An Empirical Study of Bonett's (2009) Meta-Analytic Model

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Aolin Xie

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor
Michael Harwell

October 2012

Acknowledgements

I am indebted to many people for reaching this point.

First I would like to thank my advisor Dr. Harwell. He guided me through the whole graduate study and provided me with great research opportunities. I am fortunate to have such a dedicated, enthusiastic, and patient advisor.

I would like to thank Dr. Michael Rodriguez, Dr. Ernest Davenport, and Dr. Frances Lawrenz. Thank you for serving as my committee members and providing helpful feedback on the thesis.

I'm grateful to Dr. Sheryl Larson and Dr. Brian Abery from Institute of Community Integration, University of Minnesota. They provided me with an excellent assistantship/internship which supported me through my doctoral study. It's one of the best experiences for me to learn and practice my skills during graduate school.

I am very fortunate to have worked with my former colleagues and friends Timothy Ryan and Debora Maruska from Minnesota Department of Human Services. Their kind help made my life and work in Minnesota much easier and more pleasant.

I would also like to thank my mentor Dr. Troy Chen from ATC. Inc. Troy is a very nice and caring mentor who gives me valuable suggestions about my career choice and development.

I thank Dr. Yuanjia Wang from Columbia University. She provided a lot of guidance, help, and internship opportunities during my graduate study.

Last, I would like to thank my parents Caiqing Xie and Wei Lu, and my husband LiDong Pan for their unconditional love and support.

Dedication

This dissertation is dedicated to my parents Caiqing Xie and Wei Lu, and my grandparents Yehua Zhan and Liang Lu.

Abstract

A key methodological decision in a meta-analysis has traditionally been the choice between the classic fixed-effects (FE) or random-effects (RE) models assumed to underlie effect sizes (see Hedges & Olkin, 1985). Recent work has criticized these models because of the implausibility of their underlying assumptions (Bonett, 2008, 2009; Hunter & Schmidt, 2000). Bonett (2009a) proposed a modified FE model and recommended using contrasts to compare mean effect sizes among levels of discrete moderators. This study empirically investigated the behavior of the Bonett (2009a) modified FE model and the classic FE model for interval estimation and hypothesis testing of effect size contrasts. The results suggested that the two models performed similarly well with normally distributed data. The Bonett model was robust to nonnormality combined with unequal within-study variances and unequal within-study sample sizes, whereas the classic FE model showed inflated type I error rates and lower statistical power under these conditions.

Table of Contents

List of Tables

List of Figures

Introduction

Statement of the Problem

Meta-analysis is one of the most frequently used and increasingly important methods to synthesize research across empirical studies and obtain accumulated knowledge (Suri, 2000; Lipsey & Wilson, 2001; Hunter & Schmidt, 2004). Meta-analysis provides effect sizes as "a measure of the magnitude of the strength of a relationship between independent and dependent variables" (Dunst, Hamby & Trivette, 2004), which allows cross-study comparisons by putting research findings on a common and usually standardized metric.

Under the two-group (experimental and control) conditions, standardized mean difference serves as the point estimate of effect size. The average effect of a treatment can be estimated by combining effect sizes from individual studies. Comparing mean effect sizes for levels of key (discrete) moderator variables provides information about the role and impact of the moderator. For example, in the Hyde et al. (1990) meta-analysis of gender differences in mathematics achievement, student ethnicity served as a key moderator. Ethnicity had four levels (Black, White, Hispanic, Asian), and to examine whether the same mean gender differences in mathematics achievement were observed for Black and White students a simple contrast could be estimated and tested under a classic fixed-effect (FE) approach. To examine whether White and non-White students differ in mean effect sizes, a complex contrast is needed. This complex contrast involves averaging the effect sizes from studies of non-White students (Black, Hispanic, Asian), and comparing this average effect size with that obtained from studies that included White students.

The classic FE model assumes that k studies have been deliberately selected, that population effect sizes are the same to allow average effect sizes to be estimated

and tested meaningfully, and that inferences are restricted to the sampled studies. Fixed-effect assumptions and models are widely used in educational and psychological research. However, the classic FE model has been criticized for various shortcomings (Bonett, 2008; Hunter & Schmidt, 2000; National Research Council, 1992; Schmidt, Oh, & Hayes, 2009) that include producing inflated Type I error rates, narrow confidence intervals, biased results when the k sample sizes are unequal and k population effect sizes differ, and the unrealistic assumption of equal variances within studies.

The criticisms of the classic FE model prompted calls for increasing the use of the random-effect (RE) models in meta-analysis. The RE model assumes that k studies have been randomly sampled from a clearly defined superpopulation of studies which allows unconditional inferences. However, RE models have also been criticized for shortcomings such as that the random sampling assumption of RE model is unlikely to be achieved in reality (Bonett, 2008, 2009; Hedges & Vevea, 1998), biased inferences when the assumption of equal within-study variances is violated, and biased confidence intervals for variance components when the assumption of normality is not satisfied. Both the classic FE and RE models have also been criticized for their use of empirically based weights based on variances of effect sizes, which tends to produce biased estimates since the weights and individual study effect sizes are expected to be correlated (Shuster, 2009).

Bonett (2009a) criticized the classic FE model, arguing that key underlying assumptions of homogeneity of variance within studies and equality of population effect sizes across studies were implausible. Bonett (2009) proposed a modified FE model which does not require these assumptions and made a compelling argument

that estimating and testing contrasts under this model can be a powerful explanatory tool.

Bonett (2009a) conducted Monte Carlo studies of the ability of his proposed model to control Type I error rates and produce narrow confidence intervals about a contrast for various sample sizes and patterns of population effect sizes. Bonett reported that his model outperformed the FE model with a coverage probability of .95 across conditions, whereas the probability was sometimes far below .95 for the FE model. Also, the Bonett model had better precision than the classic RE model by providing slightly narrower average confidence interval widths.

However, the evidence that the Bonett model performed well was limited to normally-distributed and homoscedastic data for a limited number of within-study sample sizes and numbers of studies. The meta-analytic literature has demonstrated the importance of studying realistic data conditions including the number of effect sizes, the average sample size per effect size, effect size magnitude, level of heterogeneity of effect sizes (Cafri, 2010; Johnson, 1995; Nortgate & Onghena, 2003), and non-normal raw data distributions (Harwell, 1997; Steel et al., 2002). The Bonett model should be examined under these data conditions.

Purpose of the Study

With little knowledge about the behavior of the Bonett (2009) model compared to the classic FE model under realistic data conditions in meta-analysis, a substantial gap in the literature exists. This study is an attempt to investigate the behavior of the Bonett modified FE model for interval estimation of contrasts of mean effect sizes and that of the classic FE model (Hedges & Olkin, 1985) to provide a basis of recommending a model to researchers.

Literature Review

Meta-analysis

A major goal in science is to provide cumulative evidence. This goal can be achieved by synthesizing research across empirical studies to obtain accumulated knowledge. Research synthesis translates the accumulated knowledge and provides direction for further research, practice and policies. One of the most frequently used and increasingly important methods of research synthesis in education is meta-analysis (Suri, 2000, Lipsey & Wilson, 2001; Hunter & Schmidt, 2004).

History. The interest in research synthesis started to grow after Fisher (1932) first proposed a method to combine p-values for the same hypothesis test from a set of independent tests (Chalmers, Hedges & Cooper, 2002). The term meta-analysis was introduced by Gene Glass to the American Educational Research Association in 1976 (Oswald, 1999) as a statistical method to integrate, review, and summarize empirical research findings in various areas. According to Glass et al. (1981), meta-analysis provides a quantitative summary of research findings and seeks to generate conclusions. It does not prejudge which study to include or exclude.

Abrami et al. (1988) suggested five purposes of meta-analysis: (a) to summarize the relationship between two constructs, (b) to examine how factors explain variability in the main relationship of interest, (c) to provide directions for future research, (d) to generate new theoretical perspectives (subject to empirical verification), and (e) to suggest future applications of the findings.

Advantages and limitations. There are several advantages of meta-analysis over individual studies. As a quantitative literature review, meta-analysis integrates results under a common metric, such as the standardized mean difference or the Pearson correlation coefficient (Kisamore, 2003; Rosenthal et al., 2006). The use of a common metric allows comparisons among individual studies conducted in various

settings with different measures (Hunter & Schmidt, 2004; Lipsey &Wilson, 2001). Meta-analysis can handle a large number of studies which are otherwise overwhelming, and helps to avoid over-interpreting differences across studies (Banda & Therrien, 2008; Lipsey & Wilson, 2001).

Cooper and Hedges (1994) concluded that meta-analysis provided more precise and unbiased relationship estimates than a single study. They also pointed out that the total sample size in meta-analysis was increased by integrating several studies, thus, the statistical power was larger which allows for more accurate and reliable statistical analysis and hypothesis testing. Besides, weights can be assigned to studies with better quality, leading to less biased estimation than an individual study.

Generally speaking, meta-analysis has more statistical power than a primary study (Matt & Cook, 1994, p. 510; Murlow, 1995, p. 4) since it reduces the standard error of the weighted average effect size to create a narrower confidence interval which, other things being equal, increases the chance of detecting nonzero population effects (Cohen & Becker, 2003). Thus, meta-analysis can reveal significant results by aggregating across studies, including those with small to moderate sample sizes that may be underpowered (Quintana & Minami, 2006). This feature allows researchers to detect the effects of interventions with more confidence in social science fields in which the population is diverse and the statistical power is usually low due to relatively small sample sizes.

Besides improving the quality of research conclusions by aggregating results across studies from different settings, meta-analysis also provides generalizable estimation of the true relationship in the population (Cooper & Hedges, 1994). Meta-analysis has good external validity since the effects can be generalized across a wide range of samples, settings, and interventions (Quintana & Minami, 2006).

Another advantage of meta-analysis is that moderator analysis allows for the identification of moderators and the investigation of how they are related to effect sizes. Since the 1980s, detecting and explaining the variance in effect sizes is considered as important as examining the mean effect size (Hunter & Schmidt, 1990). This kind of analysis is particularly important and useful since research findings are usually inconsistent, and in need of identifying the factors that explain variability in the main relationship of interest, such as sampling error and study characteristics. Also, theory building increasingly relies on analyzing moderators in meta-analysis (Miller & Pollock, 1994; Mullen, Salas, & Miller, 1991).

Despite the above mentioned advantages, meta-analysis has been criticized for several limitations. First, meta-analysis often includes studies with both good and poor designs which is sometimes referred to as "mixing apples and oranges". Mixing studies with different qualities and measures may produce mean effect sizes that are meaningless (DeCoster, 2004; Lipsey & Wilson, 2001). Second, sampling bias and the file drawer problem will lead to biased estimation in meta-analysis. The file drawer problem means that studies with negative or non-significant results are unlikely to be reported or published (Rosenthal, 1993). Therefore, the selected studies in a meta-analysis may not represent the true population. Including only published results is likely to provide biased effect sizes estimations (Banda & Therrien, 2008; Lipsey & Wilson, 2001, Rosenthal, 1998). Third, the findings in meta-analysis can be subjective since the inclusion criterion for studies is decided by researchers (DeCoster, 2004). Other concerns raised include exclusion of qualitative studies (Eysenck, 1994), exaggeration of significance levels by including many studies, and violation of the independence assumption if subjects participate in more than one study (Rosenthal, 1998).

Reviewing real-world meta-analytic work, we see that the above mentioned concerns are not groundless; however, the problems may be due to either the limitations of meta-analysis or inappropriate practice. For example, Lipsy and Wilson (1993) examined 302 meta-analytic psychological, educational, and behavioral treatment studies. Among the over 300 meta-analyses published after 1980, only six of reported negative mean effect sizes. The distribution of the 302 mean effect sizes (standardized mean differences) was "overwhelmingly positive", with 90% of them greater than 0.1 and 85% of them greater than 0.2 (Lipsy & Wilson, 1993). These authors questioned the validity of the meta-analytic findings in which almost all examined treatments had positive effects.

Several possible explanations of this result were proposed by Lipsy and Wilson (1993). First, the primary studies with potentially poor design overestimated the effect which inflated the mean effect size in the meta-analysis. Second, published studies were over sampled in the typical meta-analysis and caused biased results (Lipsy & Wilson, 1993). This situation could be caused by deficiencies in practice rather than the meta-analytic technique itself, as Lipsy and Wilson (1993) noted.

Practical issues and remedies. Methodological and practical remedies can be used to overcome the shortcomings of meta-analysis. For example, to provide better interpretations, study quality and characteristics can be coded and analyzed in a moderator analysis to investigate whether these factors influence the meta-analytic findings. Incorporating unpublished reports and dissertations helps to reduce selection bias. With the development of technology and the internet, researchers can now access unpublished reports and dissertations easier than decades ago, which makes the inclusion of non-significant findings feasible.

Models have also been developed to describe the degree of publication bias, estimate the number of unpublished studies, and adjust for the bias (Sutton et al., 2000) in order to provide better interpretations of meta-analytic results. In short, meta-analytical statistical methods have developed rapidly and important progress has been made in recent years (Cook et al., 1992; Durlak & Lipsey, 1991; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Rosenthal, 1991a, Lipsey & Wilson, 2001).

Still, with the complexity of methodology issues there is an increasing need for rigorous and sophisticated methods to ensure the validity of research reviews and provide unbiased syntheses (Chalmers, Hedges & Cooper, 2002). Additional work providing guidance to meta-analyses is needed.

