Chapter I

1.1. Introduction

Scholars have used statistical testing for research purposes since the early 1700s (Huberty, 1993). In the past 300 years, applications of statistical testing have advanced considerably, most noticeably with the advent of the computer and recent technological advances. However, much of today’s statistical testing is based on the same logic used in the first statistical tests and advanced in the early twentieth century through the work of Fisher, Neyman, and the Pearson family. Specifically, significance testing and hypothesis testing have remained at the cornerstone of research papers and the teaching of introductory statistics courses. Currently, we are in an era where the value of statistical significance testing is being challenged by many researchers. Research methodology literature in recent years has included a full frontal assault on statistical significance testing. The assault is based on whether or not statistical significance testing has value in answering a research question posed by the investigators. In a nutshell, statistical significance testing is conducted to evaluate the viability of null hypothesis by assessing how likely some observed sample statistic could have occurred as the result of random sampling variation for a given population parameter. More specifically, statistical significance testing answers the question: what is the probability of obtaining an observed sample statistic for a given or known
population parameter? The real meaning of statistical significance testing, however, has often been lost in research practice, and the importance of statistical significance tends to be greatly exaggerated. A brief review of criticisms attributed to statistical significance testing provides information about what is presented (comparison of meta-analytical approaches) in this study. In addition, information on statistical significance testing inevitably leads us to discuss other relevant problems which social and behavioral sciences are faced in the field of research. Therefore, two main sources of the problems are described as they are related to statistical significance testing as well as main purpose of present study.

1.1.1. Criticisms on significance testing

Although the use of statistical significance testing in the analysis of research data is still almost universal, it has now become a controversial practice. Two recent articles (Cohen, 1994; Schmidt, 1996) have argued strongly that the long-standing practice of reliance on significance testing is logically indefensible and retards the research enterprise by making it difficult to develop cumulative knowledge. Some of the major criticisms on statistical significance testing are described in the following paragraphs.

One of the most important contemporary criticisms emphasizes the need that researchers must evaluate the practical importance of results, along with testing for statistical significance. Kirk (1996) agreed that statistical significance testing is a necessary part of a statistical analysis. However, he asserted that the time had come to include practical significance in the results. He recommended the use of
statistical significance testing; however, it must be considered in combination with other criteria.

The second frequently expressed concern is simply that statistical significance testing identifies statistically significant results but not necessarily practically significant results. This criticism is partly a product of the relationship between the size of a study (number of participants or degree of freedom) and statistical significance. Even the smallest relationships can become statistically significant, if a large enough sample is used for the study. Thus, the argument goes, many studies are getting published and taking on disproportionate importance because they demonstrate statistical significance, even though the magnitude of the effect measured has no practical value.

The third criticism of the use of statistical significance testing is that the null hypothesis, which is fundamental to all statistical significance testing, is often misunderstood and misinterpreted (Kirk 1996). The typical null hypothesis assumes in advance that there are no differences between groups or, in the case of continuous variables; there is no relationship between the variables. The significance test then determines the probability that the reported data would occur given that there is no relationship. However, generally investigators do not want to know this probability. Instead, it would be much more useful to know the probability that there is no relationship, given the reported data. The probability that a researcher’s null hypothesis is false, given some set of data, may be quite different from the probability that these data would occur, given the null
hypothesis. Ottenbacher (1989) pointed out that this error results from a failure to consider Type II errors. Type II errors result from failing to reject a null hypothesis even though it is true. The probability to Type I errors, which is controlled in conventional significance testing, does not imply that the probability of Type II errors is controlled at similar levels. Cohen (1994) noted in this regard that significance testing “does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does”.

