Education, 6*(1), 26-36.

*Akın, S. (2008). *Anız yangınları ozon tabakasındaki incelme ve motorlu taşıtlardan kaynaklanan çevre sorunlarının probleme dayalı öğrenme yöntemi ile öğretimi*. Unpublished master thesis. Atatürk Üniversitesi. Erzurum.

*Akınoğlu, O., & Tandoğan, R. Ö. (2007). The effects of problem-based active learning in science education on students' academic achievement, attitude and concept learning. *Eurasia Journal of Mathematics, Science & Technology Education, 3*(1), 71-81.

*Alagöz, B. (2009). *Sosyal Bilgiler öğretmen adaylarında çevre bilincinin geliştirilmesinde probleme dayalı öğrenme yönteminin etkisi*. Unpublished doctoral dissertation. Gazi Üniversitesi. Ankara.

Albanese, M. A. (2000). Problem-based learning: Why curricula are likely to show little effect on knowledge and clinical skills. *Medical Education, 34*(9), 729-738.

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine, 68*(1), 52-81.

Anderson, D. R., Kahl, S. R., Glass, G. V., & Smith, M. L. (1983). Science education: A meta-analysis of major questions. *Journal of Research in Science Teaching, 20*(5), 379-385.

*Anderson, J. C. (2007). *Effect of problem-based learning on knowledge acquisition, knowledge retention, and critical thinking ability of agriculture students in urban schools*. Unpublished doctoral dissertation, University of Missouri, Columbia.

*Araz, G. (2007). *The effect of problem-based learning on the elementary school students' achievement in genetics*. Unpublished master thesis, METU, Ankara.

*Atan, H., Sulaiman, F., & Idrus, R. M. (2005). The effectiveness of problem-based learning in the web-based environment for the delivery of an undergraduate physics course. *International Education Journal, 6*(4), 430-437.

Ateş, Ö. (2009). *An analysis of the problem-based instruction in engineering education*. Unpublished doctoral dissertation, Middle East Technical University, Ankara.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1978). *Educational psychology: A cognitive view*. New York: Holt Rinehart and Winston.

Bangert-Drowns, R. L., & Rudner, L. M. (1991). *Meta-analysis in educational research*: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Barrows, H. . (1996). Problem-based learning in medicine and beyond: A brief overview. *New Directions for Teaching and Learning, 68*, 3-12.

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.

*Bayrak, B., & Bayram, H. (2011). Effects of Problem-based learning in a web environment on conceptual understanding: The subject of acids and bases. *International Online Journal, 3*(3), 831-848.

*Bayrak, R. (2007). *Probleme dayalı öğrenme yaklaşımı ile katılar konusunun öğretimi*. Unpublished doctoral dissertation, Atatürk Üniversitesi, Erzurum.

Bayraktar, S. (2000). *A meta analysis study on the effectiveness of computer assisted instruction in science education*. Unpublished doctoral dissertation, Ohio University, Ohio.

Bayraktar, S. (2002). A meta-analysis of the effectiveness of computer assistant instruction in science education. *Journal of Research on Technology in Education, 34*(2), 173-188.

*Bayram, A. . (2010). *Probleme dayalı öğrenme yönteminin ilköğretim 5. sınıf öğrencilerinin fen ve teknoloji dersi "ısı ve sıcaklık" konusunda sahip oldukları kavram yanılgılarını gidermede etkisi*. Unpublished master thesis, Selçuk Üniversitesi, Konya.

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.

*Benli, E. . (2010). *Probleme dayalı öğrenmenin fen bilgisi öğretmen adaylarının akademik başarılarına, bilgilerin kalıcılığına ve fene karşı tutumlarına etkilerinin araştırılması*. Unpublished master thesis, Gazi Üniversitesi, Ankara.

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.

Berkson, L. (1993). Problem-based learning: Have the expectations been met? *Academic Medicine, 68*(10).

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.

*Bilgin, İ., Şenocak, E., & Sözbilir, M. (2009). The effects of problem-based learning instruction on university students'performance of conceptual and

quantitative problems in gas concepts. *Eurasia Journal of Mathematics, Science and Technology Education, 5*(2), 153-164.

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

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res., 4*, 4-16.

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.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21-32.

Burns, P. K., & Bozeman, W. C. (1981). Computer assisted instruction and mathematics achievement: Is there a relationship? *Educational Technology, 21*(10), 32-39.

*Burris, S. (2005). *Effect of problem-based learning on critical thinking ability and content knowledge of secondary agriculture students*. Unpublished doctoral dissertation, University of Missouri, Columbia.

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.

*Çam, A. (2009). *Effectiveness of case-based learning instruction on students' understanding of solubility equilirium concepts*. Unpublished doctoral dissertation, METU, Ankara.

Campbell, L. O. (2009). *A meta-analytical review of Novak's concept mapping*. Unpublished doctoral dissertation, Regent University, Virginia.

Capon, N., & Kuhn, D. (2004). What's so good about problem-based learning? *Cognition and Instruction, 22*(1), 61-79.

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

*Carll-Williamson, M. P. (2003). *The effect of problem-based learning in high school students enrolled in biology for the technologies as measured by concept of self and attitude toward science*. Unpublished doctoral dissertation, University of South Carolina, Columbia.

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

*Carrio, M., Larramona, P., Banos, J. E., & Perez, J. (2011). The effectiveness of the hybrid problem-based learning approach in the teaching of biology: a comparison with lecture-based learning. *Journal of Biological Education, 45*(4), 229-235.

*Çelik, E. (2010). *Fen eğitiminde probleme dayalı öğrenme yaklaşımının öğrencilerin akademik başarısına, tutumuna, akademik risk alma düzeyine ve kalıcılığa etkisi*. Unpublished master thesis, Gazi Üniversitesi, Ankara.

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.

Chin, C., & Chia, L. G. (2005). Problem-based learning: Using ill-structured problems in biology project work. *Science Education, 90*(1), 44-67.

Christmann, E. (1997). Progressive comparison of the effects of computer-assisted instruction on the academic achievement of secondary students. *Journal of Research on Computing in Education, 29*(4), 325-337.

*Çınar, D. (2007). *İlköğretim fen eğitiminde probleme dayalı öğrenme yaklaşımının üst düzey düşünme becerilerine ve akademik risk alma düzeyine etkisi*. Unpublished master thesis, Selçuk Üniversitesi, Konya.

Clark, R. E. (1985). Evidence for confounding in computer-based instruction studies: Analyzing the meta-analyses. *Educational Technology Research and Development, 33*(4), 249-262.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah: Lawrence Erlbaum Associates, Publishers.

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

Colliver, J. A. (2000). Effectiveness of problem-based learning curricula: research and theory. *Academic Medicine, 75*(3), 259-266.

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., Hedges, L. V., & Valentine, J. C. . (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). 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.

Danielson, C. (2008). *Teaching methods*. Boston: Prentice Hall Higher Education.

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

*De Simone, C. (2008). Problem-based learning: a framework for prospective teachers' pedagogical problem solving. *Teacher Development, 12*(3), 179-191.

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).

*Demirel, M., & Turan, B. A. (2010). Probleme dayalı öğrenmenin başarıya, tutuma, bilişötesi farkındalık ve güdü düzeyine etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 38*, 55-66.

Dewey, J. (1938). *Experience and education*. New York: Macmillan.

*Dieber, J. M. (1994). *A comparison between traditional and problem based learning medical students as self-directed continuing learners*. Unpublished doctoral dissertation, Northern Illınois University, Illinois.

Dieckmann, N. F., Malle, B. F., & Bodner, T. E. (2009). An empirical assessment of meta-analytic practice. *Review of General Psychology, 13*(2), 101-115.

*Diggs, L. L. (1997). *Student attitude toward and achievement in science in a problem-based learning educational experience*. Unpublished doctoral dissertation, University of Missouri, Columbia.

Distlehorst, L. H. (1994). Responses to "Problem-based learning: Have the expectations been met?". *Academic Medicine, 69*(6), 471-472.

*Dobbs, V. (2008). *Comparing student achievement in the problem-based learning classroom and traditional teaching methods classroom*. Unpublished doctoral dissertation, Walden University, Minnesota.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction, 13*(5), 533-568.

*Downing, K., Ning, F., & Shin, K. (2011). Impact of problem-based learning on student experience and metacognitive development. *Multicultural Education and Technology Journal, 5*(1), 55-69.

*Drake, K. N., & Long, D. (2009). Rebecca's in the dark: A comparative study of problem-based learning and direct instruction/experiential learning in two 4th grade classrooms. *Journal of Elementary Science Education, 21*(1), 1-16.

Druva, C. A., & Anderson, R. D. (1983). Science teacher characteristics by teacher behavior and by student outcome: A meta-analysis of research. *Journal of Research in Science Teaching, 20*(5), 467-479.

Duffy, T. M., & Cunningham, D. J. (1996). Constructivism: Implications for the design and delivery of instruction. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology*. New York: Macmillan.

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.

*Erdem, E. (2006). *Probleme dayalı öğrenmenin öğrenme ürünlerine, problem çözme becerisine ve öz-yeterlik algı düzeyine etkisi*. Unpublished doctoral dissertation, Hacettepe Üniversitesi, Ankara.

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. (2003a). Can meta-analysis be trusted? *The Psychologist, 16*(12), 642-645.

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

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665-694.

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.

Fleming, M. L., & Malone, M. R. (1983). The relationship of student characteristics and student performance in science as viewed by meta-analysis research. *Journal of Research in Science Teaching, 20*(5), 481-495.

Flinn, C. M., & Gravatt, B. (1995). The efficacy of computer assisted instruction (CAI): A meta-analysis. *Journal of Educational Computing Research, 12*(3), 219-241.

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). Boston: McGraw-Hill Higher Education.

*Gabr, H., & Mohamed, N. (2011). Effect of problem-based learning on undergraduate nursing students enrolled in nursing administration course. *International Journal of Academic Research, 3*(1), 154-162.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research, 75*(1), 27-61.

Gijselaers, W. H. (1996). Connecting problem-based practices with educational theory. *New Directions For Teaching and Learning, 68*, 13-21.

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

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5*(1), 351-379.

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. (2000). Meta-analysis at 25. Retrieved March 12, 2011, from http://www.gvglass.info/papers/meta25.html

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.

*Günhan, B. C., & Başer, N. (2009). Probleme dayalı öğrenmenin öğrencilerin eleştirel düşünme becerilerine etkisi. *Türk Eğitim Bilimleri Dergisi, 7*(2), 451-482.

*Gürlen, E. (2011). Probleme dayalı öğrenmenin öğrenme ürünlerine, problem çözme becerisine, öz-yeterlik algı düzeyine etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 40*, 221-232.

Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting conceptual change in science: a comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly*, *28*(2), 116-159.

Haas, M. (2005). Teaching methods for secondary algebra: A meta-analysis of findings. *NASSP Bulletin*, *89*(642), 24.

Hartley, K. (2001). Learning strategies and hypermedia instruction. *Journal of Educational Multimedia and Hypermedia*, *10*(3), 285-305.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. NY: Taylor & Francis.

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.

*Hesterberg, L. J. (2005). *Evaluation of a problem-based learning practice course: Do self-efficacy, critical thinking, and assessment skills improve?* Unpublished doctoral dissertation, University of Kentucky, Lexington.

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.

Hmelo-Silver, C. E. (2004). Problem-based learning: what and how do students learn? *Educational Psychology Review, 16*(3), 235-266.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99-107.

Horton, P. B. McConney A. A. Gallo M. Woods A. L. Senn G. J., & Hamelin, D. (1993). An investigation of the effectiveness of concept mapping as an instructional tool. *Science Education, 77*(1), 95-111.

*Horzum, M. B., & Alper, A. (2006). The effect of case based learning model, cognitive style and gender to the student achievement in science courses. *Journal of Faculty of Eductaional Sciences, 39*(2), 151-175.

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.

Igel, C. C. (2010). *The effect of cooperative learning instruction on K-12 student learning: A meta-analysis of quantitative studies from 1998 to 2009*. Unpublished doctoral dissertation, University of Virginia, Virginia.

*İnel, D. (2009). *Fen ve teknoloji dersinde probleme dayalı öğrenme yöntemi kullanımının öğrencilerin kavramları yapılandırma düzeyleri, akademik başarıları ve sorgulayıcı öğrenme becerileri algıları üzerindeki etkileri*. Unpublished master thesis, Dokuz Eylül Ünüversitesi, İzmir.

*Jandric, G. H., Obadovic, D. Z., Stojanovic, M., & Rancic, I. (2011). Impacts of the implementation of the problem-based learning in teaching physics in primary schools. *The New Educational Review*, 194-204.

Jonhson, D. W., Johnson, R., & Stanne, M. B. (2000). Cooperative Learning Methods: A Meta-Analysis, [on-line]. Retrieved June 1, 2011, from The Cooperative Learning Center at The University of Minnesota website: http://www.tablelearning.com/uploads/File/EXHIBIT-B.pdf

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.

*Kaddouro, M. A. (2011). Critical thinking skills of nursing students in lecture-based teaching and case-based learning. *International Journal for the Scholarship of Teaching and Learning*, *5*(2).

Kalaian, H. A., Mulllan, P. B., & Kasim, R. M. (1999). What can studies of problem-based learning tell us? Synthesizing and modeling PBL effects on national board of medical examination performance: Hierarchical linear modeling meta-analytic approach. *Advances in Health Sciences Education, 4*(3), 209-221.

*Kanlı, E. (2008). *Fen ve teknoloji öğretiminde probleme dayalı öğrenmenin üstün ve normal zihin düzeyindeki öğrencilerin erişi, yaratıcı düşünme ve motivasyon düzeylerine etkisi*. Unpublished master thesis, İstanbul Üniversitesi, İstanbul.

*Kar, T. (2010). *Lineer cebirde probleme dayalı öğrenme yönteminin öğrencilerin akademik başarıları, problem çözme becerileri ve yaratıcılıkları üzerine etkisi*. Unpublished master thesis, Atatürk Üniversitesi, Erzurum.

*Karaöz, M. P. (2008). *İlköğretim fen ve teknoloji dersi "kuvvet ve hareket" ünitesinin probleme dayalı öğrenme yaklaşımıyla öğretiminin öğrencilerin bilimsel süreç becerileri, başarıları ve tutumları üzerine etkisi*. Unpublished master thesis, Muğla Üniversitesi, Muğla.

Kilpatric, W. H. (1918). The project method. *Teach. Coll. Rec., 19,* 319-335.

Kilpatric, W. H. (1921). Dangers and difficulties of the project method and how to overcome them: Introductory statement: Definitions of terms. *Teach. Coll. Rec., 22*, 282-288.

*Kıray, S. A., & İlik, A. (2011). Polya'nın problem çözme yönteminin fen bilgisi öğretiminde kullnaılmasına yönelik bir çalışma: kanıt temelli uygulamaya doğru. *Selçuk Ahmet Keleşoğlu Eğitim Fakültesi Dergisi, 31*, 183-202.

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.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.

Koballa, T. R., & Glynn, S. M. (2007). Attitudinal and motivational constructs in science learning. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research in science education* (pp. 75-102). Mahwah, New Jersey: Lawrance Erlbaum Associates, Publishers.

*Koçakoğlu, M. (2008). *Probleme dayalı öğrenme ve motivasyon stillerinin öğrencilerin biyoloji dersine karşı tutum ve akademik başarılarına etkisi*. Unpublished doctoral dissertation, Gazi Üniversitesi, Ankara.

*Könings, K. D., Wiers, R. W., van de Wiel, M. W. J., & Schmidt, H. G. (2005). Problem-based learning as a valuable educational method for physically

disabled teenagers. The discrepancy between theory and practice. *Journal of Development and Physical Disabilities, 17*(2), 107-117.

*Koray, Ö., Presley, A., Köksal, M. S., & Özdemir, M. (2008). Enhancing problem-solving skills of pre-service elementary school teachers through problem-based learning. *Asia-Pacific Forum on Science Learning and Teaching, 9*(2).

Kuhn, D. (2007). Is direct instruction an answer to the right question? *Educational Psychologist, 42*(2), 109-113.

Kulik, C. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior, 7*, 75-94.

Kulik, J. A. (1983). Synthesis of research on computer-based instruction. *Educational Leadership, 41*(1), 19-21.

Kulik, J. A. (1985). Effectiveness of computer-based education in elementary schools. *Computers in Human Behavior, 1*(1), 59-74.

Kulik, J. A. , Bangert, R. L., & Williams, G. W. (1983). Effects of computer-based teaching on secondary school students. *Journal of Educational Psychology, 75*(1), 19-26.

Kulik, J. A. , Kulik, C. C., & Cohen, P. A. (1980). Effectiveness of computer-based college teaching: A meta-analysis of findings. *Review of Educational Research, 50*(4), 525.

*Kuşdemir, M. (2010). *Probleme dayalı öğrenmenin öğrencilerin başarı, tutum ve motivasyonlarına etkisinin incelenmesi*. Unpublished master thesis, Mustafa Kemal Üniversitesi, Hatay.

Last, J. M. (2001). *A dictionary of epidemiology*. Oxford: Oxford University Press.

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.

*LeJeune, N. F. (2002). *Problem-based learning instruction versus traditional instruction on self-directed learning, motivation, and grades of undergraduate computer science students* Unpublished doctoral dissertation, University of Colorado, Denver.

Lemeshow, A. R., Blum, R. E., Berlin, J. A., Stoto, M. A., & Colditz, G. A. (2005). Searching one or two databases was insufficient for meta-analysis of observational studies. *Journal of Clinical Epidemiology, 58*(9), 867-873.

*Lesperance, M. M. (2008). *The effects of problem-based learning on students' critical thinking skills*. Unpublished doctoral dissertation, The University of North Carolina, Greensboro.

*Lewis, D. L. (2006). *Comparison of problem-based learning and traditional learning in biology laboratory for biology and science mejors in community colleges*. Unpublished doctoral dissertation, Arkansas State University, Arkansas.

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

Liao, Y. C. (1999). Effects of hypermedia on students' achievement: A meta-analysis. *Journal of Educational Multimedia and Hypermedia, 8*(3), 255-277.

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.

Lott, G. W. (1983). The effect of inquiry teaching and advance organizers upon student outcomes in science education. *Journal of Research in Science Teaching, 20*(5), 437-451.

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

*Lyons, E. B. (2006). *Examining the effects of problem-based learning on the critical thinking skills of associate degree nursing students in a southeastern community college*. Unpublished doctoral dissertation, Mississippi State University, Mississippi.

Marcucci, R. G. (1980). Meta-analysis of research on methods of teaching mathematical problem solving. *Dissertation Abstracts International, 41*(6).

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.

Marzano, R. J. (1998). *A Theory-based meta-analysis of research on instruction*. Colorado: Mid-continent Regional Educational Laboratory.

Marzano, R. J., Pickering, D., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. VA: Association for Supervision & Curriculum Development.

*Mathew, E. (2008). *Learning physics: A comparative analysis between instructional design methods*. Unpublished doctoral dissertation, Capella University, Minneapolis.

Maudsley, G. (1999). Do we all mean the same thing by" problem-based learning"? A review of the concepts and a formulation of the ground rules. *Academic Medicine, 74*(2), 178-185.

*McGee, M. R. (2003). *A comparison of traditional learning and problem-based learning in pharmacology education for athletic training students*. Unpublished doctoral dissertation, The University of North Carolina, Greensboro.

Minner, D. D., Levy, A. J., & Century, J. (2009). Inquiry-based science instruction-What is it and does it matter? Results fro a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474-496.

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.

*Mungin, R. E. (2012). *Problem-based learning versus traditional science instruction: achievement and interest in science of middle grades minority females*. Unpublished doctoral dissertation, Capella University, Minneapolis.

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

*Needham, M. E. (2010). *Comparison of standardized test scores from traditional classrooms and those using problem-based learning* Unpublished doctoral dissertation, University of Missouri, Kansas City.

Nesbit, J. C., & Olusola, O. A. (2006). Cooperative versus competitive efforts and problem solving *Review of Eductaional Research, 76*(3), 413-448.

Niemiec, R. P., & Walberg, H. J. (1985). Computers and achievement in the elementary schools. *Journal of Educational Computing Research, 1*(4), 435-440.

Norman, G. R., & Schmidt, H. G. (2000). Effectiveness of problem-based learning curricula: Theory, practice and paper darts. *Medical Education, 34*(9), 721-728.

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

*Nowak, J. A. (2002). *The implications and outcomes of using problem-based learning to teach middle school science*. Unpublished doctoral dissertation, Indiana University, Indiana.

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.

*Olgun, O. S. (2008). Teaching grade 5 life science with a case study approach. *Journal of Elementary Education, 20*(1), 29-44.

Onuoha, C. O. (2007). *Meta-analysis of the effectiveness of computer-based laboratory versus traditional hands-on laboratory in college and pre-college science instructions*. Unpublished doctoral dissertation, Capella University, Minneapolis.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*(2), 157-159.

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.