Despite the limitations, the advantages make meta-analysis an appealing method in social science and educational research. Meta-analysis appears to be the most effective way to combine the results from multiple studies. Since it can generate new theories from related studies, builds frameworks to explain phenomena (Hunter & Schmidt, 2004; Stanovich, 2004), and provide guidance that promote evidence-based practice (Jenson et al., 2007; Rosenthal et al., 2006), the academic recognition of meta-analysis is growing and this method has been well accepted among researchers and practitioners.

For example, a brief review by Jenson et al. (2007) of articles published for several years before 2002 showed that approximately 25% of the articles published in *Psychological Bulletin,* a major review journal in psychology, used a meta-analytic approach. Although developed primarily as a research tool in education and psychology, meta-analysis has been used in both basic and applied research and has become essential in fields such as medicine (Jenson et al., 2007). Other research areas

with increasing use of research synthesis and meta-analytic technique include advertising, agriculture, archaeology, astronomy, biology, chemistry, criminology, ecology, education, entomology, law, manufacturing, parapsychology, psychology, public policy, and zoology (Petticrew, 2001).

Effect Size

Definition. Defined as "a measure of the magnitude of the strength of a relationship between independent and dependent variables" (Dunst, Hamby & Trivette, 2004), effect sizes allow cross-study comparisons by putting research findings on a common and usually standardized metric. The standardized measures have significant advantages over the unstandardized ones since the standardization allows measures from different scales (sometimes with no intrinsic meanings, such as raw scores from math achievement tests) to be compared.

Effect size index. The choice of an effect size index is driven by considerations including: (a) effect sizes from different studies should be comparable in the sense of measuring the same thing; (b) they should be substantively interpretable and computable from information likely to be reported in studies; (c) effect size should have good technical properties such as known sampling distributions (Cooper & Hedges, 2009). More than one standardized index was developed to represent the magnitude and direction of the relationship examined. These indices generally fall into two categories: standardized differences between group means and correlation coefficients (Lipsey & Wilson, 2001). The commonly used effect size indices (standardized mean difference, log odds ratio and correlation) are readily converted from one to another to allow the comparison of studies reporting different types of effect sizes (Cooper & Hedges, 2009).

Interpretation. The effect sizes can be interpreted in different ways. For example, the standardized mean difference quantifies the degree of overlap between the distributions of observations in the experimental and control groups, when the observations are normally distributed (Glass, 1976). Converting the standardized mean difference to a correlation coefficient allows another interpretation of effect size: the squared correlation coefficient can be interpreted as the percentage of variance in the dependent variable that can be accounted for by the experimental treatment effect (Hedges & Olkin, 1985). Attempting to better describe the magnitude of effect sizes, Cohen (1988, 1992) developed guidelines for power analysis and sample size planning which is now widely used in various contexts. According to Cohen's standard, correlation coefficients (r) less than or equal to 0.10, or standardized mean difference (d) less than or equal to 0.2 are considered as small, $0.1<r\leq0.30$ or $0.2<d\leq0.5$ are viewed as medium effects, and $r\geq 0.50$ or $d\geq0.8$ are considered large.

The Classic Fixed-effects (FE) and Random-effects (RE) Models

Equations of the classic FE model. The classic FE and RE models for effect sizes of mean differences are described in several sources including Hedges and Olkin (1985), Hedges and Vevea (1998), Konstantopoulos and Hedges (2009), and Raudenbush (2009). The traditional Hedges and Olkin's meta-analytic approach typically involves two stages. At the first stage, the decision of whether or not the collected studies share a common population effect size needs to be made. If a test of the homogeneity of studies effect sizes is not statistically significant the FE model is usually used to combine mean effect sizes across studies. If the study effect sizes are not homogeneous the RE model is typically used. However, selecting between the FE

and the RE models is not a straightforward process and the test of homogeneity is not always recommended (Hardy, 1998; Shuster, 2009).

In the FE model the estimated effect size of study i (i=1…k) is typically calculated as the mean difference in treatment and control groups divided by the pooled standard deviation:

$$\hat{\delta}_i=(\hat{\mu}_{iT} - \hat{\mu}_{iC})/\hat{\sigma}_i \qquad (1)$$

where $\hat{\mu}_{iT}$ and $\hat{\mu}_{iC}$ are the estimated means for treatment and control groups, respectively, and $\hat{\sigma}_i$ is the estimated pooled standard deviation for the control and treatment groups in study i and is calculated as the square root of:

$$\hat{\sigma}_i^2 = [(n_{iT} -1)\hat{\sigma}_{iT}^2 + (n_{iC} -1)\hat{\sigma}_{iC}^2]/(n_{iT}+n_{iC}-2), \qquad (2)$$

where it is assumed that $\sigma_{iT}^2 = \sigma_{iC}^2 = \sigma_i^2$. In equation (2) n_{iT} and n_{iC} are the sample sizes in the treatment and control groups, respectively, of study i, and $\hat{\sigma}_{iT}^2$ and $\hat{\sigma}_{iC}^2$ are the estimated variances in the treatment and control groups, respectively. The variance of the effect size for study i with independent samples is

$$\text{var}(\hat{\delta}_i) = (n_{iT}+n_{iC})/(n_{iT}\times n_{iC})+ \hat{\delta}_i^2/[2(n_{iT}+n_{iC})] \qquad (3)$$

(Hedges & Olkin, 1985).

To compare mean effect sizes in the classic FE model among levels of a discrete moderator it is common to construct a 100(1-α)% confidence interval of the contrast that is calculated as

$$\sum_{j=1}^{m} c_j \hat{\bar{\delta}}_j \pm z_{\alpha/2} \left[\text{var}\left(\sum_{j=1}^{m} c_j \hat{\bar{\delta}}_j\right)\right]^{1/2}, \qquad (4)$$

where c_j is the contrast coefficient of the jth level of the discrete moderator (j = 1,2,…,m), and $\sum_{j=1}^{m} c_j=0$ and $\hat{\bar{\delta}}_j$ is the weighted estimated mean effect size (standardized mean difference) for level j across the k studies that is calculated as

$$\hat{\bar{\delta}}_j = \sum_{i=1}^{k} w_i \hat{\delta}_i/\sum_{i=1}^{k} w_i \qquad (5)$$

(Hedges & Olkin, 1985). In equation (5), w_i is the weight of each study calculated as $w_i = 1/\text{var}(\hat{\delta}_i)$ as in equation (3). The variance of the mean effect size in equation (5) is calculated as

$$\text{var}(\hat{\bar{\delta}}_j) = \left[\sum_{i=1}^{k} w_i\right]^{-1} \qquad (6)$$

(Hedges & Olkin, 1985).

Assumptions of the FE model. The classic FE model assumes that k studies have been deliberately selected, that population effect sizes are the same (or very similar) to allow average effect sizes to be meaningfully estimated and tested, and that inferences are restricted to the sampled studies. The classic FE model provides information about the results of these particular studies but provides no or limited information about the generalizability of results to similar studies (Hedges, 1998; Rosenthal et al., 2006).

Assumptions of the RE model. Unlike the classic FE model, the RE model assumes that k studies have been randomly sampled from a clearly defined superpopulation of studies which allows unconditional inferences (Hedges & Olkin, 1985). Regarding the effect size variation, the FE model is assumed to reflect within-study sampling error only, while the RE model is assumed to have two sources of variation: variance due to within-study sampling error as well as a random-effect variance, which involves differences in effect sizes due to between-study differences (Rosenthal et al., 2006).

Other FE and RE models. In addition to the Hedges and Olkin (1985) (HOd) model of mean effect sizes difference, several popular FE models also include the Hedges & Olkin (1985) approach for correlation coefficients (HOr) and the Rosenthal & Rubin (RR, 1979, 1982) method for transformed correlations. The frequently used RE models include the Hunter & Schmidt method (HS, 1990), and the DerSimonian-

Laird method (DSL, DerSimonian and Laird, 1983, 1986), which is also referred to as the Hedges & Vevea method (HV, 1998). Table 1 from Schulze (2004) compares these approaches by their effect sizes, weights, underlying models, and tests of homogeneity.

Table 1

Overview of Meta-analytic Methods (Schulze, 2004, p.186)

Method	Effect Size	Weight	Model	Homogeneity Test
HOr	z	n-3	FE	Q
HOd	d	$1/\ \mathrm{var}(\hat{\delta}_i)$	FE	Q
RR	z	n	FE	-
HS	r	n	RE	75 or 95% & Q
DSL/HV	z	$[(n-3)^{-1}+\mathrm{var}(\hat{\zeta}_i)]^{-1}$	RE	-

Note. The effect sizes include mean differences (d), correlations (r), and rs transformed to Fisher - zs (z). Fisher's transformation is used to correct a skew in the sampling distribution that occurs for extreme r values. On the other hand, Hunter and Schmidt (1990) advocated the untransformed correlation coefficient to avoid the problems arising from the Fisher's transformation (see Field, 2001 and references therein).

Performance of the FE and RE models. There are several Monte Carlo studies that have examined the performance of the FE and RE models as well as comparing their performance on estimating the mean effect sizes, associated variances, and moderating effects. Table 2 presents a summary of the results from these studies. In general the classic FE and RE models have satisfactory power and good control of Type I error rates in mean effect size estimation (Johnson et. al., 1995) as well as detecting moderator effects (Johnson et. al., 1995; Sanchez-Meca & Marin-Martinez, 1998a). These findings apply for increasing numbers of studies and/or sample sizes, different mean magnitude of effects, and varying heterogeneity of effect sizes (Johnson et. al., 1995). Both FE and RE models seemed to provide unbiased and comparable effect size estimation (Berkey et al., 1998; Brockwell & Gordon, 2001; Field, 2001; Johnson et. al., 1995) and reasonably accurate and consistent findings (confidence intervals, Type I error rates, and power) to assess moderator effect sizes when their assumptions were met (Overton, 1998).

However, the results from the two types of models showed evidence of bias when their assumptions were violated. The FE models seemed to perform well only when there was very little between-study variance (Brockwell & Gordon, 2001). It has been found consistently across studies that the FE models did not take into account the between-study variation and thus had standard errors smaller than that of the RE models (e.g. Berkey, et al., 1998; Brockwell & Gordon, 2001; Noortgate & Onghena, 2003b), especially for a smaller set of studies with small sample sizes (Noortgate & Onghena, 2003b). With the underestimated standard error, the FE models had narrower confidence intervals. This led to inflated Type I error rates in estimating mean effect sizes and detecting moderator effects (e.g. Field, 2001; Higgins & Thompson, 2004; Overton, 1998). The bias in Type I error rates, detection

probabilities, and confidence interval of the FE models were expected to increase as the within-studies error variance increases and the number of studies decreases (Overton, 1998).

The RE models, on the other hand, greatly overestimated the sampling error variance when the true between-study differences were fixed (Overton, 1998). This overestimation, which tended to decrease with increasing numbers of studies, caused large confidence intervals with unsatisfactory precision of the estimated moderator effect (Overton, 1998). These findings suggest that using a FE or RE model can lead to different results and conclusions for the same set of studies.

Table 2

Summary of Studies Examining the FE & RE Models in Estimating Mean Effect Size, Variances, and Moderator Effect

Author	Estimator	Simulation conditions	Findings
Aguinis, 2008	Detecting moderator: the HO, the HS, and the Aguinis & Pierce (AP) approaches	Number of studies; sample sizes; population moderating effect size; sources of variance (measurement error and range restriction)	Overall rank: (a) point estimates: HS > AP > HO; (b) Type I and Type II error rates regarding homogeneity tests: HS < HO < AP; (c) Type I and Type II rates regarding moderating tests: HO = HS > AP. The AP approach was recommended in general for meta-analyses including strong theory-based hypotheses and for data with severe levels of range restriction.
Berkey, et al., 1998	The RE and FE meta- regression	Population heterogeneity; sample sizes	Both FE and RE methods provided unbiased estimates of the regression coefficients. The FE models seriously underestimated the standard errors of regression coefficients under heterogeneous conditions, causing narrower CI for coefficients and biased p-values. The RE models provided estimated models and inferences that were more realistic and trustful than were the FE estimates.
Bohning, et al., 2002	Heterogeneity variance estimated by the DSL method	Number of studies; sample sizes; population heterogeneity	The DSL estimator using inverse population-averaged study-specific variances as weights was unbiased. Using estimates of the study-specific variances instead led to considerable bias. The RE approach by DSL would lead to the most different result from the FE models if there was strong heterogeneity.

Table 2

Summary of Studies Examining FE & RE Models in Estimating Mean Effect Size, Variances, and Moderator Effect (continue)

Author	Estimator	Simulation conditions	Findings
Brockwell & Gordon, 2001	Mean effect size estimation: the FE and the RE (DSL, maximum likelihood and profile likelihood) methods	Number of studies; between-study variation	The FE method only performed well when there was very little between-study variation. The RE methods generally performed better than the FE methods in terms of coverage probabilities. The RE methods all generally had coverage below the nominal level. Profile likelihood method produced the highest coverage probabilities in all cases.
Field, 2001	Mean effect size (r) estimation: the HO and the HS methods	Population effect sizes; sample sizes; number of studies	Homogenous case: comparable estimates of population effect sizes for both methods. The Type I error rates for tests of homogeneity were equally well controlled by the two methods when population effect sizes were small to medium but better controlled by the HO method for large effect sizes. Heterogeneity case: larger bias in mean effect size for the HO than the HS method. Power of both techniques was < 0.3 when the average population effect size was small. The HO method had lower power for small numbers of studies. Neither method controlled the Type I errors rates for < 15 studies.
Field, 2005	Mean and variances of effect sizes (r) estimation: the HV and the HS methods	Number of studies; sample sizes; mean population correlations; population standard deviation	The HS method produced the mean effect size estimates with the least error, although both methods were very accurate. Confidence intervals from the HS method were slightly too narrow, but became more accurate than those from the HV method as the number of studies, the size of the true correlation and the variability of correlations increased.