A fourth criticism has to do with the a priori selection of an alpha (α) level against which the probability level for each test statistic is to be compared. Whereas conventional choices of alpha levels, such as 0.5, are commonly used in an effort to balance the application of a study’s power toward avoiding both Type I and Type II errors (Olejnik 1984; Ottenbacher 1989). Most authors make little effort to actually assess the power as irrelevant (Rosenthal 1979). Moreover, the selection of a specific alpha level imposes an artificial dichotomy on a static (p) that is continuous (Kirk 1996; Thompson 1997; Young 1993). The practical difference between calculated probabilities of .049 as opposed to one of .051 is certainly not as dramatic as the dichotomous decision that only the former result is statistically significant, with all that it may imply, whereas the other is not. Frustration with this arbitrary dichotomy may encourage authors to refer to some results as “nearly significant” or “approaching significance.”
A fifth criticism of the use of significance testing involves misuse of the results. Perhaps as a consequence of the combination of previously cited criticisms, some authors seem to associate significance testing with replicability or reliability (Schmidt & Hunter 1995; Thompson 1996; 1997; Vacha Haase & Nilsson 1998), which leads to the assumption that a p value of .001, for example, is somehow more important or more impressive than a p value of .05. Certainly, p values are not a measure of the likelihood that a given result will be replicated (Cohen 1994). Nevertheless, getting a very small p value often leads to the potentially misleading description of a result as “highly” significant or as evidence of a “strong effect”, in spite of the fact that p level does not imply the strength of the relationship (Friedman 1968; Vacha-Haase & Thompson 1998).

Schmidt and Hunter (1995) also cautioned against another all too common error when reporting nonsignificant results which is also related to the magnitude of the p value. That is, some authors infer incorrectly that nonsignificance implies that there is no effect. Clearly, a nonsignificant result only indicates that the data being tested do not provide adequate evidence to reject the null hypothesis, given a particular alpha level. The nonsignificant result does not demonstrate that the null hypothesis is true.

Two additional, somewhat more technical criticisms were raised by Thompson (1993). The first criticism involves hierarchical testing within the same data set. Given the recommendation that higher order interactions should be tested in factorial ANOVA studies before main effects (Keppel 1991), Thompson has
reminded that each of these tests may represent very different distributions of the samples size across means. These differences could result in very different power to detect differences for each test. Thus, whatever a significant result may mean in an omnibus test that includes the entire sample means, it may mean something very different for a main effect or for some other specific comparison in the same data.

The second technical concern of Thompson (1993) about significance testing relates to the relationship between the sample size and the assumptions on which significance testing is based. For example, ANOVA assumes homogeneity of variances, and ANCOVA additionally assumes homogeneity of regression. In testing these assumptions, investigators conduct significance tests in hopes of not rejecting the null hypothesis. Ironically, the same large sample size that provides power against Type II errors will also increase the likelihood that the null hypothesis rejected, making the use of those significance tests more questionable.

Due to criticisms noted previously, some authors have recommended the complete elimination of significance testing (Morse 1998; Schmidt & Hunter 1995). However, most have taken the more moderate view that significance testing should be supplemented with or placed in the context of additional information.

In 1978, Carver noted that all of the criticisms of tests of statistical significance appeared to have had little effect. The situation has not changed since then. A quick perusal of educational research journals, educational and psychological statistics textbooks, and doctoral dissertations will confirm that tests
of statistical significance continue to dominate the interpretation of quantitative data in educational research. Surely one characteristics of statistical significance testing is that it is an enduring – in the face of the devastating criticism, perhaps it would be better to say, relentless-phenomenon in educational and psychological research. However, Carver (1993) offered four ways to minimize not to eliminate the importance of statistical significance testing: (a) insist on the word statistically being placed in front of significance testing, (b) insist that the results always be interpreted with respect to the data first, and statistical significance second, (c) insist on considering effect sizes (whether significant or not),and (d) require journal editors to publicize their views on the issue of statistical significance testing prior to their selection as editors.

But this is not the entire of story. By the mid-1970, apart from excessive emphasizing on significance testing as well as their misunderstanding and misusing, the behavioral and social sciences were in serious trouble. Large number of studies had accumulated on many questions that were important to theory development or social policy decisions. But results of different studies on the same question were statistically conflicting. On the other hand, some researchers conclude that more research is needed to identify the supposed interactions (moderators) that have caused the conflicting findings. After that, a large number of research studies were funded and conducted to test moderator hypothesis. When they were completed, there is now a large body of studies, but instead of being resolved, the numbers of conflicts increase. This may be the reason for
social science research, in general, for having attracted less funding in comparison to research in science. There is a chronic pessimistic feeling in the social and behavioral sciences that, when compared to the natural sciences, our progress has been exceedingly slow, if indeed there has been any progress at all. Two sources of pessimism in the social sciences and behavioral sciences are poor cumulation and small effect. Poor cumulation refers to the observation that the social sciences do not show the orderly progress and development shown by such older sciences as physics and chemistry. Poor cumulation does not seem to be due primarily to lack of replication or failure to recognize the need for replication. Indeed, the calls for further research with which researchers so frequently end their articles are carried wherever scholarly journals are read. It seems rather that it has been better at issuing such calls that at knowing what to do with the answers. There are many areas of the social sciences for which the result of many studies all addressing essentially the same question. Summaries of the results of these sets of studies, however, have not been nearly as informative as they might have been, with respect to summarized significance levels.