*Özeken, Ö. F., & Yıldırım, A. (2011). Asit-baz konusunun öğretiminde probleme dayalı öğrenme yönteminin fen bilgisi öğretmen adaylarının akademik başarıları üzerine etkisi. *Pegem Eğitim ve Öğretim Dergisi, 1*(1), 33-38.

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.

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.

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

Pillemer, D.B., & Light, R.J. (1980). Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review, 50*(2), 176-195.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*, 223-232.

Qin, Z., Johnson, D. W., & Johnson, R. T. (1995). Cooperative versus competitive efforts and problem solving *Review of Educational Research, 65*(2), 129-143.

*Rajab, A. M. (2007). *The effects of problem-based learning on the self-efficacy and attitudes of beginning biology majors* Unpublished doctoral dissertation, University of California, Irvine and Los Angeles.

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.

*Sağır, Ş. U., Çelik, A. Y., & Armağan, F. Ö. (2009). Metalik aktiflik konusunun öğretimine probleme dayalı öğrenme yaklaşımının etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 36*, 283-293.

*Şahbaz, Ö. (2010). *İlköğretim 5. sınıf fen ve teknoloji dersinde kullanılan farklı yöntemlerin öğrencilerin bilimsel süreç becerileri, problem çözme becerileri, akademik başarıları ve hatırda tutma üzerindeki etkileri*. Unpublished doctoral dissertation, Dokuz Eylül Üniversitesi, İzmir.

*Şahin, A. (2011). *Genel fizik laboratuar dersinde basit elektrik devreleri konusunun öğretilmesinde probleme dayalı öğrenme yaklaşımının öğrencilerin akademik başarılarına etkisinin incelenmesi*. Unpublished master thesis, Atatürk Üniversitesi, Erzurum.

*Şahin, M. C. (2005). *İnternet tabanlı uzaktan eğitimin etkililiği: Bir meta-analiz çalışması*. Unpublished master thesis, Çukurova Üniversitesi, Adana.

*Şalgam, E. (2009). *Fizik eğitiminde probleme dayalı öğrenme yönteminin öğrencilerin akademik başarılarına ve tutumlarına etkisi*. Unpublished master thesis, Dokuz Eylül Üniversitesi, İzmir.

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. *Meta-analysis in 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.

*Sanderson, H. L. (2008). *Comparison of problem-based learning and traditional lecture instruction on critical thinking, knowledge, and application of strength and conditioning*. Unpublished doctoral dissertation, The University of North Carolina, Greensboro.

*Saral, S. (2008). *The effect of case based learning on tenth grade students' understanding of human reproductive system and their perceived motivation*. Unpublished master thesis, METU, Ankara.

Savery, J. R. (2006). Overview of problem-based learning: Definitions and distinctions. *The Interdisciplinary Journal of Problem-based Learning, 1*(1), 9-20.

Savery, J. R., & Duffy, T. M. (2001). Problem based learning: An instructional model and its constructivist framework *CRLT Technical Report No. 16-01* Indiana: Indiana University.

Savin-Baden, M., & Major, C. H. (2004). *Foundation of problem-based learning*. Berkshire: Open University Press.

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, In-Sue, & 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.

Schmidt, H. , Loyens, S. M. M., Van Gog, T., & Paas, F. (2007). Problem-based learning is compatible with human cognitive architecture: Commentary on Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 91-97.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T. Y., & Lee, Y. H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44*(10), 25.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Ashland: Hogrefe & Huber Publishers.

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

*Scott, W. (2005). *Investigating traditional instruction and problem-based learning at the elementary level* Unpublished doctoral dissertation, Mississippi State University, Mississippi.

*Selçuk, G. S., Karabey, B., & Çalışkan, S. (2011). Probleme dayalı öğrenmenin matematik öğretmen adaylarının ölçme ve vektörler konularındaki başarıları üzerindeki etkisi. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8*(15), 313-322.

*Semerci, N. (2006). The effect of problem based learning on the critical thinking of students in the intellectual and ethical development unit. *Social Behaviour and Personality, 34*(9), 1127-1136.

*Şendağ, S. (2008). *Çevrimiçi probleme dayalı öğrenmenin öğretmen adaylarının eleştirel düşünme becerilerine ve akademik başarılarına etkisi*. Unpublished doctoral dissertation, Anadolu Üniversitesi, Eskişehir.

*Şenocak, E., Taşkesenligil, Y., & Sözbilir, M. (2007). A study on teaching gases to prospective primary science teachers thorugh problem-based learning. *Research in Science Education, 37,* 279-290.

*Serin, G. (2009). *The effect of problem based learning instruction on 7th grade students' science achievement, attitude toward science and scientific process*

skills. Unpublished doctoral dissertation, Middle East Technical University, Ankara.

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.

*Shepherd, N. G. (1998). *The probe method: A problem-based learning model's affect on critical thinking skills of fourth and fifth grade social stuides students*. Unpublished doctoral dissertation, North Carolina State University, Raleigh.

Shymansky, J. A., Kyle Jr, W. C., & Alport, J. M. (1983). The effects of new science curricula on student performance. *Journal of Research in Science Teaching, 20*(5), 387-404.

Smaldino, S. E., Russell, J. D., Heinich, R., & Molenda, M. (2005). *Instructional technology and media for learning*. New Jersey: Pearson Education Inc.

Smith, D. . (1996). *A meta-analysis of student outcomes attributable to the teaching of science as inquiry as compared to traditional methodology*. Unpublished doctoral dissertation, Temple University, Philadelphia.

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.

Smith, R. A. (2003). *Problem-based versus lecture-based medical education: A meta-analysis of cognitive and noncognitive outcomes*. Unpublished doctoral dissertation, University of Florida, Florida.

Smits, P. B. A., Verbeek, J., & De Buisonje, C. D. (2002). Problem based learning in continuing medical education: A review of controlled evaluation studies. *British Medical Journal, 324*(7330), 153.

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.

Strobel, J., & Van Barneveld, A. (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. *Interdisciplinary Journal of Problem-based Learning, 3*(1), 4.

*Sungur, S., & Tekkaya, C. (2006). Effects of problem-based learning and traditional instruction on self-regulated learning. *The Journal of Educational Research, 99*(5), 307-320.

*Sungur, S., Tekkaya, C., & Geban, O. (2006). Improving achievement through problem-based learning. *Journal of Biological Education, 40*(4), 155.

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.

Sweitzer, G. L., & Anderson, R. D. (1983). A meta-analysis of research on science teacher education practices associated with inquiry strategy. *Journal of Research in Science Teaching, 20*(5), 453-466.

Sweller, J., Kirschner, P. A., & Richard, E. C. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist, 42*(2), 115-121.

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.

*Tarhan, L., & Acar, B. (2007). Problem-based learning in an eleventh grade chemistry class: 'factors affecting cell potential'. *Research in Science and Technological Education, 25*(3), 351-369.

*Tarhan, L., Kayalı, H. A., Ürek, R. O., & Acar, B. (2008). Problem-based learning in 9th grade chemistry class: 'Intermolecular forces'. *Research in Science Education, 38,* 285-300.

Taşkesenligil, Y. (2008). Probleme dayalı öğrenme teorik temelleri. *Milli Eğitim Dergisi, 177*, 50-64.

*Taşoğlu, A. K. (2009). *Fizik eğitiminde probleme dayalı öğrenmenin öğrencilerin başarılarına,bilimsel süreç becerilerine ve problem çözme tutumlarına etkisi*. Unpublished master thesis, Dokuz Eylül Üniversitesi, İzmir.

*Tavukçu, K. (2006). *Fen bilgisi dersinde probleme dayalı öğrenmenin öğrenme ürünlerine etkisi*. Unpublished master thesis, Zonguldak Karaelmas Üniversitesi, Zonguldak.

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.

Tinoca, L. F. (2004). *From professional development for science teachers to student learning in science*. Unpublished doctoral dissertation, University of Texas, Austin.

*Tiwari, A., Lai, P., So, M, & Yuen, K. (2006). A comparison of the effects of problem-based learning and lecturing on the development of students' critical thinking *Medical Education, 40*, 547-554.

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

*Tosun, C. (2010). *Probleme dayalı öğrenme yönteminin çözeltiler ve fiziksel özellikleri konusunun anlaşılmasına etkisi*. Unpublished doctoral dissertation, Atatürk Üniversitesi, Erzurum.

Treagust, D. F. (2007). General instructional methods and strategies. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of Reserach on Science Education*. New Jersey: Lawrance Erlbaum Associates.

*Tüysüz, C., Tatar, E., & Kuşdemir, M. (2010). Probleme dayalı öğrenmenin kimya dersinde öğrencilerin başarı ve tutumlarına etkisinin incelenmesi. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 7*(13), 48-55.

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.

Uden, L., & Beaumont, C. (2006). *Technology and problem-based learning.* Hershey: Information Science Publishing.

*Ülger, K. (2011). *Görsel sanatlar eğitiminde probleme dayalı öğrenme modelinin yaratıcı düşünmeye etkisi*. Unpublished doctoral dissertation, Gazi Üniversitesi, Ankara.

*Usoh, I. I. (2003). *An investigation into effectiveness of problem-based learning in an engineering technology program at Nashville State Technical Community College* Unpublished doctoral dissertation, Tennesse State University, Nashville.

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.

*van Loggerenberg-Hattingh, A. (2003). Examining learning achievement and experiences of science learners in a problem based learning environment. *South African Journal of Education, 23*(1), 52-57.

Vernon, D. T., & Blake, R. L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine, 68*(7), 550-563.

Vernon, D. T., & Blake, R. L. (1994). Responses to "Problem-based learning: Have the expectations been met?". *Academic Medicine, 69*(6), 472-473.

Viechtbauer, W. (2007). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie/Journal of Psychology, 215*(2), 104-121.

von Glasersfeld, E. (1989). Cognition, construction of knowledge, and teaching. *Synthese, 80*(1), 121-140.

Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Walker, A., & Leary, H. (2009). A problem based learning meta analysis: Differences across problem types, implementation types, disciplines, and assessment levels. *Interdisciplinary Journal of Problem-based Learning, 3*(1), 6.

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.

Willett, J. B., Yamashita, J. J. M., & Anderson, R. D. (1983). A meta-analysis of instructional systems applied in science teaching. *Journal of Research in Science Teaching, 20*(5), 405-417.

Wise, K. C. (1996). Strategies for teaching science: What Works? *Clearing House, 69*(6), 337-338.

Wise, K. C., & Okey, J. R. (1983). A meta-analysis of the effects of various science teaching strategies on achievement. *Journal of Research in Science Teaching, 20*(5), 419-435.

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

Wood, D. F. (2003). Problem based learning. *British Medical Journal, 326*(7384), 328.