Table 2

Summary of Studies Examining the FE & RE Models in Estimating Mean Effect Size, Variances, and Moderator Effect (continue)

Author	Estimator	Simulation conditions	Findings
Hall & Brannick, 2002	Mean and variances of effect sizes (r) estimation: the HS and the HV models	Mean and standard deviation of correlation r; number of studies; attenuation; sample sizes	Both ρ and σ_ρ^2 appeared larger for the HV model. The credibility intervals for the HV model were conservative. The uncorrected HS method performed poorly when the data were attenuated. The HS method was preferable with more realistic credible intervals.
Higgins & Thompson, 2004	The FE and RE meta-regression	Population heterogeneity; number of studies; weights; number of covariates; correlation between covariates	The FE method had high type I error rates in the presence of heterogeneity. The RE method referring the standard T-statistic to a normal distribution was highly anticonservative for small numbers of studies. The problem may partly be avoided using tests based on a t-distribution for the test statistic. All RE meta-regression methods performed well on single covariates when the number of studies was large.
Johnson et al., 1995	Compare the HO, the RR, and the HS methods	Number of studies; mean effect size; sample sizes	The HO and RR methods yielded similar results. The HS method reached more conservative estimates of significance and wider confidence intervals than the other two methods. The HS method should be used only with caution.

Table 2

Summary of Studies Examining FE & RE Models in Estimating Mean Effect Size, Variances, and Moderator Effect (continue)

Author	Estimator	Simulation conditions	Findings
Oswald & Johnson,1998	Mean effect size (r) estimation from the HS method	Population effect size (r); data distribution, number of studies; sample sizes	ρ was accurate on average (slightly bias for non-normal data) and σ_ρ^2 was consistently negatively biased for all distributions. One would correctly conclude more than half the time that no moderator effects existed. However, cases of variation in ρ and especially in σ_ρ^2 indicated that results from individual meta-analyses could deviate substantially from what was found on average.
Overton,1998	The FE and RE models in testing moderator	Size of the moderator effect; heterogeneity; sample sizes	The FE models seriously underestimated and the RE models greatly overestimated sampling error variance when their basic assumptions were violated, which caused biased confidence intervals and hypothesis tests. The FE models had inflated Type I error rates in detecting a significant moderator.
Sackett et al., 1986	Mean differences in effect sizes: the HS ratio of expected to observed variance, the Callender-Osburn (CO) procedure, and a chi-square test	True population differences; number of studies; sample sizes; measurement error	Small true differences were not detected regardless of sample sizes and numbers of studies. Moderate true differences were not detected with small numbers of studies or small sample sizes. The HS procedure was consistently more powerful than the CO and chi-square procedures while having Type I error rates typically larger than .20. Both the CO and chi-square procedures had identical power and controlled the Type I error rates at the .05 level.

Table 2

Summary of Studies Examining FE & RE Models in Estimating Mean Effect Size, Variances, and Moderator Effect (continue)

Author	Estimator	Simulation conditions	Findings
Sanchez-Meca & Marin-Martinez, 1998	Detecting moderator: the T test, the HO, and the RR methods	Sample sizes; number of studies; correlation between population effect sizes and moderator	Good control of Type I error rates for all 3 procedures in all conditions. Similar and satisfactory power (especially for larger sample sizes and larger number of studies) for the HO and RR procedures and systematically lower in T test.
Van den Noortgate & Onghena, 2003a	Mean effect size estimated from the FE, the RE models based on observed effect sizes, the weighted average of the observed effect sizes, and empirical Bayes estimates	Sample sizes, number of studies; population mean effect size, variance of mean effect size.	Considering the bias and mean square error (MSE), for a small number of studies, the ordinary approach performed better; for a moderately sized dataset, the empirical Bayes approach had the best results; for a large meta-analytic dataset, the iterative approach of HO could be recommended. Precision-weighted average of observed effect sizes resulting from homogeneous studies, with optimal weights estimated using the observed effect sizes, gave a biased estimate of the mean effect. Bias was larger when the studies were heterogeneous and the bias increased with increasing numbers of studies.

Criticisms of the Classic FE and RE Models

Criticisms of the FE models. The FE models are more commonly used than the RE models in educational and psychological research in large part because they are conceptually and computationally simple and easy to manage compared to RE models (Cooper, 1997; Hunter & Schmidt, 2000; National Research Council [NRC], 1992). As an example, the journal *Psychological Bulletin* represents a prime outlet for high quality meta-analyses but only 13 out of 199 (6.5%) of the published meta-analyses between 1986 and 2006 used the RE methods (Schmidt et al., 2009).

Despite their popularity the criticisms of FE models have prompted the development of methods that do not possess various shortcomings. Bonett (2009a) proposed a modified FE model which does not require these assumptions and made a compelling argument that estimating and testing contrasts under this model can be a powerful explanatory tool.

As revealed in empirical studies, heterogeneous effect sizes also produced inflated Type I error rates in significance tests for the FE models (Brockwell & Gordon, 2001; Field, 2001; Higgins & Thompson, 2004; Hunter and Schmidt, 2000; Overton, 1998). Additionally, omitting variation in the population of studies produced smaller standard errors of the mean effect size, causing overly narrow confidence intervals (Hunter and Schmidt, 2000). Considering the universal heterogeneity of studies in meta-analysis (NRC, 1992), the FE models tend to result in false positive findings and overestimate the precision of the findings.

Similarly, the false positive findings and overestimation of precision also plague moderator effects in FE models. Reviewing and reanalyzing empirical fixed-effect meta-analytic studies published in the *Psychological Bulletin* with the RE model, Schmidt et al. (2009) concluded that most of the meta-analysis results in this

journal might be substantially in error in their statements of precision of findings. The inference is that using the FE models when it is not appropriate misrepresents cumulative evidence and complicates the actual phenomena studied, which are potentially serious consequences (Hunter and Schmidt, 2000).

Criticisms of the RE models. The criticisms of the classic FE models prompted the NRC (1992) and some researchers (e.g., Hunter & Schmidt, 2000) to call for increased use of RE models in meta-analysis. However, RE models may not be the reasonable choices as an important criticism is their unrealistic fundamental assumption of random sampling (Bonett, 2008a, 2009a; Hedges & Vevea, 1998). Bonett (2009) suggested that the random sampling assumption might never be satisfied in typical meta-analyses. Since the studies were published chronologically, the recent ones might be designed intentionally to be similar or different from the previous ones. Without knowing if the set of studies were a true random sample, the statistical inferences could not be generalized to the superpopulation. In addition, the superpopulation did not even exist or was only "imaginary" since it was implausible to describe the superpopulation in detail. Thus, Bonett argued that the RE model had limited scientific value and should not be recommended for routine use.

Other shortcomings of the RE models include the biased inferences when the assumption of equal within-study variances is not satisfied, and biased confidence intervals for variance components for even slight departures from the assumption of normality. Both the classic FE and RE models have also been criticized for their use of empirically based weights based on Var($\hat{\delta}_i$), which tends to produce biased estimates since the weights and individual study effect sizes are expected to be correlated (Shuster, 2009).

Tests of Heterogeneity

The selection between FE models or RE models is crucial, since the two types of models produce statistically different estimates, confidence intervals, and significant tests for mean effect sizes and for moderator analysis (Hunter & Schmidt, 2004; Rosenthal et al., 2006). In meta-analytic practice model selection is usually guided by tests of heterogeneity.

Performance. Table 3 shows the results of several Monte Carlo studies examining and comparing the commonly used tests of heterogeneity in recent years, including the Q test (Hedges, 1983), the Hunter-Schmidt (HS) 75%- or 90%- rule (Hunter et al., 1982), the likelihood ratio ((Hartley & Rao, 1967), and the Wald and score tests (Lehmann, 1999, Verbeke & Molenberghs, 2003). Consistently observed across these studies was that all the tests of heterogeneity had low and insufficient power, especially for small sample sizes, number of studies, small degree of heterogeneity, and non-normal data distributions. Among these tests, the Q test seemed to be relatively better than the alternative tests with larger power and tighter control of Type I error rates in most conditions, although it still did not perform satisfactorily. However, even with the inadequate performance the HS 75%- or 90%- rule was used much more often in practice than the Q test (Schulze, 2004).

The low power of these tests and the nature of model selection suggest that the test of heterogeneity could be potentially misleading and should not be the only determinant of model choice in meta-analysis (Hardy, 1998; Schulze, 2004; Steel, et al., 2002; Shuster, 2009). Inspecting relevant normal plots and clinical insight may be more relevant to both investigating and modeling heterogeneity (Hardy, 1998).

Table 3

Summary of Studies Examining the Tests of Heterogeneity

Author	Estimator	Simulation conditions	Findings
Cornwell & Ladd, 1993	The HS-75% Rule	Sample sizes; effect sizes; range restriction; measurement error; numbers of correlations	No practical bias in estimating mean ρ for small sample sizes, few correlations, and substantial measurement error. Low power and high Type I error rates to detect heterogeneity among the ρs under almost all situations.
Cornwell, 1993	The Q test, the chi-square, and the likelihood ratio test	Mean population ρ; population variances; sample sizes; number of correlations; reliability of predictor and criterion; variances of predictor; bivariate distributions	Distribution had very little effect on Type I error rates and power for the 3 tests. The Q and chi-square tests controlled the Type I error rates well and had power > 50% when k > 18. The likelihood ratio test had inflated Type I error. For all 3 tests power decreased with range restriction or measurement error when k < 18. The Q test was recommended.
Harwell, 1997	The Q test	Number of studies; study sample sizes; population heterogeneity; distributions; group sample sizes (equal, unequal) and group variances ratios (equal, unequal)	Equal effect sizes hypothesis was rejected less than expected if smaller study sample sizes were paired with larger numbers of studies. Pairing smaller variances with larger sample sizes (or vice versa) led to inflated Type I error rates (especially for skewed data). The power was also less than expected when small study sample sizes were paired with larger numbers of studies.

Table 3

Summary of Studies Examining the Tests of Heterogeneity (continue)

Author	Estimator	Simulation conditions	Findings
Huedo-Medina et al., 2006	The Q statistic using the d index (QH) and using the g index (QG); the I^2 using the d index (I^2H) and using the g index (I^2G)	Number of studies; between-study variances; within-study variances; sample sizes; raw score distributions (normal/non-normal)	Good control of the Type I error rates for QH and I^2H for non-normal and homoscedastic data. Inflated Type I error rates of QG and the I^2G for nonnormal data. Both procedures had higher power as the number of studies, the average sample size, and the between-studies variance increased. Insufficient power (< .8) for a small number of studies (k < 20) and/or average sample size (N < 80). Both procedures had similar power when the normality and homoscedasticity assumptions were not met and had higher power with the g index than with the d index.
Sagie & Koslowsky, 1993	The HS-75% rule for uncorrected r; the HS-75% rule for corrected r, the HS-95% rule for uncorrected r, the HS-95% rule for corrected r, the Q statistic, and the credibility interval	Subpopulation correlations; criterion reliability; range restriction; sample sizes; numbers of studies	When the differences between the population correlations were small, power for all techniques were relatively low. Overall, the HS rules and the Q statistic had greater power but higher Type I error rates than credibility intervals. The HS-75% rules and Q statistic were recommended.
Sanchez-Meca & Marin-Martinez, 1997	The Q test and the HS-75% and 90% procedures	Numbers of studies; sample sizes; population effect sizes	The Q test had good control of while the HS procedures had inflated Type I error rates. The HS procedures presented greater power. In all conditions, the power was very low, particularly for small sample sizes, number of studies, and small differences between the parametric effect sizes.

Table 3

Summary of Studies Examining the Tests of Heterogeneity (continue)

Author	Estimator	Simulation conditions	Findings
Sidik & Jonkman, 2007	Variance component estimator (VC), method of moments estimator (MM), maximum likelihood estimator (ML), restricted maximum likelihood estimator (REML), empirical Bayes estimator (EB), model error variance estimator (MV), and a variation of the MV estimator (MVvc)	Sample sizes; mean effect size; variance heterogeneity;	The REML and the ML and MM estimators had large biases unless the true heterogeneity variance was small. The VC estimator tended to overestimate the heterogeneity variance, but was accurate when the number of studies was large. The MV estimator was not a good estimator when the heterogeneity variance was small to moderate, but it was reasonably accurate when the heterogeneity variance was large. The two estimators MVvc and EB were found to be the most accurate in general, particularly when the heterogeneity variance was moderate to large.
Spector & Levine, 1987	The HS and U statistic for homogeneity	Population correlations; numbers of correlations; sample sizes	The Type I error rate for the HS method was high in many conditions compared with U, which was uniformly robust. Power for the HS method increased with increasing size of population differences, sample size per correlation, and number of correlations compared. The U statistic had much lower power in most conditions.