The second source of pessimism in the social sciences is the problem of small effects. Even when researchers seem to come up with a possibly replicable result, the practical magnitude of the effect is almost always small, i.e., accounts for only a trivial proportion of the variance. Thus, the complaint goes, even if some social action program works, or if some new teaching approach works, or if
psychotherapy works, the size of the effect is likely to be so small that it is of no practical consequence whatever.

1.1.2. Recommended changes in practice

At this point, research sponsors, governments, officials and the public become cynical. Even in some cases social and behavioral scientists themselves became cynical about the value of their own work, so they publish articles expressing doubts about whether behavioral and social science research is capable, in principle, of developing cumulative knowledge and providing general answers to socially important questions. In addition to the criticisms of statistical significance testing have led quantitative researchers to explore other approaches for making quantitative sense out of data because as said previously the rejection of the null hypothesis by itself is not very informative.

In an effort to quell the criticisms leveled at the social sciences, various actions have been developed in the hope that they would provide more surety in drawing conclusions. The long standing debate over statistical tests has swayed the editorial boards of leading research journals to encourage authors to supplement their statistical tests with “simple, flexible, and graphical techniques” aimed at “understanding the set of data in hand” (Cohen, 1994, p. 1001). An often-recommended technique is the use of effect sizes to describe the practical significance of a statistical test result, independent of the sample size and the measurement scale (Vaske, Gliner & Morgan, 2002). Effect size is simply a way of quantifying the effectiveness of a particular intervention, relative to some
comparison. It is easy to calculate, readily understood and can be applied to any measured outcome in Education or Social Science. It allows us to move beyond the simplistic, ‘Does it work or not?’ to the far more sophisticated, ‘How well does it work in a range of contexts?’ Moreover, by placing the emphasis on the most important aspect of an intervention – the size of the effect – rather than its statistical significance (which conflates effect size and sample size), it promotes a more scientific approach to the accumulation of knowledge. For these reasons, effect size is an important tool in reporting and interpreting effectiveness. As effect size measures are the common currency of meta-analysis which is main part of this study, it is explained in detail at the related part.

As previously described the most important problem in behavioral and social sciences is the failure to produce cumulative knowledge or poor cumulation and small effect. In one research area after another, researchers have conducted numerous studies, and in almost every case it has been found that different studies give different results. Some studies report significant findings, some do not. Some studies support the hypothesis, some do not. This situation has led many to conclude that cumulative knowledge and general principles and theories may be impossible to establish in the social sciences. The most important recent development is, therefore, the development of quantitative methods which solve this problem. These methods lift the information processing burden from the reviewer by quantitatively integrating findings across studies while simultaneously correcting for the effects of statistical and measurement artifacts which distort
study findings. Various methods for aggregating data across studies have been developed in the hope that aggregate data analysis would provide the social sciences more surety in drawing conclusions. Many of the earliest methods of aggregation were based on literature reviews or narrative subjective review. Conclusions were drawn based on the reviewers’ overall perceptions of what each study added to the current knowledge in the area. However, such qualitative analyses left many unanswered questions because of the potential for bias (Mason, 2003). In many research literatures there were not only conflicting findings, there were also large numbers of studies. This combination made the standard narrative subjective review a nearly impossible task—one shown by research on human information processing to be far beyond human capabilities. How does one sit down and make sense of, say, 210 conflicting studies? (Hunter & Schmidt, 2004).