*Yalçınkaya, E. (2010). *Effect of case based learning on 10th grade students' understanding of gas concepts, their atttude and motivation*. Unpublished master thesis, METU, Ankara.

Yalçınkaya, E. (2012). Is case-based learning an effective teaching strategy to challange students' alternative conception regarding chemical kinetics? *Research in Science and Technological Education, 30*(2), 151-172.

*Yaman, S. (2005). Fen bilgisi öğretiminde probleme dayalı öğrenmenin mantıksal düşünme becerisinin gelişimine etkisi *Türk Fen Eğitimi Dergisi, 2*(1), 56-70.

*Yaman, S., & Yalçın, N. (2005a). Fen bilgisi öğretiminde probleme dayalı öğrenme yaklaşımının yaratıcı düşünme becerisine etkisi. *İlköğretim-Online, 4*(1), 42-52.

*Yaman, S., & Yalçın, N. (2005b). Fen eğitiminde probleme dayalı öğrenme yaklaşımının problem çözme ve öz-yeterlik inanç düzeylerinin gelişimine etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 29*, 229-236.

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

*Yuan, H., Kunaviktikul, W., Klunklin, A., & Williams, B. A. (2008). Improvement of nursing students' critical thinking skills thorugh problem-based learning in the People's Republic of China: A quasi-experimental study. *Nursing and Health Sciences, 10*, 70-76.

*Yurd, M. (2007). *İlköğretim 5. sınıf fen ve teknoloji dersinde probleme dayalı öğrenme yöntemi ile bil-iste-öğren stratejisi kullanılarak geliştirilen bil-iste-örnekle-öğren stratejisinin öğrencilerin kavram yanılgılarının giderilmesine v derse karşı tutumlarına etkisi*. Unpublished master thesis, Mustafa Kemal Üniversitesi, Hatay.

*Yurd, M, & Olgun, O. S. (2008). Probleme dayalı öğrenme ve bil-iste-öğren stratejisinin kavram yanılgılarının giderilmesine etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 35*, 386-396.

APPENDIX A

FIRST DRAFT OF CODING SHEET

Coder Name:

Study No:

Title of the Study:

Author(s) (Surname, Name):

1. **Publication Date:**
2. **Publication Type:**
   a. Journal Article b. Master Thesis c. Doctoral Dissertation
3. **Research Design:**
   a. True Experimental
   b. Quasi-experimental (with random selection)
   c. Quasi-experimental (without random selection)
   d. Other (Specify):
4. **Country:**
   a. Turkey b. United States c. Other (Specify):
5. **Demography:**
   a. Urban b. Suburban c. Rural d. Unspecified
6. **Type of Teaching Strategy:**
7. **Subject Area:**
   a. Physics b. Chemistry c. Biology
   d. General Science e. Other (Specify):

8. **School Level:**

   a. Primary b. Secondary c. College

   d. University e. Other (Specify):

9. **Private or Public (State) School:**

   a. Private School b. State School c. Mixed

10. **School Type:**

    a. Elementary School b. Anatolian High School

    c. Vocational School d. Regular High School d. Other (Specify):

11. **Grade Level:**

12. **Characteristic of Population:**

    a. Gifted Students b. Students with Learning Difficulties

    c. Normal

13. **(If your answer is "c" for the previous question) Ability Level of Student:**

    a. Below average b. Average c. Above average d. Mixed

    e. Unspecified

14. **Gender:**

    a. All Female b. All Male c. Mixed d. Unspecified

15. **Length of Treatment (in months, weeks, days, or hours):**

16. **Sample Size:**

17. **Type of the Assessment Instrument 1:**

    a. Standardized Test b. Researcher Developed Test c. Both

    d. Unspecified

    Type of the Assessment Instrument 2:

    a. Standardized Test b. Researcher Developed Test c. Both

    d. Unspecified

    Type of the Assessment Instrument 3:

    a. Standardized Test b. Researcher Developed Test c. Both

    d. Unspecified

18. Teacher Effect:

a. Same teacher for both control and experimental group

b. Different teachers

c. Unspecified

19. Researcher Effect:

a. Researcher is one of the teachers

b. Researcher is not any of the teachers

c. Researcher is the only teacher

20. Teacher Training Period (in months, weeks, days, or hours):

21. Implementation Time:

22. Statistical Analysis Technique Used:

23. Type of Outcome:

a. Achievement

b. Motivation in Science (Specify if necessary):

c. Attitudes towards Science

d. Other (Specify):

e. Combination (Specify the components):

24. Science Achievement Outcome:

a. Science Process Skills

b. Concept Learning

c. Problem Solving

d. Other (Specify):

e. Combination (Specify the components):

f. Unspecified

25. Level of Control over Threats to Internal Validity

	Unacceptable (0)	Poor (1)	Average (2)	Good (3)	Very Good (4)	Not enough information
Subject Characteristics						
Loss of Subjects (Mortality)						
Location						
Instrument Decay						

Data Collector Characteristics
Data Collector Bias
Testing
Extraneous Events (History)
Maturation
Attitude of Subjects
Regression
Implementation

TOTAL SCORE:

26. Level of Treatment Fidelity

a. Unacceptable b. Poor c. Average d. Good

d. Very Good e. Not Enough Information

27. Level of Treatment Verification

a. Unacceptable b. Poor c. Average d. Good

d. Very Good e. Not Enough Information

28. Reliability Coefficients of the Instruments:

a. Instrument 1:

b. Instrument 2:

c. Instrument 3:

29. Average Difficulty level of the instruments:

a. Instrument 1:

b. Instrument 2:

c. Instrument 3:

30. Average Distinctiveness of the instruments:

a. Instrument 1:

b. Instrument 2:

c. Instrument 3:

31. Number of items in the instruments:

a. Instrument 1:

b. Instrument 2:

c. Instrument 3:

32. Time Limit to Complete the Instruments:

a. Instrument 1:

b. Instrument 2:

c. Instrument 3:

33. Study Results:

a. Effect Size(s):

b. p(s):

c. t(s):

d. F(s):

e. Omega Square(s):

TEST 1

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 2

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 3

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 4

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 5

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

Reference:

APPENDIX B

SECOND DRAFT OF CODING SHEET

Coder Name:

Study No:

Title of the Study:

Author(s) (Surname, Name):

34. Publication Date:

35. Publication Type:

a. Journal Article b. Master Thesis c. Doctoral Dissertation

36. Research Design:

a. True Experimental
b. Quasi-experimental (with random selection)
c. Quasi-experimental (without random selection)
d. Other (Specify):

37. Country:

a. Turkey b. United States c. Other (Specify):

38. Demography:

a. Urban b. Suburban c. Rural d. Unspecified

39. Type of Teaching Strategy:

40. Subject Area:

a. Physics b. Chemistry c. Biology
d. General Science e. Other (Specify):

41. Topic:

Unit:

Subject:

42. School Level:

a. Primary b. Secondary c. College

d. University e. Other (Specify):

43. Private or Public (State) School:

a. Private School b. State School c. Mixed

44. School Type:

a. Elementary School b. Anatolian High School

c. Vocational School d. Regular High School d. Other (Specify):

45. Grade Level:

46. Characteristic of Population:

a. Gifted Students b. Students with Learning Difficulties

c. Normal

47. (If your answer is "c" for the previous question) Ability Level of Student:

a. Below average b. Average c. Above average d. Mixed

e. Unspecified

48. Gender:

a. All Female b. All Male c. Mixed d. Unspecified

49. Socio Economic Status:

a. Below average b. Average c. Above average d. Mixed

e. Unspecified

50. Length of Treatment (in months, weeks, days, or hours):

51. Sample Size:

52. Type of the Assessment Instrument 1:

a. Standardized Test b. Researcher Developed Test c. Both

d. Unspecified

Type of the Assessment Instrument 2:

a. Standardized Test b. Researcher Developed Test c. Both

d. Unspecified

Type of the Assessment Instrument 3:

a. Standardized Test b. Researcher Developed Test c. Both

d. Unspecified

53. Application Time of Posttest:

a. Just After Treatment b. Delayed Test c. Both

d. Unspecified

54. Teacher Effect:

a. Same teacher for both control and experimental group

b. Different teachers

c. Unspecified

55. Researcher Effect:

a. Researcher is one of the teachers

b. Researcher is not any of the teachers

c. Researcher is the only teacher

56. Teacher Training Period (in months, weeks, days, or hours):

57. Is group working used?

a. Yes b. No c. Unspecified

58. (If your answer is "a" for the previous question) Group Size:

59. Implementation Time:

60. Statistical Analysis Technique Used:

61. Type of Outcome:

a. Achievement

b. Motivation in Science (Specify if necessary):

c. Attitudes towards Science

d. Other (Specify):

e. Combination (Specify the components):

62. Science Achievement Outcome:

g. Science Process Skills

h. Concept Learning

i. Problem Solving

j. Other (Specify):

k. Combination (Specify the components):

l. Unspecified

63. Level of Control over Threats to Internal Validity

Threats to Internal Validity: Subject Characteristics, Loss of Subjects (Mortality), Location, Instrument Decay, Data Collector Characteristics, Data Collector Bias, Testing, Extraneous Events (History), Maturation, Attitude of Subjects, Regression, Implementation

Unacceptable:

None of the threats to internal validity was controlled

Poor:

1-3 of the threats to internal validity was controlled

Average:

4-6 of the threats to internal validity was controlled

Good:

7-9 of the threats to internal validity was controlled

Very Good:

All of the threats to internal validity was controlled

64. Level of Treatment Fidelity

a. Unacceptable b. Poor c. Average d. Good

d. Very Good e. Not Enough Information

65. Level of Treatment Verification

a. Unacceptable b. Poor c. Average d. Good

d. Very Good e. Not Enough Information

66. Reliability Coefficients of the Instruments:

a. Instrument 1:
b. Instrument 2:
c. Instrument 3:

67. Average Difficulty level of the instruments:

a. Instrument 1:
b. Instrument 2:
c. Instrument 3:

68. Average Distinctiveness of the instruments:

a. Instrument 1:

b. Instrument 2:

c. Instrument 3:

69. Number of items in the instruments:

a. Instrument 1:

b. Instrument 2:

c. Instrument 3:

70. Time Limit to Complete the Instruments:

a. Instrument 1:

b. Instrument 2:

c. Instrument 3:

71. Study Results:

a. Effect Size(s):

b. p(s):

c. t(s):

d. F(s):

e. Omega Square(s):

TEST 1

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 2

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 3

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 4

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 5

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

Reference:

APPENDIX C

FINAL VERSION OF THE CODING SHEET

Coder Name:

Study No:

Title of the Study:

Author(s) (Surname, Name):

1. **Publication Year:**
2. **Publication Type:**

   ___Journal Article ___Master Thesis ___Doctoral Dissertation

   ___Other (Specify): ……..
3. **Country:**

   ___Turkey ___United States ___Other (Specify): ……
4. **Research Design:**

   ___True Experimental

   ___Quasi-experimental (with randomly assigned clusters)

   ___Quasi-experimental (without randomly assigned clusters)

   ___Other (Specify): ……..
5. **Research Model:**

   ___Post-test only control group design

   ___Pre-test post-test control group design

   ___Solomon four group design

   ___Factorial design

   ___Other (Specify): ……..

6. **Sampling Method**

___Random Sampling

___Simple Random Sampling

___Stratified Random Sampling

___Cluster Random Sampling

___Two stage Random Sampling

___Other (Specify): ……..

___Nonrandom Sampling

___Systematic Sampling

___Convenience Sampling

___Purposive Sampling

___Other (Specify): ……..

___Unspecified

7. **Demography:**

___Urban ___Suburban ___Rural ___Unspecified

8. **Subject Area:**

___Physics ___Chemistry ___Biology ___General Science

___Unspecified ___Other (Specify): ……..

9. **Topic:**

Unit: ……..