Table 3

Summary of Studies Examining the Tests of Heterogeneity (continue)

Author	Estimator	Simulation conditions	Findings
Takkouche et al., 1999	The Q statistics, the weighted least square, $Z^2_{WLS.R}$, Z^2_K , the likelihood ratio test, and the bootstrap versions of them	Population heterogeneity; numbers of studies; data distributions	The Q statistic and the bootstrap versions of the other tests gave the correct type I error but all of the tests had low statistical power, especially when the number of studies was small. Considering the validity, power, and computational ease, the Q statistic was the best choice. Under exponential distribution, the Q statistics had lower power than that in the normal setting.
Viechtbauer, 2007	The Q test, the likelihood ratio, and the Wald and score tests	Number of effect sizes; mean population effect size; within-study sample sizes; population heterogeneity	The Q test kept the tightest control of the Type I error rates, especially for large sample sizes. The other homogeneity tests did not control the Type I error rates adequately. When using raw correlation as the effect size and the average sample size within studies was low, increasing the number of effect sizes resulted in inflated Type I error rates. The power to detect heterogeneity depends on the number of effect sizes, the sample sizes within the studies, and the amount of heterogeneity.

Weighting Schemes

Two weighting schemes. There are two commonly used weighting schemes in meta-analysis to combine study effect sizes: sample size weights (Hunter & Schmidt, 2004) and the inverse variance weights (Hedges & Vevea, 1998). Both of these attempt to give greater weight to studies with better precision. The rationale for sample size weights is that studies with larger sample sizes are expected to be estimating the true value more precisely than studies with smaller sample sizes, and thus should be assigned larger weights (Hedges & Olkin, 1985; Hunte & Schmidt, 2004). Inverse variance weights describe the degree of precision associated with sampling error and underlying population effect size variability. In the Hedges & Olkin FE model the weight is calculated as the inverse of the variance of an effect size, which takes into account the sample size of each study (see equation 5). As discussed by Gersten et al. (2005), the weighted effect size takes into consideration: (1) the number of studies conducted on a specific intervention, (2) the sample size of the study, and (3) the magnitude and consistency of effects. These three elements are essential to evidence based practice.

Unit weights. Unit weights are another method to combine effect sizes. This method does not assign different weights to studies with different precision or quality. Bonett (2008, 2009) adopted and advocated the use of unit weights in his fixed-effect meta-analytic model, and suggested that unit weights may show a smaller mean squared error. Bonett argued that for a small number of studies, when the effect sizes are inconsistent across studies some highly weighted studies may be distant from the population mean but strongly influence the estimated mean. Assigning larger weights to

these studies may result in poor estimation of population mean effect size. Shuster (2009) similarly suggested that any empirical weighting other than unit weights risk serious bias for targeted population parameters because of the potential for a few studies to dominate estimation.

Performance of weighting schemes. A simulation study by Sanchez-Meca and Marin-Martinez (1998b) showed that the inverse variance weights method was systematically more efficient than sample size weights, and the lower variability in the inverse variance weights provided more accurate estimates of population effect size. This study also found that in both homogeneous and heterogeneous cases the inverse variance weights provided slightly higher bias than the sample size weights, especially when the sample size was small. Although this bias was not practically important, Sanchez-Meca and Marin-Martinez (1998b) recommended that researchers apply the inverse variance weights with caution when the sample size in studies in a meta-analysis is less than 30.

A recent study by Brannick, Yang and Cafri (2010) compared the inverse variance weights, sample sizes weights, and unit weights using Monte Carlo simulation in which the population values were derived from published meta-analytic studies. For the RE model, which these authors argued is a better representation of realistic meta-analysis despite Bonett's objections, the inverse of variance weights yielded generally the most accurate estimates of the population standardized mean difference. The advantage of the inverse of variance weights over the unit weights increased as the underlying effect size variance increases. The authors concluded that the sample size weights method was preferable for correlation effect sizes. For effect sizes of mean differences, the inverse of

variance weights (as used in the Hedges & Olkin FE model) performed best; the unit weights provided an estimate of the overall mean that was nearly as good but did not provide as accurate an estimate of the underlying random-effects variance.

Ordinary Least Square (OLS) vs. Weighted Least square (WLS)

Early meta-analytic work used ordinary least square (OLS) to estimate parameters in moderator analyses of the relationship between study characteristics and effect sizes (Hedges & Olkin, 1985). However, the OLS procedure has several problems including the fact that the "homogeneous variance" assumption might not hold and empirical information about model specification is not available (Hedges & Olkin, 1985).

The classic FE model uses the weighted least square (WLS) procedure, which applies to both continuous and discrete independent variables. In additional to estimating the model parameters, WLS also leads to tests if a model adequately explains the variation in effect sizes by providing large sample tests of significance and an explicit test of the specification of the model (Hedges & Olkin, 1985). A study by Steel et al. (2002) reported that WLS had the best overall estimates and was the only method that actually converged toward the true moderator effect size as the number of studies increased for varying multicollinearity and heterogeneous conditions. These authors also found that WLS was associated with tests of greater power in detecting continuous moderators over other methods including OLS.

Bonett's (2009a) Interval Estimation Model for Standardized Mean Differences Expressed as a Contrast

Assumptions and equations. As noted earlier Bonett (2009a) proposed an interval estimation model for mean effect sizes expressed as a contrast that he claimed does not possess the deficiencies of the classic FE and RE models. Specifically, the proposed model does not assume random sampling of studies (meaning that inferences are strictly conditional on sampled studies), does not require equal variances within studies or assume that population effect sizes are equal, and does not employ weighted averages based on Var($\hat{\delta}_i$) thus circumventing the criticism of Shuster (2009a).

In Bonett's (2009a) model the effect size of study i (i=1…k) for level j (j=1…m) of a discrete moderator is calculated using equation (1) except that the denominator ($\hat{\sigma}_i$) is calculated as the square root of

$$\hat{\sigma}_i^2 = (\hat{\sigma}_{iT}^2+\hat{\sigma}_{iC}^2)/2, \qquad (7)$$

The variance of $\hat{\delta}_i$ for independent treatment and control groups is

$$\text{var}(\hat{\delta}_i)=[\hat{\delta}_i^2(\hat{\sigma}_{iT}^4/df_{iT}+\hat{\sigma}_{iC}^4/df_{iC})/8\hat{\sigma}_i^4+(\hat{\sigma}_{iT}^2/df_{iT}+\hat{\sigma}_{iC}^2/df_{iC})/\hat{\sigma}_i^2 \qquad (8)$$

where the degrees of freedom are $df_{iC}=n_{iC}-1$ and $df_{iT}=n_{iT}-1$, and n_{iT} and n_{iC} are the sample sizes in the treatment and control groups, respectively. Equation (8) allows within-study variances to be unequal.

To compare and contrast mean effect sizes among different levels of a discrete moderator a 100(1-α)% confidence interval can be calculated as:

$$\sum_{j=1}^{m} c_j \hat{\bar{\delta}}_j \pm z_{\alpha/2}\left[\text{var}\left(\sum_{j=1}^{m} c_j \hat{\bar{\delta}}_j\right)\right]^{1/2}, \qquad (9)$$

where c_j is the contrast coefficient of level j of the moderator and $\hat{\bar{\delta}}_j$ is the mean effect size (standardized mean difference) obtained from level j across k studies and is calculated as:

$$\hat{\bar{\delta}}_j = k^{-1} \sum_{i=1}^{k} b_i \hat{\delta}_i, \qquad (10)$$

where k is the number of studies in level j of a discrete moderator and b_i is an approximate bias adjustment which is typically Hedges (1981) adjustment of [l - 3/(4n_i - 9)] where n_i is the total sample size in study i. The variance of the mean effect size for level j is calculated as

$$\text{var}(\hat{\bar{\delta}}_j) = \left[k^{-2} \sum_{i=1}^{k} b_i^2 \text{var}(\hat{\delta}_i)\right]^{1/2} \qquad (11)$$

(Bonett, 2009a).

The key differences between Bonett's (2009a) modified FE model and the classic FE model are in how they combine effect sizes and in the calculation of the variance of effect sizes. In Bonett's method a simple average (unit weights) is used to calculate mean effect size (see equation 10) whereas in the classic FE model the mean effect size is calculated using a weight equal to the inverse of Var($\hat{\delta}_i$) (see equation 5). Bonett's model also allows within-study variances to be unequal (see equation 8) whereas the classic FE approach assumes these parameters share a common value (see equation 2).

New classification scheme of meta-analysis methods. Bonett (2010) proposed a new classification scheme of meta-analysis methods based on three basic statistical models: the constant coefficient model, the varying coefficient model, and the random

coefficient model. The constant coefficient model and the varying coefficient model are both FE models. Besides the meta-analytic confidence intervals for mean differences, Bonett also proposed meta-analytic models for correlation coefficients (2008). His meta-analytic models for effect sizes of mean difference and correlation are both based on varying coefficient models. The computation of confidence interval for mean effect size and linear contrasts under these models were described in detail in Bonett's (2008a, 2008b, 2009a, 2009b, 2010) papers.

Performance of the Bonett's model. Bonett (2009a) conducted Monte Carlo studies of the ability of his proposed FE model of mean differences to control Type I error rates and produce narrow confidence intervals of a contrast for various sample sizes and patterns of population effect sizes. Bonett (2009) reported that his model always had a coverage probability of .95 across conditions, while the classic FE model could have a coverage probability far below .95. Bonett also reported that his model had average confidence interval width that was slightly narrower than that of the classic RE model, suggesting better precision. However, Bonett's (2009a) results were limited to normally distributed and homoscedastic data for a limited number of within-study sample sizes and numbers of studies.

Chen and Peng (2012) conducted a simulation study to compare Bonett's and several other models to construct confidence intervals for a single standardized linear contrast of means in one-way fixed-effects between-subjects univariate ANOVA designs. They discovered that Bonett's model had adequate coverage probabilities when there was moderate violation of normality and equal variance assumptions. The confidence interval

width tended to be wider compared to other methods examined in this study, which were noncentral confidence intervals, symmetric percentile bootstrap confidence intervals, and bias-corrected and accelerated confidence intervals. These authors recommended the Bonett's model for moderate nonnormal data and unequal variances conditions.

Factors affect the performance of meta-analytic estimators. A review of Monte Carlo studies in meta-analytic research (as shown in Table 2 and 3) suggests that several factors influence the performance of meta-analytic estimators and tests including skewness of the data and heteroscedasticity (e.g., Brannick et al, 2010; Hardy & Thompson, 1998; Harwell, 1997; Sanchez-Meca & Marin-Martinez, 1998a; Steel et al. 2002). In particular, skewed data, which are commonly observed and examined in educational and psychological studies, can impact meta-analytic estimation and hypothesis testing (e.g. Brannick et al, 2010; Cornwell, 1993; Hardy & Thompson, 1998; Harwell, 1997; Huedo-Medina et al., 2006; Micceri, 1989; Oswald & Johnson, 1998; Sanchez-Meca & Marin-Martinez, 1998a; Steel et al. 2002; Takkouche et al., 1999). For example, Harwell's (1997) Monte Carlo study of the Hedge's homogeneity test (Q test) found that non-normal (skewed) data affected both Type I error rates and power. Specifically, when the distribution was skewed, the Type I error rates were inflated under unequal variance conditions across various study sample sizes and numbers of studies. Similarly, somewhat lower power of the Q test under non-normal data was also observed in the Takkouche et al. (1999) study. In the Steel et al. (2002) study, it was reported that the effect of moderators was underestimated for skewed data and heterogeneity of effect sizes.

Distributions of effect sizes also appear to frequently show significant skewing. Sanchez-Meca and Marin-Martinez (1998) concluded that many meta-analyses have a large number of effect sizes near a certain mean value but also show evidence of a few outliers that are responsible for large skew values. The impact on estimation and hypothesis testing of such skewing can be important, especially for the RE model.

Summary and Objectives. In summary, the classic FE and RE models have both been criticized because these models can perform poorly when their assumptions are violated. In real-world meta-analytic practice, data often show certain degrees of skewness, violating the normality assumption of the FE and RE models and potentially leading to problematic results and incorrect conclusions.

Bonett (2009a) proposed a fixed-effect model which has fewer restrictions than the classic FE model. The inverse variance weights are not adopted in this modified model and thus its formulas are less complex than the classic FE model. However, there is little research examining its behavior under realistic meta-analytic conditions on factors affecting meta-analytic model performance. Given the scarceness of literature, it is important to investigate the behaviors of both the FE and the Bonett models under non-normal distributions which frequently occurred in realistic meta-analysis. Thus, this study is trying to provide practitioners with more statistical evidence in choosing among meta-analytic models. In contrast to the classic FE model, which calculates the variance of effect sizes based on normal approximation, the Bonett model adjusts for the degrees of freedom and can be considered an exact method. It is expected that the Bonett model will perform better under skewed data distributions. The random-effects models are not

examined in this study because: 1) The Bonett model is a fixed-effects model, and 2) Bonett criticized the RE models for the random sampling assumptions and suggested that the RE models should not be used (2009a).

Methods

Design of the Simulation Study

Parameters of the simulation study. A factorial design was adopted for the Monte Carlo study with five independent variables. In all cases the within-study effect size consisted of a treatment and control group comparison. Informed by an examination of meta-analyses published in the *Review of Educational Research* and *Psychological Bulletin* between 2004-2009, the conditions of the Monte Carlo study were (i) numbers of studies ($k_1 = 30$, $k_2 = 60$, and $k_3 = 120$, representing small, medium, and large number of studies typically observed in meta-analysis), (ii) within-study sample sizes ($n_{iT} = n_{iC} = 15$, or $n_{iT} = 15$ and $n_{iC} = 30$, representing equal or unequal within-study sample sizes), (iii) raw data distribution (normal distribution with mean of 0 and variance of 1, gamma distribution with scale of 1 and rate of 1, chi-square with df = 4, chi-square with df = 8, and Laplace distribution with location of 0 and scale of 1. The shapes of these non-normal distributions deviate from normal distributions with different degrees), and (iv) within-study variances (1:1, or 3:1, representing typical equal or unequal within-study variances). Altogether $3 \times 2 \times 5$ by 2 = 60 conditions were studied for Type I error rates and power, respectively.