1.1.3. Meta-analysis as a technique for integrating research findings

Starting in the late 1970s, new methods of combining findings across studies on the same subject were developed. These methods were referred to collectively as META-ANALYSIS a term coined by Glass (1976). Gene Glass (1976) presented what he called “meta-analysis” as a way to combine the results of multiple studies in a quantitative way. He and a colleague analyzed over 400 studies designed to assess the effectiveness of psychotherapy. He was able to show that, on average, across a large number of studies, therapy made a significant difference in the client outcomes. Glass (1976) provided this example to show
how meta-analysis could be used to compute an average effect size across studies. Glass also demonstrated that such averaged effect sizes could be used to find conclusions among opposing findings. Prior to meta-analysis, most methods for summarizing studies failed to incorporate the effect-size statistics and instead simply summarized the findings on a categorical basis (i.e., significant vs. not). Traditional methods of literature review focus on statistical significance testing, which is problematic because significance testing is highly dependent on sample size. Meta-analysis changes the focus from significance to the direction and magnitude of the effects across studies.

An effect-size statistic is the index used to represent study findings in direction and magnitude (Lipsey and Wilson, 2001). Meta-analysis is essentially the survey research approach by which the effect size of the research studies is surveyed, weighted and compared.

Glass’s meta-analytic approach caught the eye of many psychologists and remains well cited in the social sciences. A quick search of a social science database revealed over 2200 published articles using or discussing meta-analysis published between 1981 and 2001. Of these, over 1800 have been published since 1995 to 2001 (Field, 2001). Clearly, the use of meta-analysis is still accelerating and tends to produce itself an enormous amount of research articles (Schulze, 2004). However, it must be added in this context that along with this rising interest in the development and applications of meta-analysis there have been several attempts to modify the approach, technique or alternatives to it and consequently
this question that which approach or technique is best, has been arisen (Glass, 2000). Approaches of meta-analysis have changed over the years and continue to evolve (e.g., Kalaian & Raudenbush, 1996).

1.1.4. Methods of Meta-analysis (approaches toward meta-analysis)

Meta-analysis approaches fall into three broad categories. The purely descriptive approaches (the Glass approaches and study effects meta-analysis approaches) paint a descriptive picture of what is in the research literature but do not attempt to analyze, correct for, or otherwise address any of the artifacts that distort study findings. Next are meta-analysis approaches that address only the artifact of sampling error. These include the homogeneity test-based approaches of Hedges and Olkin (1985) and Rosenthal and Rubin (1982). These approaches do not address the effects of artifacts other than sampling error. In particular, they do not address measurement error. Finally, there are meta-analysis approaches that address and correct for the effects of not only sampling error but also a variety of other artifacts that distort study results. These approaches estimate the results that would have been obtained had all the studies been conducted in a methodological unflawed manner. These approaches called psychometric meta-analysis approaches. Hunter & Schmidt (1990) approach; Callender and Osburn (1980) approach have made important contribution in psychometric approaches. All the approaches have now been labeled “meta-analysis” but each approach has its own specific idiosyncrasies. Several recent studies have compared theses approaches. Johnson, Mullen and Salas (1995) compared the Hedges-Olkin, Rosenthal-Rubin
and Hunter-Schmidt meta-analytic approaches by manipulating a single data set. Mason (2003) compared Vacha-Haase and Hunter & Schmidt approaches for meta-analyzing reliability coefficients. Schulze, (2004) in his book compared traditional approaches with modern one toward meta-analysis. Some of these studies have been reviewed in detail in chapter II. The approaches that have compared in this study are Glass approach (descriptive approach) and Hunter & Schmidt approach (psychometric approach). The reason for studying the Glass and Hunter & Schmidt approaches is that they are the two techniques most frequently used in applied psychology, education and related fields (Johnson et al, 1995 cited in Aguinis 2007). Moreover, APA Monitor reports that 600 to 800 meta-analysis have been done in the area of psychology since meta-analytic techniques was developed, with the primary methods used being the Hunter-Schmidt approach as psychometric approach and the Glass approach as classical approach (Adler, 1990). These approaches have applicability in many other areas of research including research in education (Hall, 1991).

1.1.5. Glass meta-analytical approach (Descriptive approach)

For Glass, the purpose of meta-analysis is descriptive, the goal is to paint a very general, broad, and inclusive picture of a particular research literature. Glassian meta-analysis often combines studies with somewhat different independent variables and different dependent variables. Glassian meta-analysis approach has three primary properties:
1. A strong emphasis on effect sizes rather than significant levels. Glass believed the purpose of research integration is more descriptive than inferential and that the most important descriptive statistics are those that indicate most clearly the magnitude of effects. Glassian meta-analysis typically employs estimates of the Pearson $r$ or estimates of $d$. The initial product of a Glassian meta-analysis is the mean and standard deviation of observed effect sizes or correlations across studies.