Subject: ……..

10. **School Level:**

___Primary ___Secondary ___College ___University

___Unspecified ___Other (Specify): ……..

11. **Private or Public (State) School:**

___Private School ___State School ___Mixed ___Unspecified

12. **(If school level is "Secondary") School Type:**

___Anatolian High School ___Vocational School

___Regular High School ___Mixed ___Unspecified

13. **Grade Level:** ……..

14. **Sample Size:** ……..

15. **Age (mean in years):** ……..

16. Characteristic of Population:

___Gifted Students ___Students with Learning Difficulties ___Normal

17. (If your answer is "Normal" for the previous question) Ability Level of Student:

___Below average ___Average ___Above average

___Mixed ___Unspecified

18. Gender:

___All Female ___All Male ___Mixed ___Unspecified

19. Socio Economic Status:

___Below average ___Average ___Above average

___Mixed ___Unspecified

20. Total Length of Treatment (in months, weeks, days, or hours):

……..

21. Is length of treatment same for the control and experimental conditions?

___Yes

___No (Indicate them separately):

For experimental group: ……..

For control group: ……..

___Unspecified

22. Type of Outcome:

___Achievement

___Motivation in Science (Specify if necessary):

___Attitudes towards Science

___Science Process Skills

___Other (Specify): ……..

23. Science Achievement Outcome:

___Conceptual Understanding ___Quantitative Problem Solving

___Both ___Unspecified

24. How are dependent variables measured?

___Using paper-pencil test

___including only objective type questions

___including only open-ended questions

___including both objective type and open-ended questions

___no information about type of questions

___Using process assessment

___Using product assessment

___Unspecified

25. Type of Teaching Methods:

……..

26. PBL is used as

___a curriculum model

___a teaching method

___Unspecified

27. Is there any method integrated to PBL?

___Yes (specify):

___No

___Unspecified

28. Is group working used?

___Yes ___No ___Unspecified

29. (If your answer is "Yes" for the previous question) Group Size:

……..

30. Have the problem statements been aligned with students' interests?

___Yes (specify how?)

___No

___Unspecified

31. Teacher Effect:

___Same teacher for both control and experimental group

___Different teachers

___Unspecified

32. Researcher Effect:

__Researcher is one of the teachers __Researcher is not any of the teachers

__Researcher is the only teacher __Unspecified

33. Teacher Training Period (in months, weeks, days, or hours):

……..

34. Background information about the teachers:

a. **Teaching Experience (in years):** ……..

b. **Does he/she have a Master or PhD Degree?** ……..

35. Implementation Year: ……..

36. Type of the Assessment Instrument 1:

__Pre-existing Test __Researcher Developed Test __Adapted Test

__Unspecified

Type of the Assessment Instrument 2:

__Pre-existing Test __Researcher Developed Test __Adapted Test

__Unspecified

Type of the Assessment Instrument 3:

__Pre-existing Test __Researcher Developed Test __Adapted Test

__Unspecified

37. Application Time of Posttest:

__Just After Treatment__Delayed Test __Both __Unspecified

38. Internal Reliability Coefficients of the Instruments:

a. Instrument 1: ……..

b. Instrument 2: ……..

c. Instrument 3: ……..

39. Average Item Difficulty for Each Instrument:

a. Instrument 1: ……..

b. Instrument 2: ……..

c. Instrument 3: ……..

40. Average Item Discrimination for Each Instrument:

a. Instrument 1: ……..

b. Instrument 2: ……..

c. Instrument 3: ……..

41. Number of items in the instruments:

a. Instrument 1: ……..
b. Instrument 2: ……..
c. Instrument 3: ……..

42. Time Limit Given to Complete the Instruments:

a. Instrument 1: ……..
b. Instrument 2: ……..
c. Instrument 3: ……..

43. Inferential Statistical Analysis Technique Used:

……..

44. The Extent to Which the Assumptions of Effect Size Estimation Have Been Met:

___None
___Normality
___Homogeneity of variances
___Independence of observations
___Unspecified

45. Level of Control over Threats to Internal Validity

___Unacceptable ___Poor ___Average ___Good ___Very Good

46. Level of Treatment Verification

___Unacceptable ___Poor ___Average ___Good ___Very Good
___Unspecified

47. Are the results for males and females provided separately?

___Yes (specify study results separately in the next item)
___No

48. Study Results:

Statistical Analysis I:

……..

Statistical Analysis II:

……..

Statistical Analysis III:

……..

TEST 1

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 2

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 3

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 4

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

TEST 5

	Control	Treatment 1	Treatment 2	Treatment 3
Number of Students				
Mean				
Standard Deviation				

49. Full Reference of the Study:

APPENDIX D

CODING MANUAL

Directions:

The coding sheet consists of 49 items on eight pages. For the items with multiple choices, you are expected to select one (or more than one for some items) which is (are) the most appropriate for the study you are coding. For some of the items, you are expected to write short answers on the spaces provided. If there is not enough information provided by the authors about what is asked on the item, label it as "unspecified" by selecting or writing it explicitly.

The following instructions start with a clear explanation what you are expected to do for each item and then (if necessary) some important points are highlighted on the "be aware of that" part. Please, read and try to follow the instructions as strictly as possible to be able to establish high inter-coder reliability.

1. Publication Year:

Write the publication year of the study.

Be aware of that:

Implementation year may be different from publication year and for this item publication year are asked to be written.

2. Publication Type:

Indicate whether it is a journal article, thesis, dissertation or other kind of publications like a presentation in a conference or meeting or an ERIC document etc…

3. Country:

Indicate the country where the study has been implemented.

Be aware of that:

The country in which the study has been published may be different from the one it has been implemented. Be careful that, in this item, "country" refers to the one the research has been implemented.

4. Research Design:

Decide whether the research has been designed as a true or quasi experiment. Make your decision based on explanation about the details of the research. Your decision may not be same with what the author(s) indicates about the type of research design.

Be aware of that:

Random assignment means that every participant in the experiment has an equal chance of being assigned to experimental or control group while random selection means that every member of the population has an equal chance of being selected for the sample.

Random assignment is essential for true experimental design.

Randomly assigned clusters mean that not the individual participants but the clusters (i.e. classes) are assigned randomly to any of experimental or control group conditions.

5. Research Design Details:

Select appropriate research design for the study.

6. Sampling

Select appropriate sampling procedure for the study.

Be aware of that:

If sampling procedure consists of more than one stage and different sampling methods are used in different stages, select the most nonrandom one in any of these stages.

7. Demography:

Select the appropriate demographic information for the sample of the research.

Be aware of that:

Please, do not make subjective judgments about this item. Just code as what is stated by authors on the paper.

8. Subject Area:

Select the appropriate subject area for the study.

Be aware of that:

Even if the subject is labeled as general science (or similar terms) on the paper, code this item as physics, chemistry or biology if it is obvious that the topic is related to one of them. For example, if the topic is "force and motion", label the subject as physics even if it is called as general science (or science and technology) on the paper by the authors.

9. Topic:

Write the topic in exactly the same way it is reported on the paper.

10. School Level:

Select the appropriate school level the study has been implemented.

11. Private or Public (State) School:

Indicate whether the research has been implemented in a private or state school.

Be aware of that:

If both private and state schools are included in study, select "mixed" option.

12. (If the school level is secondary) School Type:

If the school level is secondary, please select the appropriate school type for the study.

Be aware of that:

Select "mixed" option, if more than one school type exists in the study.

13. Grade Level:

Indicate the grade level of sample of the study.

Be aware of that:

If more than one grade level exists, indicate all of them separately.

14. Sample Size:

Record the sample size of study.

Be aware of that:

Sample size refers to total number of the participants in both control and experimental conditions included in the main study. If there are different types of sample (i.e. from different school types or levels), record sample size for each type separately. Please note that sample size covers only the participants in the main study, so (if exists) exclude the sample in the pilot study.

15. Age (mean in years):

Write down the average age of the participants of the study if it is provided on the paper.

16. Characteristic of Population:

Indicate whether the population of the study is special (i.e. gifted students or students with learning difficulties).

Be aware of that:

Please, do not make subjective judgments about this item. Just code as what is stated on the paper.

17. (If your answer is "Normal" for the previous question) Ability Level of Student:

If the population of the study is labeled as normal (not special), then indicate the ability level of the students based on the information provided by the authors.

Be aware of that:

Please, do not make subjective judgments about this item. Just code as what is stated by authors on the paper. Select the "unspecified" option if it is not specified by the authors explicitly.

18. Gender:

Select the appropriate choice for the gender of the sample.

19. Socio Economic Status:

Select the appropriate choice for the socio economic status of the sample.

Be aware of that:

Please, do not make subjective judgments about this item. Just code as what is stated on the paper by the authors. Select the "unspecified" option if it is not specified by the authors explicitly.

20. Total Length of Treatment (in months, weeks, days, or hours):

Record total length of treatment, which is the time interval between beginning and end of the implementation, as it is stated on the paper.

Be aware of that:

Do not forget to specify the unit (i.e. month, week, day or hour).

21. Is length of treatment same for the control and experimental conditions?

Please check whether the length of treatment is same for both control and experimental conditions.

Be aware of that:

If length of treatment is not same for the control and experimental conditions, record them separately and do not forget to specify the unit (i.e. month, week, day or hour. Select the "unspecified" option if it is not specified by the authors explicitly.

22. Type of Outcome:

Put a sign on the space provided before the outcome(s), which has been measured as dependent variable of the study.

Be aware of that:

It is possible that more than one type of outcome has been measured in a study. If so, you are expected to put a sign more than once.