Data generation. All data generation and computation were conducted in R using the following steps:

The analysis of Type I error rates.

A. Generate raw data for treatment and control groups with specific distributions and sample sizes for a study.

For raw data from a normal distribution with equal within-study sample sizes and equal within-study variances, the mean values of treatment and control groups with sample size n_t and n_c were generated with mean of 0 and variance of 1. Raw data from other distributions with unequal within study variances were generated similarly with certain manipulations. For example, if the raw data distribution is chi-square with 4 degrees of freedom, the distribution of means will follow χ_4^2. To generate unequal within study variances ($variance_c$:$variance_t$=1:3), the raw data of the treatment group was multiplied by the square root of 3. The means of the control and treatment groups were set to be 0 to allow for the effect size of a single study to be 0.

B. Repeat this process k times to generate raw data for k studies in a hypothetical meta-analysis. In this step, the raw data were generated k times to create 30, 60 and 120 studies separately.

C. Compute effect size and variance of effect sizes for the k studies in a hypothetical meta-analysis under the classic FE and the Bonett models, respectively, based on equations in the literature review.

Assuming the k studies are clustered into 3 groups, with each group representing a level of a hypothetical discrete moderator, compute mean effect sizes (d_1, d_2 and d_3) and variances of effect sizes. The contrasts will be conducted among the 3 groups.

D. Compute pairwise contrast of the null hypothesis: $d_1=d_2$.

The k studies were equally divided into 2 groups for the simple/pairwise contrast. For example, to divide k=30 studies into 2 groups, the 1st to the 15th studies were clustered into group 1, and the 16th to the 30th studies were clustered into group 2. The confidence intervals of the contrast under the classic FE model and the Bonett model were then computed.

E. Compute complex contrast of the null hypothesis: $d_1-(d_2+d_3)/2=0$.

The k studies were equally divided into 3 groups for the complex contrast. For example, when k=30, group 1, 2 and 3 had 10 studies in each group. The confidence intervals of the contrasts under the classic FE model and the Bonett model were then computed.

F. Repeat the above process 2000 times to generate data for 2000 hypothetical meta-analytic studies for each experimental condition. The confidence intervals under the classic FE model and the Bonett model were also computed, respectively, for each replication.

G. Across the 2000 replications, compute the Type I error rates and average widths for the confidence interval under the FE model and the Bonett model, respectively.

The analysis of power. The power analysis was conducted for a simple contrast among different levels of a hypothetical moderator. The results of the complex contrasts are provided for normally distributed data only since the pattern was expected to be similar to that of the simple contrasts.

H. Determine the mean difference of effect sizes between the two levels of the hypothetical moderator in a simple contrast.

Since a power value of 0.8 is usually considered as satisfactory, the mean difference between two levels of a moderator will be calculated to allow a power value of 0.8 based on the Hedges & Olkin's classic FE model.

The mean effect sizes difference will be estimated under the following specific condition:

1) normal distribution,
2) k=30 (each level of the moderator included 15 studies),
3) $n_{iT}=n_{iC}=15$,
4) $\sigma^2_T:\sigma^2_C=1:1$.

The mean effect size difference obtained will be used in the power analysis for all conditions of the Monte Carlo study, under both the classic FE and Bonett's models, to allow between-model comparisons of power across various study conditions such as different data distributions.

The power calculation for moderators in meta-analysis was demonstrated by Hedges & Pigott (2004). As the first step in this procedure, the noncentrality parameter λ_B was computed as the following:

For a two-group contrast, $\lambda_B = \frac{(\hat{\bar{\delta}}_1 - \hat{\bar{\delta}}_2)^2}{\text{var}(\hat{\bar{\delta}}_1)+\text{var}(\hat{\bar{\delta}}_2)}$ (12),

The 95% critical value for the central chi-square distribution with 1 degree of freedom is 3.841. Denote q as the value of the cumulative distribution of the noncentral chi-square at critical value 3.841, with 1 degree of freedom and a noncentrality parameter

of λ_B. To achieve power of 0.8, set q as q = 1- 0.8 = 0.2. Using R, calculate the noncentrality parameter λ_B as approximately 7.85 (pchisq (3.841, 1, λ_B) = 0.2).

Using equations (1)-(6) and (12), when k = 30 ($m_1 = m_2 = 15$) and $n_{ic} = n_{it} = 15$, for data following a standard normal distribution, the following mean effect size difference is obtained

$$\hat{\bar{\delta}}_2 - \hat{\bar{\delta}}_2 = 0.3752.$$

In summary, the mean effect size difference between the two levels of the simple contrast should be 0.3752 to achieve 80% statistical power to detect the true between-group difference for the 30 studies, with normally distributed raw data, equal within-study sample sizes, and equal within-study variances.

I. Generate raw data for treatment and control groups with specific distributions and sample sizes, for studies in level 1 and studies in level 2 of the moderator as described in step C separately. Raw data for studies in in level 1 were generated as described in step A above. Raw data for studies in level 2 were generated similarly with certain manipulations, to allow a mean effect size difference of 0.3752 between studies in level 1 and 2.

For example, for a study in level 1(say, D_1), the means of the treatment and control groups will be 0 so that an effect size of a single study in D_1 will be 0. To generate unequal within variances ($variance_c$: $variance_t$ = 1:3), the raw data of the treatment group was multiplied by the square root of 3. For a study in D_2, a value of 0.3752 was added to each generated raw value for the treatment group. This produces an

effect size of 0.3752 for this single study in D_2. Setting up the effect sizes of studies in D_1 to be 0 and of studies in D_2 to be 0.3752 allows a mean effect size difference of 0.3752 between the two levels of a moderator.

J. Repeat this process k/2 (k = 30, 60, or 120) times to generate raw data for k/2 studies in D_1 and k/2 studies in D_2 of the moderator, respectively, for the simple contrast with equal numbers of studies in each level. Altogether the raw data of k studies will be generated for a hypothetical meta-analysis.

K. The effect sizes (d_1, d_2) and their variances will be estimated for the k studies in a hypothetical meta-analysis using the raw (simulated) data under the classic FE and the Bonett models, respectively, based on equations in the literature review. The comparison between d_1 and d_2 effect sizes allows pairwise contrasts to be estimated corresponding to differences in average effect size between the two levels of a moderator.

L. Estimate confidence intervals of the simple contrasts, $\hat{\psi} = d_1 - d_2$ ($\hat{\psi}$ is the estimated contrast), under the classic FE model and the Bonett models.

M. For normally distributed data only, the complex contrasts can be estimated by equally dividing the k studies into 3 groups, with each group representing a level of the moderator (see scenario in step E). The effect sizes (d_1, d_2, and d_3) of each group and their variances were computed under the classic FE and the Bonett models. Then the complex contrasts were estimated as $\hat{\psi} = d_1 - (d_2 + d_3)/2$ and confidence intervals for these contrasts computed.

N. Replicate the above process 2000 times to generate data for 2000 hypothetical meta-analytic studies for each power condition. Compute the confidence intervals under the classic FE model and the Bonett models for each replication.

O. Across the 2000 replications, compute the power and average widths for the confidence interval under the FE model and the Bonett models, respectively.

Dependent Variables

The analysis of Type I error rates. Type I error rates were calculated as the proportion of times the Bonett or the classic FE model rejected a true null hypothesis across the 2000 replications. Error rates near .05 were implied that a test/model is performing well.

Measures of estimation precision were the confidence intervals (CI) for the pairwise and complex contrasts among levels of a hypothetical discrete moderator, the coverage probabilities (the proportion of the Cl covering the true value, which equals to 1-type I error rate) of the CI, and confidence interval widths as estimated across the 2000 replications within each of the conditions. The narrower the average confidence interval the more precise the estimation. The coverage probabilities were expected to be close to 0.95 if a test performs well. The Type I error rates and the coverage probabilities allowed the evaluation of which method performs better under which condition.

The analysis of power. Power was calculated as the proportion of times that the Hedges & Olkin classic FE or the Bonett model successfully rejected a false null hypothesis across the 2000 replications. Power greater or equal to 80% was considered sufficient. Since the mean effect sizes difference between the two groups of the simple

contrast was set to allow a power value of 0.8 for the classic FE model, power of around 0.8 was expected for this model under this condition (normally distributed data with equal within study sample sizes and variances for k = 30).

The CI for the pairwise contrasts and the complex contrasts (for normally distributed data only) among levels of a hypothetical discrete moderator, and the confidence interval widths, were estimated across the 2000 replications within each of the conditions.

Results

Descriptive statistics of Raw Data Generated

Descriptive statistics of raw data generated by R are reported in this section to examine if the software generated data follow the theoretical distributions. These statistics are based on five samples with sample sizes of 50,000 each (with arbitrarily selected seed = 100) following the five distributions (normal distribution with mean of 0 and variance of 1, gamma distribution with scale of 1 and rate of 1, chi-square with df = 4, chi-square with df = 8, and Laplace distribution with location of 0 and scale of 1 respectively).

The density plots of the five samples are shown in Figure 1. In general, the empirical distributions follow the theoretical shapes. The normal and Laplace distributions were approximately symmetric, while the gamma and chi-square distributions with degrees of freedoms of 4 and 8 were skewed as expected.

Table 4 displays the empirical and theoretical moments of generated data and goodness-of-fit test results. Examining the mean, variance, skewness, and kurtosis for

each of the five distributions, it was found that the empirical and theoretical values were very similar to each other. For example, the normal distribution generated with mean of 0 and variance of 1 had an empirical mean and variance that were the same as the theoretical values, and the empirical skewness and kurtosis that were extremely close to the theoretical values of 0. The goodness-of-fit (Kolmogorov-Smirnov) test produced a p-value of 0.8 which was not statistically significantly at $\alpha = 0.05$ level, suggesting that the generated normal sample did follow the normal distribution with mean of 0 and variance of 1 as specified. Similar patterns can be observed for the other 4 distributions. With p-values of 0.1, 0.95, 0.85, and 0.06, it can be concluded that the 4 samples generated by R had empirical distributions that were approximately the same as their theoretical distributions.

Table 4

The Theoretical and Empirical Distribution of Generated Data and Goodness-of-fit tests results

	Empirical				Theoretical				
Distribution	Mean	Variance	Skewness	Kurtosis	Mean	Variance	Skewness	Kurtosis	P-value
Normal	0.00	1.00	-0.01	-0.01	0	1	0	0	0.80
Gamma (1, 1)	1.00	1.02	2.06	6.47	1	1	2	6	0.95
Chi-sq df = 4	4.02	8.01	1.40	2.91	4	8	1.41	3	0.10
Chi-sq df = 8	8.00	16.09	1.03	1.63	8	16	1	1.50	0.85
Laplace (0,1)	-0.01	2.01	-0.01	3.14	0	2	0	3	0.06