2. Glassian meta-analysis implicitly assumes that the observed variability in effect sizes is real and should have some substantive explanation. There is no attention to sampling error variance in the effect sizes. The substantive explanations are sought in the varying characteristics of the studies (e.g., sex or mean age of subjects, length of treatment, and more). Study characteristics that correlate with study effects are examined for their explanatory power. The general finding in applications of Glassian meta-analysis has been that few study characteristics correlate significantly with study outcomes.

3. A strongly empirical approach to determining which aspects of studies should be coded and tested for possible association with study outcomes. Glass de-emphasized the role of theory in determining which variables should be tested as potential moderators of study outcome (Hunter & Schmidt, 1999).
1.1.6. Hunter & Schmidt meta-analytical approach (psychometric approach)

Hunter and Schmidt advocate a single approach (a random effect approach) based on their belief that fixed–effect models are inappropriate for real-world data and the type of inferences that researchers usually want to make (Field, 2001). As previously mentioned meta-analysis is used as a way of trying to ascertain the true effect sizes by combining effect sizes from individual studies. There are two way to conceptualize this process: fixed effects and random effect models. In essence, in the fixed effect conceptualization, the effect size in the population is assumed to be the same for all studies included in a meta-analysis (Field, 2001). The alternative possibility is that the population effect sizes vary randomly from study to study. To summarize, in the random effect model studies in the meta-analysis are assumed to be only a sample of all possible studies that could be done on a given topic whereas in the fixed effect model the studies in meta-analysis are assumed to constitute the entire universe of studies (Field, 2001).

In its fullest form, Hunter and Schmidt's approach to meta-analysis combines some of the best features of other approaches. All studies related to a given topic are gathered, regardless of quality. The distribution of effect sizes is corrected for sampling error, measurement error, range restriction, and other systematic artifacts. If the remaining variance is still large, effect sizes are grouped into subsets according to preselected study features, and each subset is meta-analyzed separately. Ideally, the meta-analysis should estimate true treatment effects under conditions typical of those represented in the studies and predict treatment effects
under conditions determined by the reviewer. Unfortunately, this technique requires substantial information from individual studies for accurate correction of effect sizes. This information is not always available in research reports. This approach is the only meta-analysis approach that takes into account both statistical and measurement artifacts (Bangert-Drowns, Robert L. & Rudner, Lawrence M. (1991).

1.1.7. Glass meta-analytical approach vs. Hunter& Schmidt meta-analytical approach

Glass technique and the Hunter& Schmidt technique are similar in many ways but they differ in several key ways. They are similar in that they recommend using every available study, published or unpublished, in meta-analysis (Hall, 1991). However, they differ in three specific areas: effect size formula, correction for sampling error, and correction for measurement error in the dependent variable (Hall, 1991). A brief discussion of the Glass and the Hunter-Schmidt techniques for each of these areas are as follows.

1.1.7.1. Effect Size

Both the Glass and Hunter-Schmidt techniques calculate an effect size. The effect size measures the average performance of the experimental group in relation to the control group. The effect size is calculated by subtracting the mean of the control group from the mean of the experimental group divided by the standard deviation. Glass proposes using the control group standard deviation because it is unaffected by the treatment (Glass, 1981). On the other hand, Hunter- Schmidt
propose using the pooled within group standard deviation because it has only half the error of the control group standard deviation (Hunter and Schmidt, 1990). In both approaches it is the overall mean effect size which is published as representing the size of the effect.