23. Science Achievement Outcome:

Identify whether the assessment tool to measure science achievement outcome (if exists) focus on conceptual understanding, quantitative problems or both.

24. How are dependent variables measured?

Indicate whether dependent variables are measured in the study by using paper-pencil test, or assessing process and product (i.e. outcome) as well. Please indicate the type of questions included in the paper-pencil test if it is used.

Be aware of that:

It is possible that more than one type of measurement has been used in a study. If so, you are expected to put a sign more than once.

25. Type of Teaching Methods:

Write down the teaching methods used in both experimental and control group conditions, like "traditional instruction vs. problem based learning ".

Be aware of that:

Please, use exactly the same phrases for teaching methods as it is indicated on the paper. For example, if the authors call the teaching method as case-based learning rather than problem based learning (PBL), write it exactly in the same way.

26. PBL is used as:

Decide whether PBL is used as a teaching model or a teaching method. Please make your decision based on the following definitions:

Teaching model refers to an instructional design or a curriculum based on PBL, in which different teaching methods can be used. For example, sometimes in medical education, PBL is used as a teaching model, which shapes all curricula in a medical school. On the other hand, teaching method has a more specific meaning with clear (basic) steps.

27. Is there any method integrated to PBL?

Please check whether there exists any other teaching method or strategy integrated to PBL. If exists, specify the integrated method on the space provided

after “Yes”. If you cannot be sure because of insufficient information, select the choice of “unspecified”.

28. Is group working used?

Indicate whether group working is used during PBL.

Be aware of that:

In this item, group working refers to the fact that more than one student study together during PBL. It does not have to be necessarily cooperative or collaborative learning.

29. (If your answer is “Yes” for the previous question) Group Size:

If group working is used in the study, specify the group size, which may be an exact number (3 students per group) or an interval (3-5 students per group).

Be aware of that:

If no group working is used or there exists no information whether group working is used or not, leave this item as blank (meaning not applicable). However, if you know that group working is used in the study but there is no information about group size; label the item as “unspecified”.

30. Have the problem statements been aligned with students’ interests?

This item aims to clarify whether students’ interests have been investigated before the problem statements have been established. That is, label this item as “yes” if the contexts for the problems have been decided according to the interests of the students in the sample group.

31. Teacher Effect:

Indicate whether both control and experimental groups have been instructed by the same teacher.

Be aware of that:

This item does not aim to discriminate whether the researcher is one of the teachers. If both control and experimental groups have been instructed by the same researcher, we should select “same teacher for both control and experimental group” option.

32. Researcher Effect:

Indicate whether researcher(s) has been involved in any of the control or experimental groups as a teacher.

Be aware of that:

If the researcher(s) has been involved in the groups beside the regular teacher to assist him/her, still label the item as "researcher is one of the teachers". However, if the researcher(s) has been involved in the groups just to observe the lessons (for treatment verification or any other purpose) and has not been taken part into instruction, then select "researcher is not any of the teachers".

33. Teacher Training Period (in months, weeks, days, or hours):

Write down the duration for teacher training about the implementation of this specific study.

Be aware of that:

Please do not forget to specify the unit (i.e. month, week, day or hour).

34. Background Information about the Teachers:

Write down teaching experience of the teachers in years and indicate whether they have a master or PhD degree.

Be aware of that:

If there is more than one teacher in the study, complete the necessary information for each of them separately indicating whether he/she teaches in the control or experimental group or both.

35. Implementation Year:

Record when the research has been implemented (i.e., in 2002, or in the fall semester of 2005)

Be aware of that:

Implementation time is not duration of the treatment but it is when the implementation has been taken place. Please note that it is not the publication year and record as "unspecified" if it is not specified by the authors explicitly.

36. Type of the Assessment Instrument:

Select the appropriate type of assessment instrument.

Be aware of that:

"Pre-existing test" refers to the tests that have already been developed by other researchers and available in the literature. These tests do not have to be standardized ones. Just being pre-existing is enough to label the test as pre-existing test.

"Researcher developed test" means the test has been developed by the authors for this study. The test had not been available in the literature before this study and it is totally original, not an adaptation of the pre-existing test.

"Adapted test" refers to the tests that have been adapted from one or more pre-existing tests for this study by the authors. However, the adapted version of the pre-existing test has not been used before for another study.

If there are more than three assessment instruments, please add extra rows for them to code necessary information.

37. Application Time of Posttest:

Record the application time of the posttest in reference to the treatment.

Be aware of that:

"Both" means posttest has been used not only just after treatment but also as a delayed test.

38. Internal Reliability Coefficients of the Instruments:

Record reliability coefficients for each instrument used in the study separately.

Be aware of that:

Please indicate (if provided on the paper) which reliability coefficient you report, that is KR-20, KR-21 or Cronbach's alpha. If there are more than three assessment instruments, please add extra rows for them to code necessary information.

39. Average Item Difficulty for Each Instrument:

Record average difficulty level for each instrument used in the study separately.

40. Average Item Discrimination for Each Instrument:

Record average distinctiveness for each instrument used in the study separately.

41. Number of items in the instruments:

Record number of items for each instrument used in the study separately.

42. Time Limit Given to Complete the Instruments:

Record time limit given to complete each instrument used in the study separately. Please *label as "unspecified" if it is not specified by the authors explicitly.*

43. Inferential Statistical Analysis Technique Used:

Indicate the inferential statistical analysis technique(s) used to check statistical significance of the test results of the related dependent variables.

Be aware of that:

If more than one has been used in the study, write all statistical analysis techniques separately indicating which ones for which dependent variable.

44. The extent to which the assumptions of effect size estimation have been met:

Decide the extent to which the assumptions of effect size estimation have been met according to the following criteria:

None: No information whether any of the assumptions have been checked or met.

Normality: "Normality" assumption have been checked and met.

Homogeneity of variances: "Homogeneity of variances" assumption have been checked and met.

Independence of observations: "Independence of observations" assumption have been checked and met.

Be aware of that:

That the authors have not mentioned how they conclude that the assumptions of effect size estimation have been met does not necessarily mean that they have not checked these assumptions. In this item study quality and reporting quality interfere and it is impossible to distinguish which one results into providing no information about assumptions. So, to be able to get an idea about statistical conclusion validity, we assume that if they check the assumptions, they report the details on the paper.

Put a sign on the space provided before each item if and only if the authors state explicitly that they have checked the related assumption and conclude that it has been met.

45. Level of Control over Threats to Internal Validity

Decide the extent to which threats to internal validity have been controlled using the following list of possible threats and criteria:

Threats to Internal Validity: Subject Characteristics, Loss of Subjects (Mortality), Location, Instrument Decay, Data Collector Characteristics, Data Collector Bias, Testing, Extraneous Events (History), Maturation, Attitude of Subjects, Regression, Implementation

Unacceptable: None of the threats to internal validity was controlled

Poor: 1-3 of the threats to internal validity were controlled

Average: 4-6 of the threats to internal validity were controlled

Good: 7-9 of the threats to internal validity were controlled

Very Good: All of the threats to internal validity were controlled

Be aware of that:

That the authors do not mention how they have controlled the possible threats to internal validity does not necessarily mean that they have not done anything for these threats. In this item, study quality and reporting quality interfere and it is impossible to distinguish which one results into providing no information about internal validity. So, to be able to get an idea about the degree of internal validity, we assume that if they have controlled any of these threats, they report the details

on the paper. Thus, any finding from this item is limited by what is reported on the paper by the authors.

46. Level of Treatment Verification

In this item, you are expected to decide the level of treatment verification using the information about the procedure on the paper according to the following criteria:

Unacceptable: None of the lessons in any of the groups have been observed for treatment verification.

Poor: Less than half of the lessons in only experimental group have been observed by only researcher(s).

Average: Approximately half of the lessons in only experimental group have been observed by only researcher(s).

Good: Approximately half of the lessons in both experimental and control group have been observed by only researcher(s).

Very Good: At least half of the lessons in both experimental and control group have been observed by at least an expert other than researcher(s) with or without researcher(s).

Be aware of that:

If the authors give no information whether they have conducted treatment verification, please do not make subjective judgment based on procedure part and label the item as "not enough information".

47. Are the study results for males and females provided separately?

Indicate whether the study results for males and females are provided separately. If they are presented separately, record the results for males and females as well in the scope of the next item.

48. Study Results:

Record the results for each instruments administered during the study.

Be aware of that:

Even you record the effect size calculated by the authors, please code as much information as possible about the study results like effect size, p, t, F, eta square,

omega square, pre-test post-test correlation values for each inferential statistics analysis besides descriptive results for each assessment.

49. Full Reference of the Study:

Write down full reference of the study according to APA Style.

APPENDIX E

DESCRIPTIVE DATA FOR THE ITEMS IN THE CODING SHEET

Variable	*Groups*	*Number of Studies*	*Percentage (%)*
Publication Year	1994-1999	3	3
	2000-2005	14	16
	2006-2012	71	81
Country	Canada	1	1
	China	1	1
	Egypt	1	1
	Hong Kong	2	2
	Malaysia	1	1
	Netherlands	1	1
	Serbia	1	1
	South Africa	1	1
	Spain	1	1
	Turkey	54	62
	United Arab E.	1	1
	USA	23	26
Research Design	Quasi experimental with RAC	48	54
	Quasi experimental without RAC	20	23
	True experimental	20	23

Variable	*Groups*	*Number of Studies*	*Percentage (%)*
Research Model	Counterbalanced design	2	2
	Post-test only control group design	7	8
	Pre-test post-test control group design	79	90
Sampling Procedure	Cluster random sampling	1	1
	Convenience sampling	76	86
	Purposive sampling	6	7
	Simple random sampling	1	1
	Stratified random sampling	1	1
	Volunteer sampling	3	3
Demography	Urban	54	61
	Rural	4	5
	Unspecified	30	34
Subject Area	Biology	14	16
	Chemistry	18	20
	General Science	4	5
	Physics	26	30
	Others	26	30
School Level	College	2	2
	Post-graduate	2	2
	Primary	26	30
	Secondary	17	19
	Undergraduate	41	47
Private or Public School	Private school	5	6
	Public school	52	59
	Unspecified	31	35
Sample Size	0-50	22	25
	51-100	43	49
	101-150	14	16
	151-200	4	5
	Over 200	5	6