Note. The goodness-of-fit test is the Kolmogorov-Smirnov test

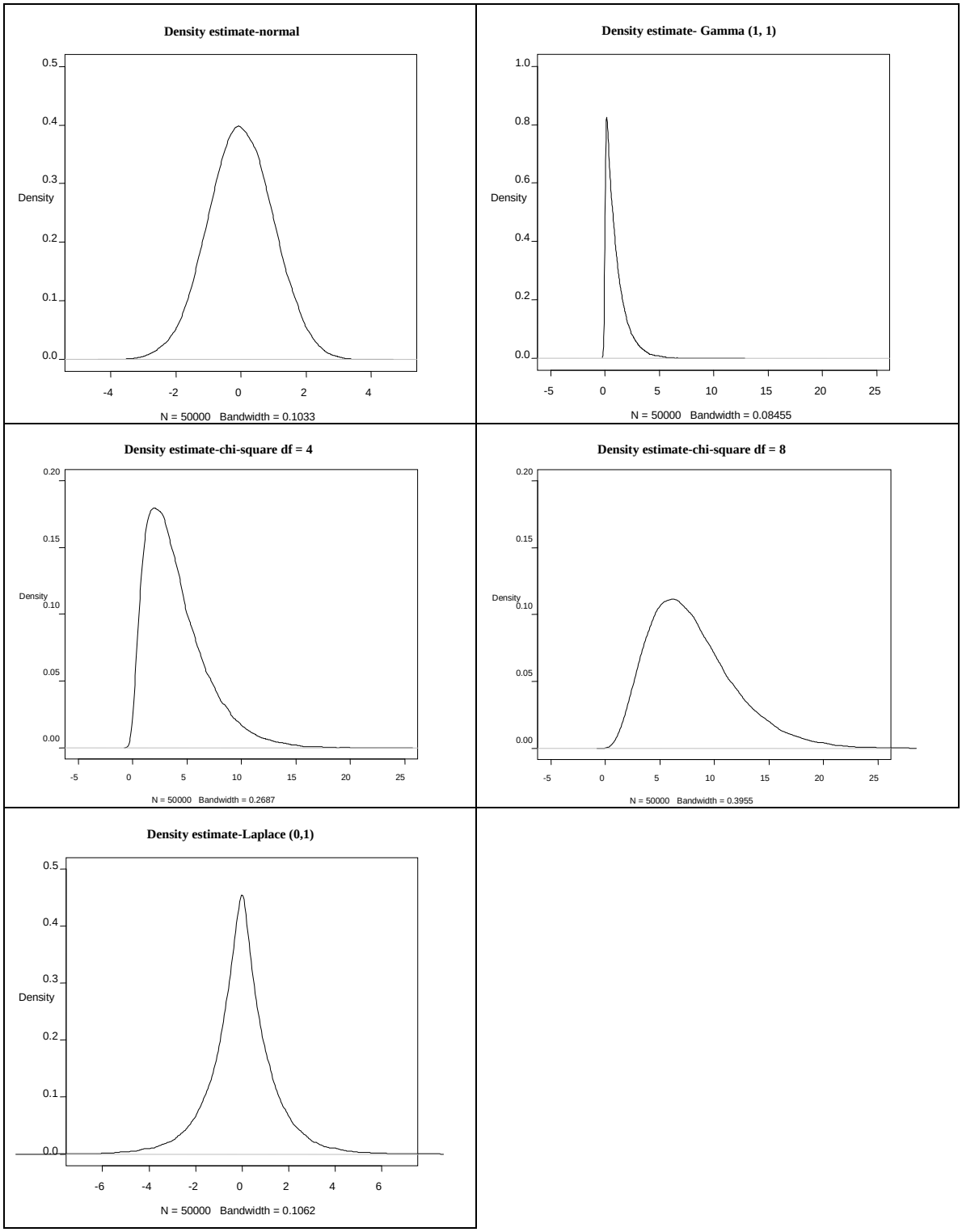


Figure 1. The Density Plots of Empirical Distribution of Raw Data Generated

The Effect Sizes Variances in the FE and the Bonett Models

In this section, we compare the empirical effect size variances computed using the FE and the Bonett models under two levels of within-study sample sizes (equal/unequal or $n_{iT} = n_{iC} = 15$; $n_{iT} = 15$, $n_{iC} = 30$) and 2 within-study variances ratios (equal/unequal or 1:1/1:3), and 3 numbers of studies (k = 30, 60 and 120). Tables 5 to Table 9 show the means of the first (Q_1), second (Q_2), and the third (Q_3) quartiles of the empirical effect size variances across the 2000 replications for a normal, gamma, chi-square distribution with 4 or 8 degrees of freedom, and Laplace distributions respectively. The effect size variances parameters for the FE and the Bonett models, respectively, are: 1) equal sample sizes and equal variances: 0.1333 and 0.1429; 2) equal sample sizes and unequal variances: 0.1333 and 0.1429; 3) unequal sample sizes and equal variances: 0.1000 and 0.1059; 4) unequal sample sizes and unequal variances: 0.1000 and 0.096.

Table 5 shows that under a normal distribution across various numbers of studies the Bonett model had slightly larger mean quartiles of effect size variances than that of the FE model under the following conditions: equal within-study sample sizes paired with equal within-study variances, equal within-study sample sizes paired with unequal within-study variances, as well as unequal within-study sample sizes paired with equal within-study variances. The only condition under which the Bonett model had slightly smaller mean quartiles of effect size variances was unequal within-study sample sizes paired with unequal within-study variances. For gamma and chi-square distributions with 4 and 8 degrees of freedom, respectively, as presented in Table 6, Table 7 and Table 8, the conditions under which the Bonett model had slightly larger mean quartiles of effect size

variances than that of the FE model were: equal within-study sample sizes paired with equal within-study variances, equal within-study sample sizes paired with unequal within-study variances, and unequal within-study sample sizes paired with unequal within-study variances. When pairing unequal within-study sample sizes with equal within-study variances, the FE model had slightly larger Q_1 but smaller Q_2 and Q_3 values than the Bonett model.

Similar to chi-square distributions, for the Laplace distribution the Bonett model also had mean quartiles of effect size variances slightly larger than the FE model under the same conditions of equal within-study sample sizes paired with equal within-study variances, equal within-study sample sizes paired with unequal within-study variances, and unequal within-study sample sizes paired with unequal within-study variances. Combining unequal within-study sample sizes with equal within-study variances, the Bonett model had slightly smaller Q_1 and larger Q_2 and Q_3 values than the FE model under Laplace distributions. Generally speaking, across different numbers of studies, in most conditions the effect sizes variances of the Bonett model were slightly larger than that of the FE model for all distributions. The patterns were similarly observed among the 4 skewed distributions across numbers of studies.

The histograms of the distributions of the effect size variances' quartiles across 2000 replications under different conditions across various numbers of studies can be found in Figure 2 through Figure 6. Most of the quartiles' distributions were slightly skewed. The FE and the Bonett models had similar distributions of effect size variances' quartiles in most conditions.

Table 5

Mean Quartiles of Effect Sizes Variance ~ Normal (0, 1)

		Equal sample sizes, Equal variances			Equal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1336	0.1344	0.1364	0.1336	0.1344	0.1364
	Bonett	0.1432	0.1441	0.1464	0.1434	0.1448	0.1482
k_2=60	FE	0.1336	0.1344	0.1364	0.1336	0.1344	0.1365
	Bonett	0.1432	0.1441	0.1463	0.1434	0.1448	0.1482
k_3=120	FE	0.1336	0.1344	0.1364	0.1336	0.1344	0.1364
	Bonett	0.1432	0.1441	0.1464	0.1434	0.1448	0.1482
		Unequal sample sizes, Equal variances			Unequal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1001	0.1006	0.1015	0.1001	0.1003	0.1009
	Bonett	0.1011	0.1068	0.1124	0.0753	0.0772	0.0797
k_2=60	FE	0.1001	0.1005	0.1015	0.1001	0.1003	0.1009
	Bonett	0.1010	0.1067	0.1124	0.0753	0.0772	0.0797
k_3=120	FE	0.1001	0.1005	0.1015	0.1001	0.1003	0.1009
	Bonett	0.1010	0.1068	0.1124	0.0753	0.0772	0.0797

Table 6

Mean Quartiles of Effect Sizes Variance ~ Gamma (1, 1)

		Equal sample sizes, Equal variances			Equal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1337	0.1345	0.1365	0.1352	0.1383	0.1428
	Bonett	0.1432	0.1443	0.1469	0.1453	0.1497	0.1572
k_2=60	FE	0.1336	0.1345	0.1365	0.1352	0.1382	0.1427
	Bonett	0.1432	0.1443	0.1469	0.1452	0.1496	0.1571
k_3=120	FE	0.1337	0.1345	0.1365	0.1352	0.1382	0.1426
	Bonett	0.1432	0.1443	0.1469	0.1452	0.1496	0.1569
		Unequal sample sizes, Equal variances			Unequal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1002	0.1006	0.1015	0.1015	0.1037	0.1071
	Bonett	0.0958	0.1048	0.1149	0.1174	0.1293	0.1410
k_2=60	FE	0.1002	0.1006	0.1015	0.1015	0.1037	0.1071
	Bonett	0.0957	0.1048	0.1148	0.1174	0.1293	0.1410
k_3=120	FE	0.1002	0.1006	0.1015	0.1015	0.1037	0.1071
	Bonett	0.0956	0.1047	0.1148	0.1172	0.1293	0.1410

Table 7

Mean Quartiles of Effect Sizes Variance ~ Chi-square (df = 4)

		Equal sample sizes, Equal variances			Equal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1336	0.1345	0.1364	0.1381	0.1428	0.1491
	Bonett	0.1432	0.1442	0.1466	0.1490	0.1559	0.1660
k_2=60	FE	0.1336	0.1345	0.1365	0.1382	0.1428	0.1492
	Bonett	0.1432	0.1442	0.1467	0.1491	0.1559	0.1661
k_3=120	FE	0.1336	0.1345	0.1365	0.1381	0.1428	0.1492
	Bonett	0.1432	0.1442	0.1467	0.1490	0.1559	0.1660
		Unequal sample sizes, Equal variances			Unequal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1002	0.1006	0.1015	0.1040	0.1075	0.1121
	Bonett	0.0980	0.1056	0.1139	0.1240	0.1357	0.1477
k_2=60	FE	0.1002	0.1006	0.1015	0.1040	0.1075	0.1122
	Bonett	0.0979	0.1055	0.1139	0.1240	0.1355	0.1477
k_3=120	FE	0.1002	0.1006	0.1015	0.1040	0.1074	0.1122
	Bonett	0.0979	0.1055	0.1137	0.1239	0.1354	0.1476

Table 8

Mean Quartiles of Effect Sizes Variance ~ Chi-square (df = 8)

		Equal sample sizes, Equal variances			Equal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1336	0.1345	0.1364	0.1450	0.1519	0.1612
	Bonett	0.1432	0.1442	0.1465	0.1580	0.1683	0.1824
k_2=60	FE	0.1336	0.1345	0.1365	0.1450	0.1520	0.1613
	Bonett	0.1432	0.1442	0.1465	0.1580	0.1684	0.1826
k_3=120	FE	0.1336	0.1345	0.1364	0.1449	0.1519	0.1612
	Bonett	0.1432	0.1442	0.1465	0.1580	0.1683	0.1825
		Unequal sample sizes, Equal variances			Unequal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1001	0.1006	0.1015	0.1095	0.1148	0.1214
	Bonett	0.0994	0.1060	0.1130	0.1341	0.1469	0.1609
k_2=60	FE	0.1001	0.1006	0.1015	0.1095	0.1147	0.1214
	Bonett	0.0994	0.1061	0.1131	0.1341	0.1469	0.1609
k_3=120	FE	0.1001	0.1006	0.1015	0.1096	0.1147	0.1214
	Bonett	0.0995	0.1061	0.1131	0.1343	0.1470	0.1610

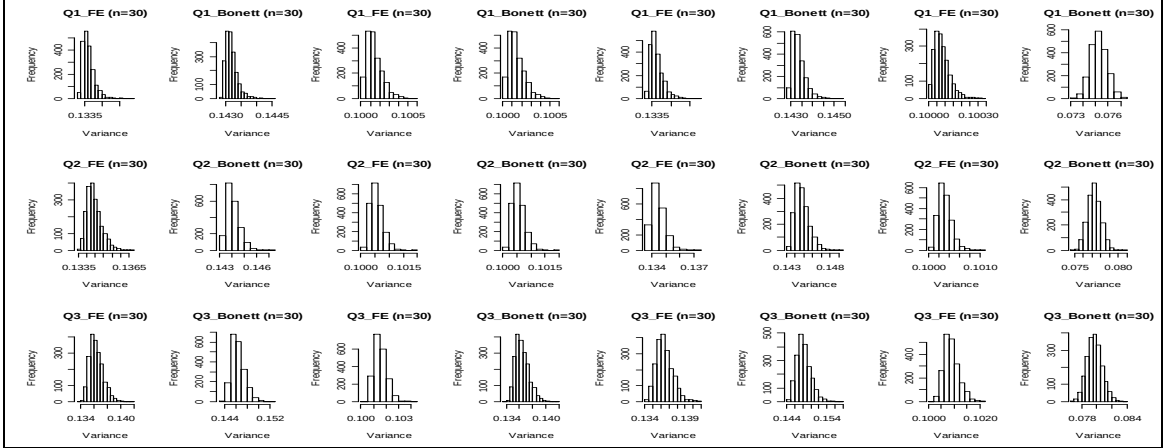
Table 9

Mean Quartiles of Effect Sizes Variance ~ Laplace (0, 1)

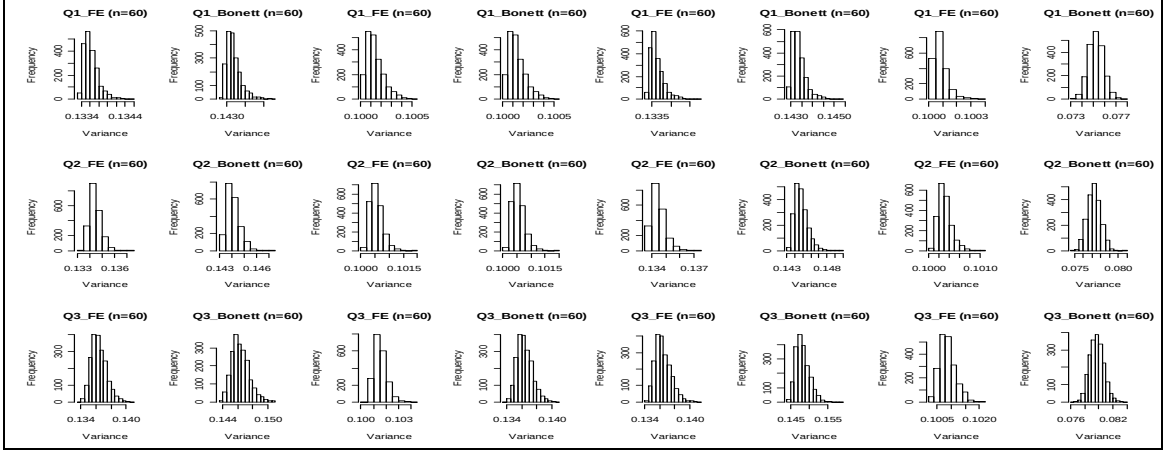
		Equal sample sizes, Equal variances			Equal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1336	0.1345	0.1364	0.1336	0.1345	0.1365
	Bonett	0.1432	0.1442	0.1466	0.1433	0.1444	0.1471
k_2=60	FE	0.1336	0.1345	0.1364	0.1336	0.1345	0.1365
	Bonett	0.1432	0.1442	0.1466	0.1433	0.1444	0.1471
k_3=120	FE	0.1336	0.1345	0.1364	0.1336	0.1345	0.1365
	Bonett	0.1432	0.1442	0.1466	0.1433	0.1444	0.1471
		Unequal sample sizes, Equal variances			Unequal sample sizes, Unequal variances		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
k_1=30	FE	0.1001	0.1004	0.1010	0.1001	0.1004	0.1012
	Bonett	0.1001	0.1068	0.1137	0.1200	0.1257	0.1305
k_2=60	FE	0.1001	0.1006	0.1015	0.1002	0.1008	0.1021
	Bonett	0.0978	0.1060	0.1143	0.1181	0.1255	0.1316
k_3=120	FE	0.1001	0.1006	0.1015	0.1002	0.1008	0.1021
	Bonett	0.0978	0.1059	0.1143	0.1181	0.1255	0.1316