1.1.7.2. Correction for Sampling Error

In addition to calculating an effect size, Hunter and Schmidt recommend testing the variance of the overall mean effect size for sampling error. This is accomplished by calculating the overall mean effect size error variance and dividing it by the variance of the overall mean effect size. The hypothesis tests whether or not the ratio of the error variance is error variance to the variance is .75 or greater. If so, then it is assumed that the rest of the variation between study effect sizes is due to other types of error (Hunter and Schmidt, 1990). If, however, the ratio is less than .75, then further analysis is recommended by the Hunter and Schmidt to determine variables within the studies that are causing the effect sizes to differ significantly from each other. These variables such as sample variables, design characteristics are called as moderator variables. Hunter and Schmidt recommend using Pearson correlation to determine the relationship between the study effect sizes and various study characteristics that assumed to be moderator variables (Hunter and Schmidt, 1990). Glass recommends using Pearson correlations, ANOVAs, or regression analysis to locate moderator variables but does not recommend testing for sampling error (Glass, 1981).
The 75% Rule. A procedure unique to the Hunter and Schmidt approach is the so-called 75%-rule originally proposed by Schmidt and Hunter (1997). The reasoning behind this rule is as follows. The development of the Hunter and Schmidt approach was done with validity coefficients as the main effect size of interest and personnel selection as the most important field of application in mind. Validity coefficients are supposed to be influenced by a series of mainly methodological factors of which many can in principle be corrected for. However, in the most application of meta-analysis all the information necessary to correct for the artifactual factors is not available so that variance in observed effect sizes due to uncorrected artifactual influences is always presumed to remain. The component supposed to account for the largest amount of observed variance is sampling error. If observed variance is larger than expected by sampling error, and then there may be variance in effect sizes left to be explained (Schulze, 2004).

**1.1.7.3. Correction for Measurement Error**

Hunter and Schmidt (1990) believe that measurement error can affect the overall mean effect size. They state that measurement error inflates the standard deviation and thus lowers the value of the effect size. To correct this, they recommend dividing the effect size by the square root of the reliability coefficient of the dependent variable measures. The correction should increase the value of the effect size. This is especially true in the case where the overall mean effect size is not significantly different from zero, but after the correction, become significant. Glass does not include any correction formulas for measurement error.
The purpose of this study is comparing these two well noticed meta-analytical approaches based on the key differences mentioned above with single data set. Creativity has been selected for application meta-analysis to investigate instructional programs which might have influence on creativity.

1.2. Need of the Study

Today meta-analysis is an increasingly popular research technique. Many discoveries and advances in cumulative knowledge are being made not only by those who do primary research studies, but also by those who use meta-analysis to discover the latent meaning of existing research literature.

As mentioned before, meta-analysis can handle the synthesis of a large number of studies, which may not be feasible in a conventional literature review. As a result, meta-analysis is particularly useful when research in a field produces confounding or vague results (Wolf, 1986), or when there are large numbers of studies on the topic of interest. By facilitating generalization of the knowledge gained through individual studies, meta-analysis promotes replication of research. The literature on educational studies is diverse in both its foci and its methods. This diversity is largely due to the importance of educational settings in the development of human thought. Meta analysis can help to organize this diverse literature, thereby making sense of the findings. Additionally, meta-analysis help to overcome the problem of small sample size that plagues many areas of educational psychology, such as gifted and creative programming (Asher, 2003 as cited in O’mara, et al, 2005). In general meta-analysis can make a significant
contribution to educational research. From the perspective of output, meta-analysis has the ability to produce concise, easily interpretable findings regarding conceptual issues, program effectiveness, and research design strategies. That is, a single meta-analysis has the capacity to inform and appeal to educational theoreticians, practitioners, policymakers, and researchers alike. This in turn can promote a consonance within the (rather segregated) educational community, potentially leading to more effective, concentrated efforts. For these reasons, and also for the overall value of meta-analysis discussed above, meta-analysis can make a significant contribution to educational research.

It seems that meta-analysts often choose a specific meta-analytic technique based on habit, the availability and their familiarity with a specific software package or usage trends in specific topic areas rather than the relative merits of the available approaches (Aguinis, Sturman, and Pierce, 2008). The fact that meta-analysts do not often justify or explain their choice for a specific meta-analytic approach may in part be because of the lack of the guidelines in the literature regarding which strategies to use under which conditions, which technique produces the most accurate estimates of moderating effect magnitude (Aguinis, Sturman, and Pierce, 2008). So far, five most highly cited meta-analyses that have been published in review of educational research (according to ISI web of science) are: classroom structure, reinforcement and motivation, literacy, sex-differences, intellectual abilities and school learning (O’mar, 2005).
It seems that in the field of creativity as an important area in education there are no or few meta-analytic researches in Iran (setting of the study). On the other hand, large number of studies incorporated in this study had conflicts with each other in terms of their results in impact of instructional programs on creativity. We need to draw together and analyze the findings from many studies to determine if instructional programs deliver a profound impact on creative ability in general. The scope has been limited to effectiveness of instructional programs in fostering creativity in general as measured by Torrance Thinking Creativity Test.