Variable	*Groups*	*Number of Studies*	*Percentage (%)*
Characteristic of Population	Gifted students	2	2
	Students with learning disabilities	0	0
	Normal	86	98
Ability Level of Students	Below average	2	2
	Average	40	47
	Above average	2	2
	Unspecified	42	49
Gender	Only female	1	1
	Only male	0	0
	Mixed	77	88
	Unspecified	10	11
Socio Economic Status	Below average	3	3
	Average	5	6
	Above average	2	2
	Unspecified	78	89
Length of Treatment	0-5	32	36
	6-10	26	30
	11-15	10	11
	Over 15	9	10
	Unspecified	11	13
Is the length of treatment same or not?	Yes	32	36
	No	0	0
	Unspecified	56	64
How are dependent variables measured?	Both objective type and open-ended questions	21	24
	Only objective type questions	63	72
	Only open-ended Questions	3	3
	Unspecified	1	1

Variable	*Groups*	*Number of Studies*	*Percentage (%)*
PBL Mode	A curriculum model	18	20
	A teaching method	70	80
Is there any method integrated to PBL?	Yes	5	6
	No	42	48
	Unspecified	41	47
Is group work used?	Yes	77	88
	No	0	0
	Unspecified	11	12
Group Size	4-6	44	50
	7-9	13	15
	Over 9	7	8
	Unspecified	24	27
Have the problem statements been aligned with students' interests?	Yes	1	1
	No	38	43
	Unspecified	49	56
Teacher Effect	Different teachers	24	27
	Same teachers	41	47
	Unspecified	23	26
Researcher Effect	Not any of teachers	36	41
	One of teachers	10	11
	The only teacher	19	22
	Unspecified	23	26
Type of Assessment Instrument	Adapted	5	6
	Pre-existing	42	48
	Researcher-developed	41	47
Application Time of Post-test	Delayed test	1	1
	Just after treatment	82	93
	Both	5	6

APPENDIX F

CODER RELIABILITY DATA

The items, which have been agreed upon in the first and second coding by the researcher, are labeled as "1" while the ones, which have been coded differently, are represented by "0" in the table

Item	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Study 7	Study 8	Study 9	Study 10
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	1	0	1	1	1	1	0	1	1	1
7	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1

Item	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Study 7	Study 8	Study 9	Study 10
11	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1	1	1
17	0	1	1	1	1	1	1	1	0	1
18	1	1	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1
21	1	1	1	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1	1	1	1
23	1	1	1	1	1	1	1	1	1	1
24	1	1	1	1	1	1	1	1	1	1
25	1	1	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1	1	1
29	1	1	1	1	1	1	1	1	1	1
30	1	1	0	1	1	1	0	1	1	1

Item	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Study 7	Study 8	Study 9	Study 10
31	1	1	1	1	1	1	1	1	1	1
32	1	1	1	1	1	1	1	1	1	1
33	1	1	1	1	1	1	1	1	1	1
34	1	1	1	1	1	1	1	1	1	1
35	1	1	1	1	1	1	1	1	1	1
36	1	1	1	0	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1	1	1
38	1	1	1	1	1	1	1	1	1	1
39	1	1	1	1	1	1	1	1	1	1
40	1	1	1	1	1	1	1	1	1	1
41	1	1	1	1	1	1	1	1	1	1
42	1	1	1	1	1	1	1	1	1	1
43	1	1	1	1	1	1	1	1	1	1
44	1	1	0	1	1	1	1	1	1	1
45	0	1	1	1	1	1	1	1	1	0
46	1	1	1	1	1	1	1	1	1	1
47	1	1	1	1	1	1	1	1	1	1
48	1	1	1	1	1	1	1	1	1	1
Score	46	47	46	47	48	48	46	48	47	47
AR	0.958	0.979	0.958	0.979	1.000	1.000	0.958	1.000	0.979	0.979
AR (Average)	0.979									

APPENDIX G

INTER-CODER RELIABILITY DATA

The items, which have been coded consistently by two coders, are labeled as "1" while the ones, which have been coded differently, are represented by "0" in the table.

	Coder 1		Coder 2		Coder 3		Coder 4		Coder 5		Coder 6		Coder 7	
Item No	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	0	1	1	1	1	1	1	1	1
4	1	1	1	0	1	1	1	1	1	0	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	0	1	1
6	0	1	0	0	1	1	1	1	1	1	1	0	0	0
7	0	1	1	1	0	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	0	1	1	1	1	1	1
10	1	1	1	0	1	1	1	1	1	1	1	1	1	1
11	1	1	0	0	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	0	1	1	1	1	1	1	1	1	1

	Coder 1		Coder 2		Coder 3		Coder 4		Coder 5		Coder 6		Coder 7	
Item No	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Item No	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	0	1	1	1	1	1	1	1
15	1	1	0	1	1	0	0	1	0	0	1	0	1	1
16	1	1	1	1	1	1	0	1	1	1	1	1	1	1
17	0	1	1	1	1	1	1	0	1	0	0	1	0	0
18	1	0	0	1	1	1	1	0	1	1	1	1	0	0
19	0	1	1	1	1	1	0	1	1	1	1	1	1	1
20	1	1	1	0	1	1	1	1	1	1	1	1	1	1
21	0	0	1	1	1	1	1	1	1	1	1	1	1	0
22	1	1	0	1	1	1	1	1	1	1	0	1	1	1
23	1	1	1	1	1	1	1	1	1	1	0	0	1	1
24	1	1	1	1	0	1	1	1	0	0	1	1	1	1
25	1	1	1	1	1	1	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1	1	1	1	1	1	0
27	1	1	1	1	1	1	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1	1	1	1	1	1	1
29	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30	0	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1
32	1	1	1	0	1	0	1	1	1	1	0	0	1	0
33	1	1	1	1	1	1	1	1	1	1	1	1	1	1
34	1	1	1	1	1	1	1	1	0	1	1	1	1	1
35	1	1	1	1	1	1	1	1	0	1	1	1	1	1

	Coder 1		Coder 2		Coder 3		Coder 4		Coder 5		Coder 6		Coder 7	
Item No	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
36	1	1	1	0	1	1	0	0	1	1	0	1	1	1
37	1	1	1	1	1	1	1	1	1	0	1	1	1	1
38	1	1	0	1	1	1	1	1	1	1	1	1	1	1
39	1	1	1	1	1	1	1	1	1	1	1	1	1	1
40	1	1	1	1	1	1	1	1	1	1	1	1	1	1
41	1	1	1	1	1	1	1	1	1	1	1	1	1	1
42	1	1	1	1	1	1	1	1	1	1	1	1	1	1
43	1	1	1	1	1	1	1	1	1	1	1	1	1	1
44	1	1	1	1	1	1	1	1	1	1	1	1	1	1
45	1	0	0	0	1	1	0	1	1	1	1	1	1	1
46	1	1	1	0	0	1	1	1	0	1	1	1	1	1
47	1	1	1	1	1	1	1	1	1	1	0	1	1	1
48	0	0	1	0	1	1	1	1	1	1	0	1	1	1
Score	41	44	41	38	44	45	42	43	43	43	41	43	45	42
AR	0.854	0.917	0.854	0.792	0.917	0.938	0.875	0.896	0.896	0.896	0.854	0.896	0.938	0.875
AR	0.885		0.823		0.927		0.885		0.896		0.875		0.906	
AR (Average)	0.885													

APPENDIX H

LIST OF EFFECT SIZES REVEALED FROM PRIMARY STUDIES

Study	Standardized mean difference	SE	Hedge's g	SE
Adalı (2005)	1.33	0.24	1.32	0.23
Adalı (2005)	1.03	0.23	1.02	0.23
Adiga and Adiga (2011)	0.91	0.19	0.91	0.19
Akın (2008)	0.70	0.27	0.69	0.27
Akın (2008)	0.77	0.27	0.76	0.27
Akınoğlu and Tandoğan (2007)	0.64	0.29	0.63	0.29
Alagöz (2009)	-0.18	0.25	-0.17	0.25
Alagöz (2009)	-0.45	0.25	-0.45	0.25
J. C. Anderson (2007)	0.04	0.19	0.04	0.19
Araz (2007)	0.21	0.14	0.21	0.14
Atan, Sulaiman, and Idrus (2005)	0.40	0.30	0.40	0.30
B. Bayrak and Bayram (2011)	1.77	0.32	1.75	0.31
R. Bayrak (2007)	0.75	0.23	0.74	0.23
R. Bayrak (2007)	0.99	0.23	0.98	0.23
R. Bayrak (2007)	0.65	0.23	0.64	0.22
Bayram (2010)	0.36	0.26	0.36	0.25
Benli (2010)	2.21	0.31	2.18	0.31
Benli (2010)	0.02	0.25	0.02	0.24
Benli (2010)	0.91	0.26	0.90	0.26
Benli (2010)	2.56	0.35	2.53	0.35
Benli (2010)	2.68	0.33	2.65	0.33
Benli (2010)	0.75	0.24	0.74	0.24
Bilgin, Şenocak, and Sözbilir (2009)	0.46	0.23	0.46	0.23
Burris (2005)	-0.11	0.17	-0.11	0.17
Çam (2009)	0.51	0.26	0.51	0.26

Çam (2009)	0.73	0.27	0.72	0.27
Çam (2009)	2.44	0.34	2.41	0.33
Carll-Williamson (2003)	0.13	0.20	0.13	0.20
Carll-Williamson (2003)	0.11	0.20	0.11	0.20
Carrio et al. (2011)	-0.10	0.21	-0.10	0.21
Çelik (2010)	-0.30	0.31	-0.30	0.30
Çelik (2010)	0.64	0.32	0.62	0.31
Çelik (2010)	-0.24	0.31	-0.24	0.30
Çelik (2010)	0.81	0.32	0.80	0.32
Çınar (2007)	1.15	0.28	1.14	0.27
Çınar (2007)	3.10	0.38	3.06	0.37
Çınar (2007)	1.29	0.28	1.27	0.28
Demirel and Turan (2010)	0.74	0.32	0.72	0.31
Demirel and Turan (2010)	0.87	0.32	0.85	0.32
Demirel and Turan (2010)	1.08	0.33	1.06	0.33
Demirel and Turan (2010)	3.74	0.51	3.67	0.50
De Simone (2008)	0.69	0.24	0.68	0.23
Dieber (1994)	0.74	0.40	0.72	0.39
Diggs (1997)	1.54	0.23	1.53	0.23
Diggs (1997)	0.63	0.20	0.62	0.20
Dobbs (2008)	-0.17	0.17	-0.17	0.17
Downing, Ning, and Shin (2011)	1.11	0.19	1.10	0.19
Drake and Long (2009)	0.29	0.37	0.28	0.36
Erdem (2006)	0.73	0.23	0.72	0.23
Erdem (2006)	0.38	0.23	0.38	0.23
Gabr and Mohamed (2011)	1.16	0.13	1.16	0.13
Gabr and Mohamed (2011)	0.77	0.13	0.77	0.13
Günhan and Başer (2009)	1.72	0.35	1.69	0.34
Gürlen (2011)	0.73	0.23	0.72	0.23
Gürlen (2011)	0.38	0.23	0.38	0.23
Hesterberg (2005)	-0.37	0.21	-0.37	0.21
Hesterberg (2005)	0.06	0.22	0.05	0.22
Horzum and Alper (2006)	1.39	0.27	1.37	0.26
İnel (2009)	0.24	0.31	0.24	0.31
Jandric, Obadovic, Stojanovic, and Rancic (2011)	0.85	0.19	0.84	0.19