Equal sample sizes Equal variances	Equal sample sizes Unequal variances	Unequal sample sizes Equal variances	Unequal sample sizes Unequal variances

k_1=30



k_2=60



k_3=120

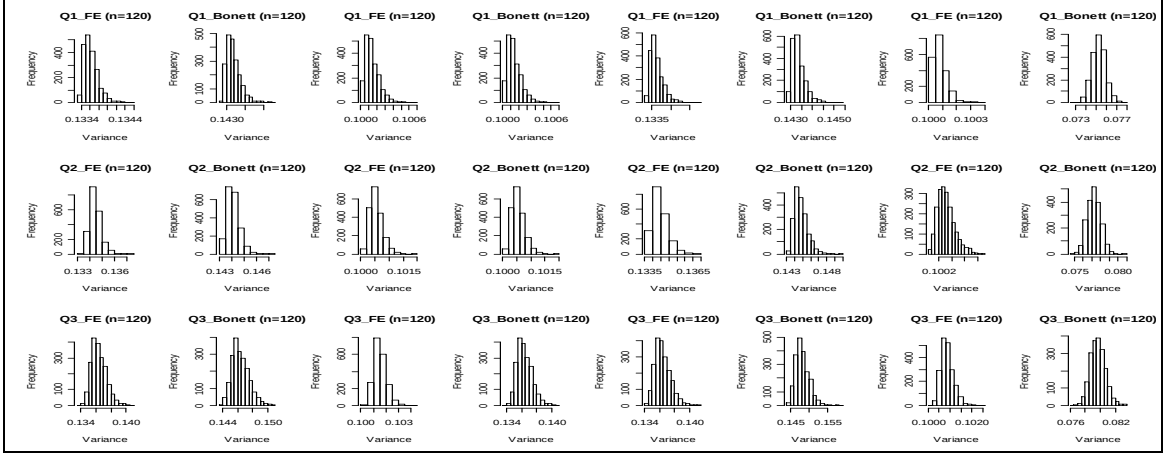


Figure 2. Comparing the Variance Quartiles between the FE and the Bonett Models (Normal)

Equal sample sizes Equal variances	Equal sample sizes Unequal variances	Unequal sample sizes Equal variances	Unequal sample sizes Unequal variances
k_1=30			
k_2=60			
k_3=120			

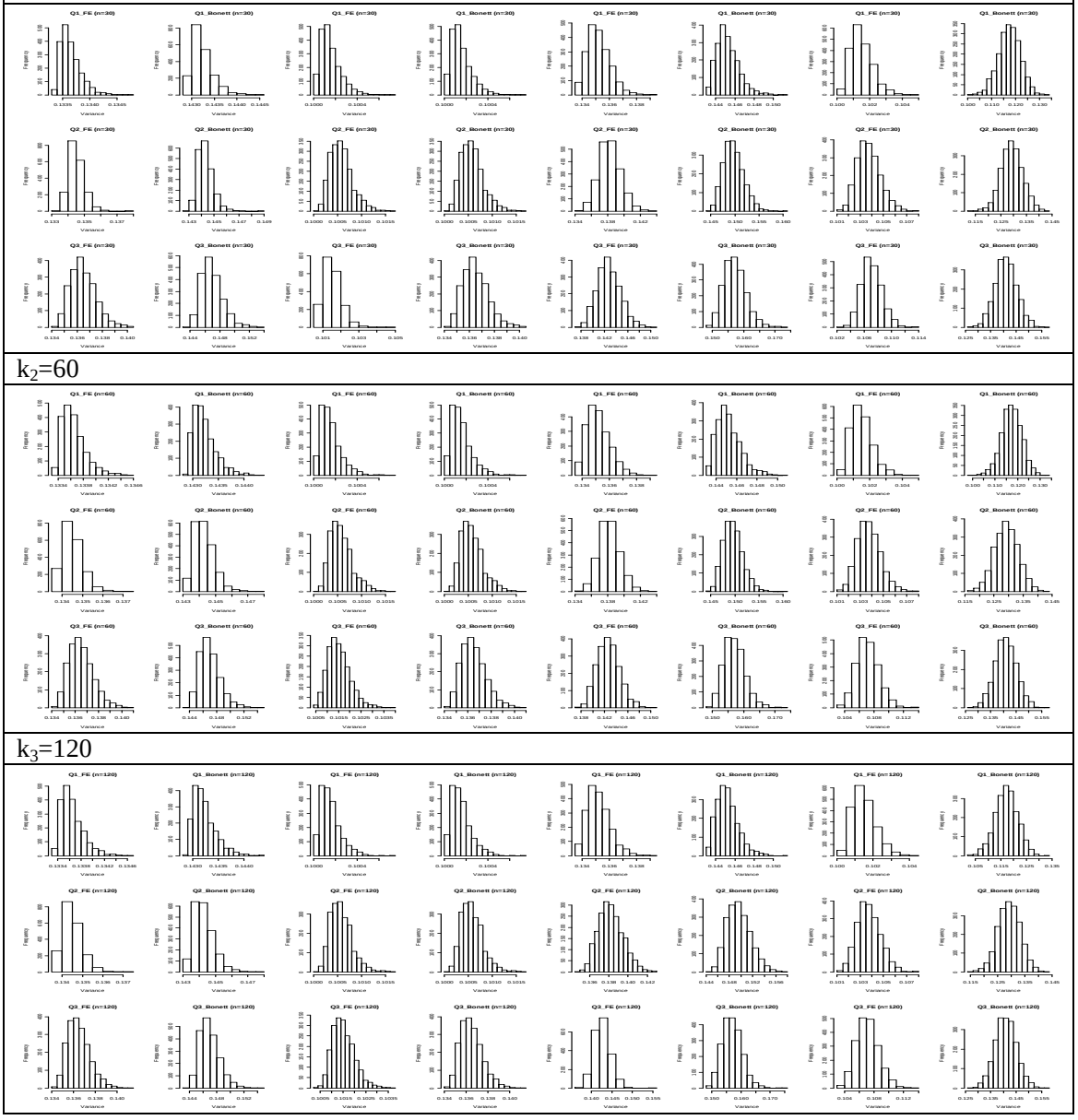
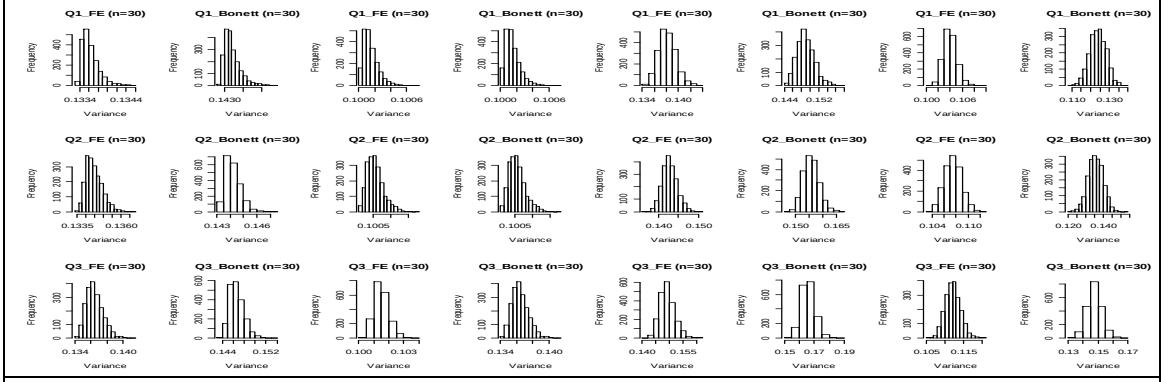


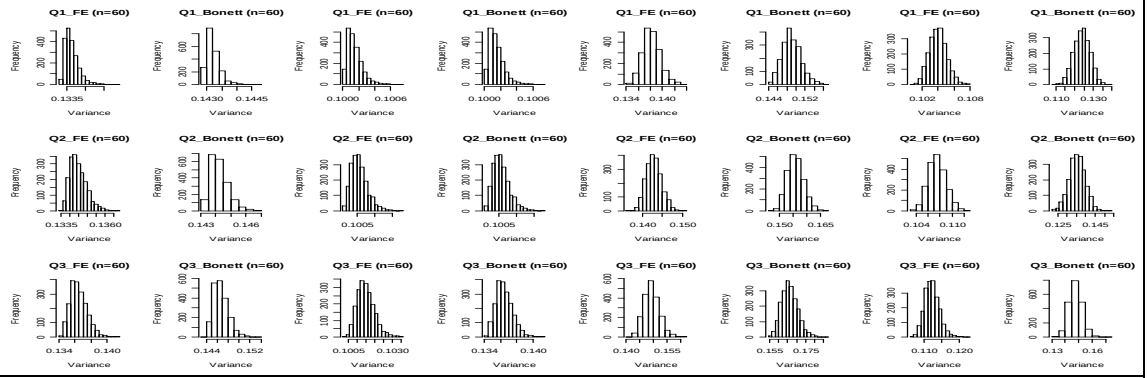
Figure 3. Comparing the Variance Quartiles between the FE and the Bonett Models (Gamma (1, 1))

Equal sample sizes Equal variances	Equal sample sizes Unequal variances	Unequal sample sizes Equal variances	Unequal sample sizes Unequal variances

k_1=30



k_2=60



k_3=120

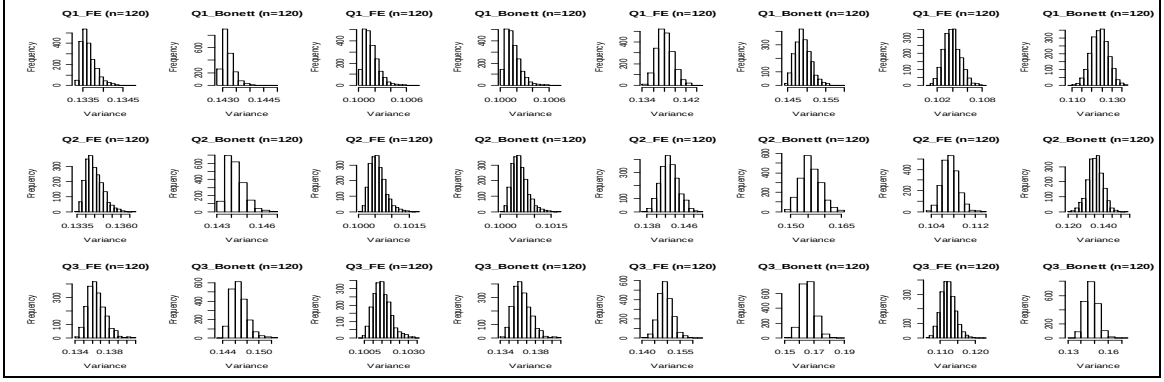
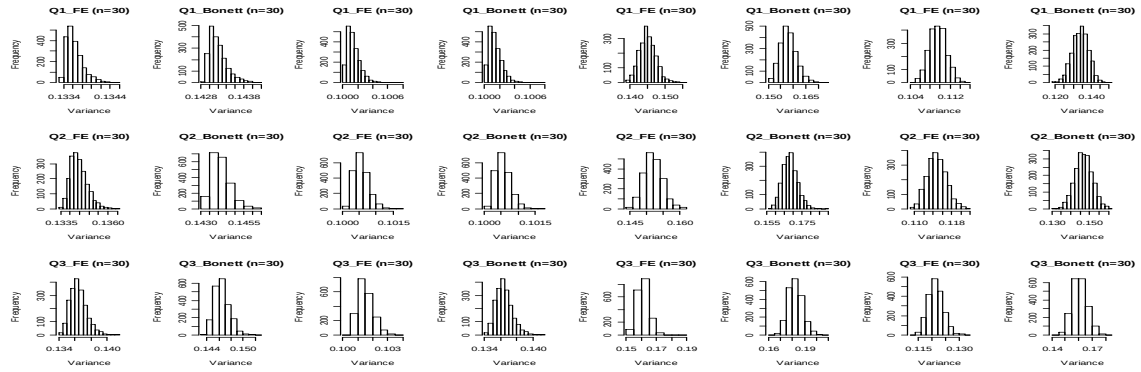


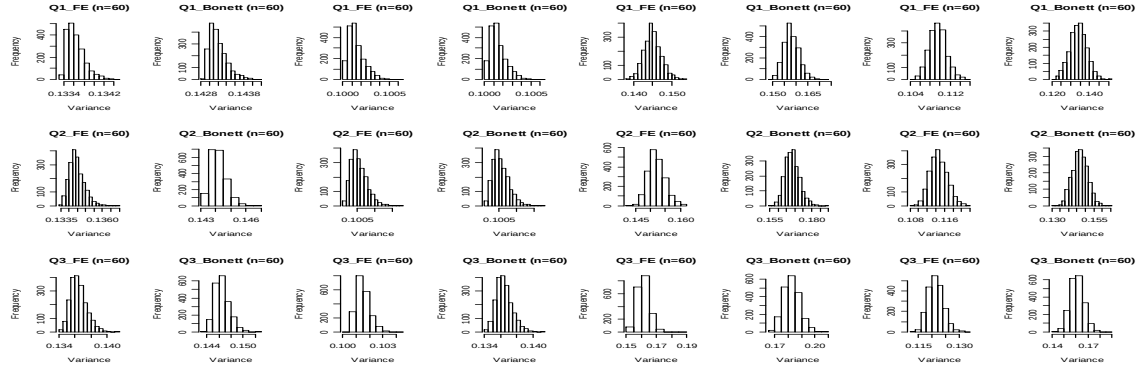
Figure 4. Comparing the Variance Quartiles between the FE and the Bonett Models (Chi-square df = 4)