The main purpose of the present study is to introduce most appropriate meta-analytic approach in educational researches comparing Glass (1985) approach and Hunter & Schmidt (2004) approach based on the selected criteria.

1.3. Area for application of meta-analysis approaches

In addition to, creativity has been selected as area for application of meta-analysis approaches. Creativity is an integral part of any understanding of human education and psychology. In the last 50 years, researchers have extolled the virtues of creativity regarding the intellectual, educational, and talent development of children both generally and in specific content areas. Moreover, educators, and psychologists from diverse specialties have noted creativity’s contributions in areas as diverse as workplace leadership, adult vocational and life success, healthy psychological functioning, coping, and emotional growth, maintenance of healthy and loving relationship. Even more on this, educational and school psychologists, have called for explorations of the role of creativity can play in reducing violence
and promoting conflict resolution. These are but a few of the areas in which creativity is being applied, but they provide a sense of the extraordinary breadth of applications that educational psychologists may pursue.

However, there appears to be no shortage of area in which creativity can be applied to improve people’s lives. Although the potential applications of creativity are well documented, observers over the past several decades, from Guilford (1950) to Sternberg and Lubart (1999), have noted that this potential is rarely fulfilled. Classrooms generally do not appear to be creativity-fostering places, primarily due to the biases of researchers, their negative assumptions, and contrary research findings which creates an atmosphere that severely restricts ability and desire to apply creativity. Hence, researchers are now required to report effect sizes and in some cases confidence intervals to overcome this problem. This meta-analysis synthesizes results of the studies on creativity in the setting of study (Iran) which most probably has not been done yet.

1.4. Statement of the Problem

This study was undertaken a meta-analytical attempt to integrate research results regarding effectiveness of similar instructional programs on creativity in general and its specific components. Two most popular meta-analysis approaches in social sciences were applied for integrating research results as well as calculating effect sizes to find out if they made statistically any differences. This study entitled:
“Comparison of Two Meta-Analytical Approaches in generalization of the effectiveness of instructional programs on development of creativity in general and its components in particular”

1.5. Objectives

The present study focused on the following objectives:

➢ Primary or main objectives;

1. To calculate mean effect sizes of the studies using the Glass meta-analysis approach and Hunter-Schmidt meta-analysis approach before correction for measurement error and to apply Cohen’s power table to classify and compare them.

2. To compare the overall mean effect sizes calculated using the Glass meta-analysis approach and Hunter-Schmidt meta-analysis approach before correction for measurement error.

The objective was examined with respect to same data set on:

• Studies on effectiveness of instructional programs on developing creativity in general,

• Studies on effectiveness of instructional programs on developing special components of creativity in particular.

3. To calculate mean effect sizes of the studies using the Glass meta-analysis approach and Hunter-Schmidt meta-analysis approach after correction for measurement error and to apply Cohen’s power table to classify and compare them.
4. To compare the overall mean effect sizes calculated using the Glass meta-analysis approach and Hunter-Schmidt meta-analysis approach after correction for measurement error.

The objective was investigated with respect to same data set on:

- Studies on effectiveness of instructional programs on developing creativity in general,
- Studies on effectiveness of instructional programs on developing special components of creativity in particular.

➢ Secondary objective:

- To find out the effect of Instructional programs on creativity -in general- across the studies using Comprehensive Meta-analysis Software (CMA).

1.6. Research questions

Four research questions were formulated for the present study as follow:

1. Is there any difference in the classifications made using Cohen’s power table of the effect sizes calculated using Glass and Hunter-Schmidt approaches before correction for measurement error?

2. Is there any difference in the classifications made using Cohen’s power table of the effect sizes calculated using Glass and Hunter-Schmidt approaches after correction for measurement error?

3. How does the Glass meta-analytic approach significantly differ from the Hunter & Schmidt meta-analytic approach when applied to same data set?
4. Do instructional programs enhance individual creativity? If so, how powerful is the size of effect of instruction for improving creativity?

1.7. Hypotheses

In the line with the objectives stated above, three hypotheses were formulated for this study. They are as follows:

**H1**: The overall mean effect sizes calculated using the pooled within group standard deviation in a Hunter & Schmidt meta-analysis approach (without correction for measurement error) will not significantly differ from that in a Glass meta-analysis approach which uses the control group standard deviation.