Jandric et al. (2011)	1.32	0.20	1.32	0.20
Kaddouro (2011)	1.48	0.23	1.47	0.23
Kanlı (2008)	2.24	0.37	2.20	0.36
Kanlı (2008)	0.42	0.29	0.41	0.29
Kanlı (2008)	0.32	0.29	0.32	0.29
Kar (2010)	0.49	0.24	0.49	0.24
Kar (2010)	-0.26	0.24	-0.26	0.23
Karaöz (2008)	1.73	0.37	1.70	0.36
Karaöz (2008)	0.39	0.32	0.39	0.31
Karaöz (2008)	1.87	0.37	1.83	0.37
Kıray and İlik (2011)	1.50	0.18	1.50	0.18
Koçakoğlu (2008)	-0.44	0.18	-0.44	0.18
Koçakoğlu (2008)	-0.47	0.19	-0.47	0.18
Könings, Wiers, van de Wiel, and Schmidt (2005)	-0.37	0.37	-0.36	0.36
Koray, Presley, Köksal, and Özdemir (2008)	0.93	0.23	0.92	0.23
Kuşdemir (2010)	0.02	0.28	0.02	0.27
Kuşdemir (2010)	1.23	0.30	1.21	0.30
Kuşdemir (2010)	1.20	0.30	1.18	0.30
LeJeune (2002)	0.49	0.51	0.47	0.48
LeJeune (2002)	0.11	0.50	0.10	0.47
LeJeune (2002)	-0.21	0.50	-0.20	0.47
Lesperance (2008)	0.43	0.44	0.42	0.42
D. L. Lewis (2006)	0.40	0.10	0.40	0.10
Lyons (2006)	-0.12	0.27	-0.12	0.27
Mathew (2008)	0.54	0.27	0.53	0.26
McGee (2003)	-0.05	0.52	-0.05	0.49
McGee (2003)	1.35	0.57	1.27	0.54
Mungin (2012)	0.12	0.35	0.12	0.34
Needham (2010)	0.01	0.38	0.01	0.37
Nowak (2002)	-0.84	0.25	-0.83	0.25
Olgun (2008)	1.33	0.24	1.32	0.23
Olgun (2008)	1.03	0.23	1.02	0.23
Özeken and Yıldırım (2011)	0.65	0.21	0.65	0.21
Rajab (2007)	0.74	0.25	0.74	0.25
Rajab (2007)	0.57	0.25	0.57	0.25
Sağır, Çelik, and Armağan (2009)	0.93	0.25	0.92	0.24
Şahbaz (2010)	0.83	0.25	0.82	0.25
Şahbaz (2010)	0.55	0.24	0.54	0.24
Şahbaz (2010)	0.91	0.25	0.90	0.25
A. Şahin (2011)	0.37	0.23	0.37	0.23

M. C. Şahin (2005)	1.26	0.18	1.25	0.18
M. C. Şahin (2005)	0.64	0.17	0.64	0.17
Şalgam (2009)	1.28	0.26	1.27	0.25
Şalgam (2009)	0.02	0.23	0.02	0.23
Sanderson (2008)	0.17	0.46	0.16	0.44
Saral (2008)	0.68	0.23	0.67	0.23
Scott (2005)	-1.10	0.32	-1.08	0.32
Selçuk, Karabey, and Çalışkan (2011)	0.96	0.27	0.95	0.27
Semerci (2006)	0.84	0.27	0.82	0.27
Şendağ (2008)	0.66	0.32	0.65	0.32
Şenocak, Taşkesenligil, and Sözbilir (2007)	0.41	0.20	0.41	0.20
Şenocak, Taşkesenligil, and Sözbilir (2007)	0.50	0.20	0.50	0.20
Serin (2009)	-0.16	0.21	-0.16	0.21
Serin (2009)	0.28	0.21	0.28	0.21
Serin (2009)	0.02	0.21	0.02	0.21
Shepherd (1998)	1.22	0.37	1.19	0.36
Sungur et al. (2006)	1.42	0.29	1.40	0.28
Sungur and Tekkaya (2006)	0.46	0.26	0.45	0.26
Sungur and Tekkaya (2006)	0.58	0.26	0.57	0.26
Tarhan and Acar (2007)	2.30	0.41	2.25	0.40
Tarhan, Kayalı, Ürek, and Acar (2008)	1.01	0.24	1.00	0.24
Taşoğlu (2009)	0.19	0.30	0.19	0.29
Tavukçu (2006)	1.59	0.26	1.57	0.26
Tavukçu (2006)	0.72	0.23	0.72	0.23
Tavukçu (2006)	1.28	0.25	1.27	0.24
Tavukçu (2006)	1.03	0.24	1.02	0.24
Tiwari, Lai, So, and Yuen (2006)	0.56	0.24	0.56	0.23
Tiwari, Lai, So, and Yuen (2006)	0.61	0.27	0.61	0.26
Tiwari, Lai, So, and Yuen (2006)	0.32	0.25	0.32	0.25
Tosun (2010)	1.04	0.25	1.03	0.25
Tosun (2010)	0.43	0.25	0.42	0.24
Tüysüz, Tatar, and Kuşdemir (2010)	2.13	0.35	2.10	0.34
Tüysüz et al. (2010)	2.49	0.37	2.46	0.36
Ülger (2011)	0.71	0.24	0.70	0.24
Usoh (2003)	0.14	0.23	0.14	0.23

van Loggerenberg-Hattingh (2003)	0.22	0.17	0.22	0.17
Yalçınkaya (2010)	1.21	0.19	1.21	0.19
Yalçınkaya (2010)	0.33	0.18	0.32	0.18
Yalçınkaya (2010)	0.38	0.18	0.37	0.18
Yalçınkaya (2012)	0.86	0.29	0.85	0.28
Yaman (2005)	0.34	0.14	0.34	0.14
Yaman and Yalçın (2005b)	0.28	0.14	0.28	0.14
Yaman and Yalçın (2005b)	0.29	0.14	0.28	0.14
Yaman and Yalçın (2005a)	0.52	0.14	0.52	0.14
Yuan, Kunaviktikul, Klunklin, and Williams (2008)	0.63	0.30	0.61	0.30
Yurd and Olgun (2008)	1.36	0.22	1.35	0.22
Yurd (2007)	1.05	0.21	1.04	0.21

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Üstün, Ulaş
Nationality: Turkish (TC)
Date and Place of Birth: June 4, 1980, Mersin
Marital Status: Married
e-mail: ulasustun@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
Ph.D.	METU, Physics Education	2012
MS	METU, Physics Education	2005

WORK EXPERIENCE

Year	Place	Enrollment
2006-2011	METU, SSME	Research Assistant

FOREIGN LANGUAGES

- English (ÜDS: 98.75; KPDS: 91)

PUBLICATIONS

Köseoğlu, F., Tümay, H., & Üstün, U. (2010). Bilimin doğası öğretimi mesleki gelişim paketinin geliştirilmesi ve öğretmen adaylarına uygulanması ile ilgili tartışmalar. *Kırşehir Eğitim Fakültesi Dergisi, 11*(4), pp.129-162.

Üstün, U. (2010). The comparison of Finnish and Turkish Curricula. *Procedia - Social and Behavioral Sciences, 2*(2), pp. 2789-2793.

Üstün, U. & Eryılmaz, A. (2010). Which definition(s) of weight do we teach? Which one is correct? In Taşar, M.F. & Çakmakcı, G. (Eds.), *Contemporary science education research: International perspectives*, pp. 179-184. Ankara, Turkey: Pegem Akademi.

Gegenfurtner, A., Laine, E., & Üstün, U. (2010). *Moderating effects on the relation between training motivation and behavior change*. Paper presented at the 5th EARLI SIG 14 Learning and Professional Development Conference, Munich.

Üstün, U. & Eryılmaz, A. (2010). Finlandiya'nın PISA'da Fen okuryazarlığı alanındaki başarısının sebepleri ve alınabilecek dersler. *IX. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi,* 23-25 Eylül, İzmir, p.122.

Taşdelen, U., Üstün, U., Küçüker, S., & Özdem, Y. (2010). Öğretmen ve öğretmen adaylari için bilimin doğası öğretimi mesleki gelişim paketi - Proje tabanli öğrenme etkinlikleri. *IX. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi,* 23-25 Eylül, İzmir, p.130.

Damar, S. Y., Üstün U., & Eryılmaz A. (2009). Cognitive strategies used by pre-service physics teachers. Bilsel, A. & Garip, M. U. (Eds.), *Frontiers in Science Education Research*, pp. 311-316. Famagusta, North Cyprus: Eastern Mediterranean University Press.

Üstün, U., Eryılmaz, A. & Gülyurdu, T. (2008). Needs assessment for Turkish high school physics curriculum design and development, *GIREP MTL Workshop, Physics Curriculum Design, Development and Validation Conference CD, available at* http://lsg.ucy.ac.cy/girep2008/s_u.htm.

Damar, S. Y., Üstün U., & Eryılmaz A. (2008). Promoting pre-service physics teachers' meta-cognitive skills through self-evaluation. *GIREP MTL Workshop, Physics Curriculum Design, Development and Validation Conference CD, available at* http://lsg.ucy.ac.cy/girep2008/v_z.htm.

Üstün U. & Eryılmaz A. (2008). Hangi ağırlık tanımını öğretiyoruz? Hangisi doğru? *VIII. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi*, 27-29 Ağustos, Bolu, p.412.

Üstün U., Damar, S. Y., & Eryılmaz A. (2008). Öğretmen adaylarının lise seviyesindeki fizik ve pedagojik alan bilgilerini artırmak için açılan derslerin verimliliği. *VIII. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi*, 27-29 Ağustos, Bolu, p.200.

Üstün, U., Eryılmaz, A. & Gülyurdu, T. (2008). Lise Fizik öğretim programının geliştirilmesi için ihtiyaç analizi çalışmaları. *VIII. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi*, 27-29 Ağustos, Bolu, p.314.

Üstün U., Damar, S. Y., & Eryılmaz A. (2008). Fizik kavramlarının yaşam temelli verilmesi ile ilgili uygulama çalıştayı**.** *VIII. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi*, 27-29 Ağustos, Bolu, p.299.