**H2**: The overall mean effect sizes will be significantly different in a Hunter & Schmidt meta-analysis approach (with correction for measurement error) than those in a Glass meta-analysis approach.

**H3**: The effect sizes indicating the effect of instructional programs on improving creativity in general will not be significantly different from zero in size when estimates are made using Comprehensive Meta-Analysis Software.

1.8. Definition terms of the study

**Meta-Analysis**: The summarization of empirical studies using quantitative methods; the effect size results form empirical studies are collected and a summary statistics is calculated.

**Effect size**: A point value that indicates the strength and direction of the relationship of interest in the research study. The effect size is what makes meta-analysis possible because it is a “statistical standardization of the study findings
such that the resulting numerical values are interpretable in a consistent fashion across all the variables and measures.” There are many different effect sizes, which are dependent on the statistical methods employed in the research.

**Standardized effect size:** In contrast to a raw effect size, which indicates the strength and direction of a relationship in the units of the scale that the relationship was measured, the standardized effect size is independent of the scale used to measure the relationship. An example of a raw effect size is simply the difference between two means. In contrast an example of a standardized effect size is the difference between two means divided by the standard deviation.

**Fixed- effect models:** Models that assume there is a “true” effect size that underlies all of the studies and that they differ among each other only because of sampling error (Streiner, 2003).

**Random- effect models:** make the assumption that there is a population effect sizes, from which the studies in the meta-analysis are a random sample. The reason that this distinction is important is that, in many situations, the two types of analysis yield different results (Streiner, 2003).

**Publication Bias:** Publication bias is defined as a bias against negative findings on the part those involved in deciding whether to publish a study. Given the apparent prevalence of positive findings in the published literature, meta-analyses are more likely to identify and include studies with positive outcomes, thereby, potentially skewing meta-analytic findings towards a positive mean effect size.
**Instructional Program**: Although instruction is a broad concept, however, in this study it means a set of procedures and materials (teaching approach) organized around specific educational objectives which lead to increase student creative performance.

**Creativity**: Creativity is the ability to produce work that is both novel (e.g. original, unexpected, and statistically infrequent) and appropriate. This study focused only on studies that utilized the TORRANCE THINKING CREATIVITY TEST. Thus, Creativity in general and its special components in particular were limited to Torrance Thinking Creativity Test.

**Measurement Error**: The extent to which there are discrepancies between survey results and the true value of what the survey researcher is attempting to measure. There are several possible sources of error here. Respondents may report inaccurate information because they do not have the required information, due to carelessness, or because they do not understand the question asked. Alternately, respondents may provide accurate information, but errors are introduced in the data processing stage due to keypunching, coding, or programming errors. Since it is often not possible to determine the "true value" of what one is trying to measure, precise estimates of measurement error are usually not possible. However, techniques exist for obtaining some information about the likely extent of measurement error.

**Measurement Error Correction**: Measurement error inflates the standard deviation and thus lowers the value of the effect size. To avoid the deflated effect
size, Hunter & Schmidt recommend dividing the effect size by the square root of the reliability coefficient of the dependent variable measure. In this study the related formula for correction was used as below:

\[
\text{Measurement error correction} = \frac{ES}{\sqrt{r}}
\]

**Artifact (error):** Hunter and Schmidt described a variety of adjustment to individual effect sizes that meta-analysis may wish to make. These include adjustment for the unreliability of the variable(s) involved in the effect size, for restriction in the range of those variables, for dichotomization of continuous variables, and for imperfect construct validity. The objective is to permit meta-analyst to come as close as possible to estimating the magnitude of the relationship represented in an effect size as it would appear under ideal research circumstances. Of the Hunter and Schmidt adjustments, the one most likely to be useful and usable is the correction for attenuation due to unreliability of the variables used in the effect size. This adjustment is applicable to the standardized mean difference and correlation coefficient effect sizes.

**1.9. Delimitations of the study**

The scope of the study is limited to all the studies done in Iran up to 2008 which are documented in Iranian Universities Digital Libraries and Information and Documentation Center under headings ‘Journal articles, Dissertations, and Research projects’. However, dissertations were both at master and doctoral level. The study is further delimitated to those studies exploring the relationship...
between instructional programs and development of creativity through experimental approach.