Equal sample sizes Equal variances	Equal sample sizes Unequal variances	Unequal sample sizes Equal variances	Unequal sample sizes Unequal variances

k_1=30



k_2=60



k_3=120

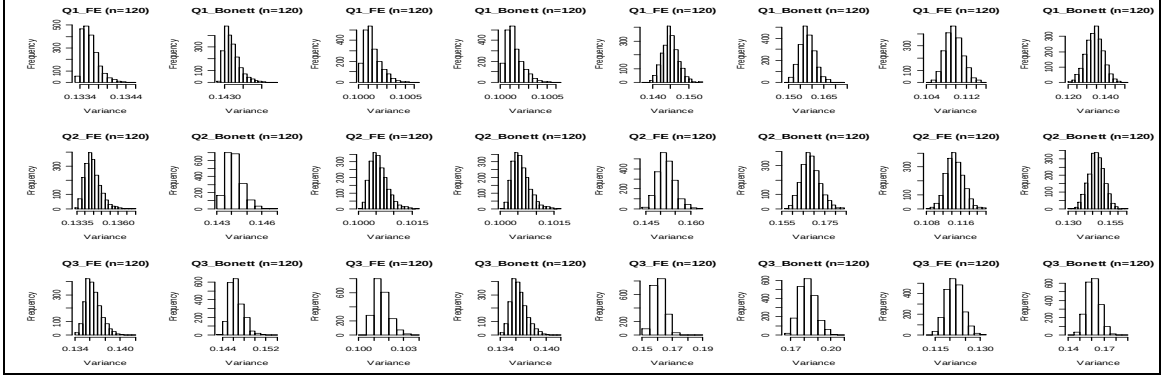
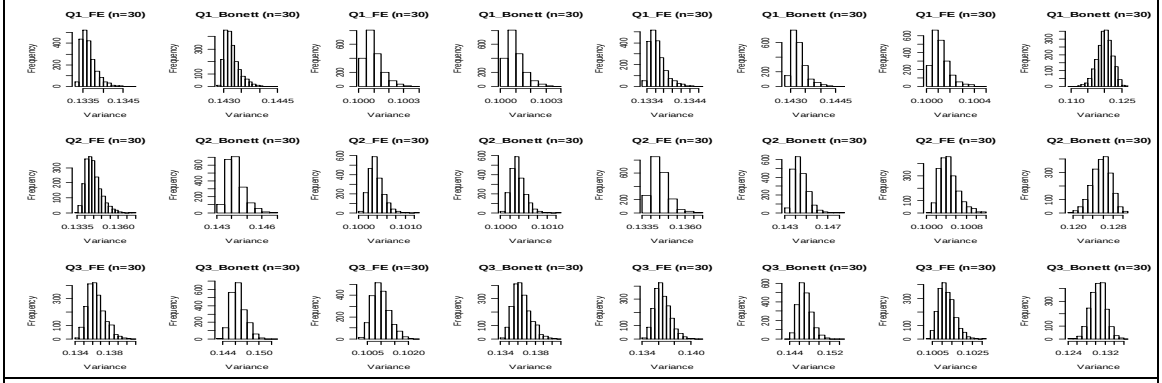


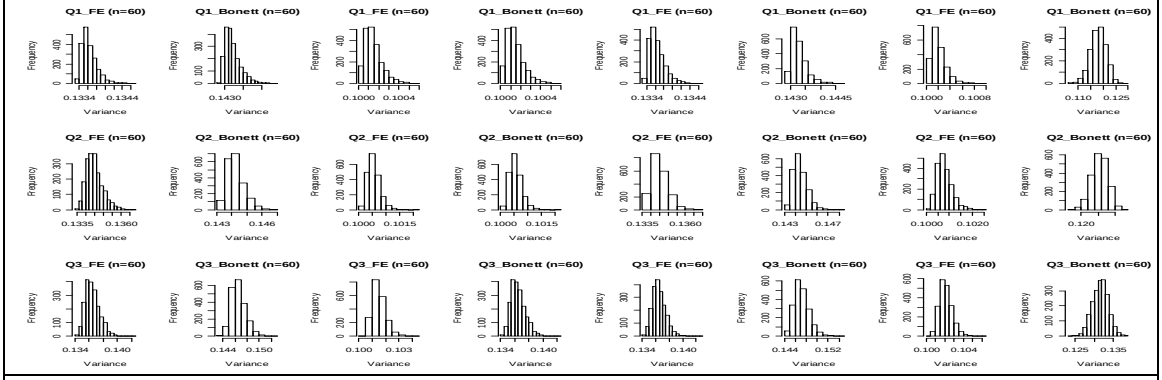
Figure 5. Comparing the Variance Quartiles between the FE and the Bonett Models (Chi-square df = 8)

Equal sample sizes Equal variances	Equal sample sizes Unequal variances	Unequal sample sizes Equal variances	Unequal sample sizes Unequal variances

k_1=30



k_2=60



k_3=120

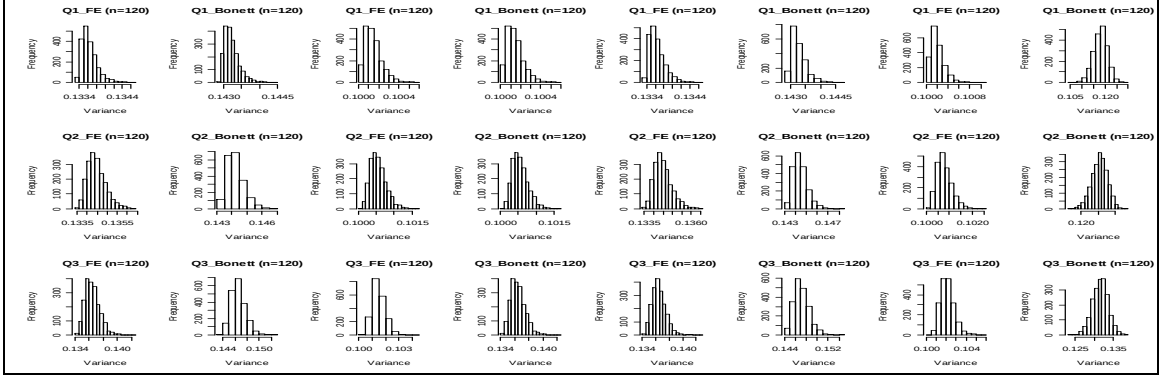


Figure 6. Comparing the Variance Quartiles between the FE and the Bonett Models (Laplace)

Comparing Type I Error Rates, Coverage Probabilities of Confidence Intervals and Confidence Interval Widths

To assess sampling error in the empirical Type I error rates for the simulation study confidence intervals are computed as CI = 0.05 +/- $1.96\times\sqrt{\frac{pq}{n}}$ = 0.05 +/- $1.96\times\sqrt{\frac{0.05\times0.95}{2000}}$ = 0.05 +/- 0.0049. Thus, empirical Type I error rates larger than 0.055 will be considered as inflated, and less than 0.045 as deflated. In meta-analytic practice, a Type I error may lead to more serious consequences than a Type II error. As a result, in this study, the models having inflated (rather than deflated) Type I error are considered as more unsatisfactory.

In this section the average coverage probabilities, Type I error rates, and the average confidence interval widths of the FE and the Bonett models for both simple and complex contrasts are summarized for the five distributions across the various data conditions examined. The results are reported in Table 10 through Table 14.

Shown in Table 10 are the results for a normal distribution. The results of the simple contrast and the complex contrast were very similar to each other. Comparing the FE model and the Bonett model, it was observed that the average confidence interval coverage probabilities were similar for the two methods in most conditions. With average confidence interval coverage probabilities around 0.95 or not far below 0.95, both of the methods performed well.

The method that produces smaller confidence interval width provides more precise estimation and thus should be preferred. The average confidence interval widths

for the FE model were similar to while slightly smaller than that of the Bonett model in most cases, which was consistent with the findings of Bonett's (2009) Monte Carlo study. It suggests that in these cases the FE model had a small advantage in providing estimates compared to the Bonett model. However, the Bonett model somewhat outperformed the FE model under the conditions of unequal within-study variances and unequal within-study sample sizes when $k_1 = 30$ and $k_3 = 120$. It was found in the previous section that under these conditions the Bonett model produced somewhat smaller effect size variances estimations than the FE model. When the smaller effect sizes variances were used to compute the confidence intervals, narrower confidence intervals were obtained. With increasing numbers of studies, both models provided narrower confidence intervals, suggesting more precise estimates.

To sum up, when the raw data distribution is normal the FE model and the Bonett model both perform well in most conditions with very similar coverage probabilities and confidence interval widths, suggesting their similar abilities in estimating and testing contrasts. The precision increases as the number of studies increases for both models. These findings were generally consistent with Bonett's (2009) findings that the proposed model performed well for normally distributed data.

Table 10

Type I Error Rates of the FE and the Bonett Models (Normal)

			Average coverage		Type I error		Average width	
			FE	Bonett	FE	Bonett	FE	Bonett
$\sigma^2_T:\sigma^2_C$			Distribution: Normal(0, 1)					
			Simple contrast					
1:1	k_1=30	$n_{iT}=n_{iC}$=15	0.958	0.960	0.042	0.040	0.527	0.531
		n_{iT}=15;n_{iC}=30	0.959	0.959	0.042	0.042	0.455	0.459
	k_2=60	$n_{iT}=n_{iC}$=15	0.956	0.957	0.044	0.043	0.373	0.376
		n_{iT}=15;n_{iC}=30	0.952	0.951	0.048	0.050	0.322	0.324
	k_3=120	$n_{iT}=n_{iC}$=15	0.952	0.952	0.048	0.048	0.264	0.266
		n_{iT}=15;n_{iC}=30	0.947	0.948	0.053	0.052	0.228	0.230
3:1	k_1=30	$n_{iT}=n_{iC}$=15	0.956	0.953	0.044	0.048	0.527	0.532
		n_{iT}=15;n_{iC}=30	0.981	0.958	0.020	0.042	0.454	0.419
	k_2=60	$n_{iT}=n_{iC}$=15	0.955	0.954	0.045	0.047	0.373	0.376
		n_{iT}=15;n_{iC}=30	0.975	0.947	0.025	0.053	0.321	0.296
	k_3=120	$n_{iT}=n_{iC}$=15	0.953	0.954	0.048	0.046	0.264	0.266
		n_{iT}=15;n_{iC}=30	0.978	0.944	0.023	0.056	0.227	0.209
			Complex contrast					
1:1	k_1=30	$n_{iT}=n_{iC}$=15	0.951	0.950	0.050	0.050	0.559	0.564
		n_{iT}=15;n_{iC}=30	0.953	0.953	0.047	0.047	0.483	0.487
	k_2=60	$n_{iT}=n_{iC}$=15	0.951	0.951	0.049	0.049	0.395	0.399
		n_{iT}=15;n_{iC}=30	0.948	0.949	0.052	0.051	0.341	0.344
	k_3=120	$n_{iT}=n_{iC}$=15	0.945	0.946	0.055	0.055	0.280	0.282
		n_{iT}=15;n_{iC}=30	0.949	0.949	0.051	0.051	0.241	0.244
3:1	k_1=30	$n_{iT}=n_{iC}$=15	0.957	0.958	0.043	0.042	0.559	0.567
		n_{iT}=15;n_{iC}=30	0.982	0.951	0.019	0.049	0.482	0.444
	k_2=60	$n_{iT}=n_{iC}$=15	0.956	0.954	0.045	0.046	0.395	0.399
		n_{iT}=15;n_{iC}=30	0.975	0.955	0.026	0.046	0.341	0.314
	k_3=120	$n_{iT}=n_{iC}$=15	0.944	0.946	0.056	0.054	0.280	0.282
		n_{iT}=15;n_{iC}=30	0.976	0.946	0.024	0.055	0.241	0.222

Table 11

Type I Error Rates of the FE and the Bonett Models (Gamma (1, 1))

			Average coverage		Type I error		Average width	
			FE	Bonett	FE	Bonett	FE	Bonett
$\sigma^2_T:\sigma^2_C$			Distribution: *Gamma* (1, 1)					
			Simple contrast					
1:1	k_1=30	$n_{iT}=n_{iC}$=15	0.959	0.958	0.042	0.043	0.527	0.532
		n_{iT}=15;n_{iC}=30	0.945	0.946	0.055	0.054	0.455	0.458
	k_2=60	$n_{iT}=n_{iC}$=15	0.947	0.948	0.054	0.053	0.373	0.376
		n_{iT}=15;n_{iC}=30	0.936	0.936	0.065	0.065	0.322	0.324
	k_3=120	$n_{iT}=n_{iC}$=15	0.949	0.950	0.051	0.051	0.264	0.266
		n_{iT}=15;n_{iC}=30	0.956	0.956	0.044	0.045	0.228	0.229
3:1	k_1=30	$n_{iT}=n_{iC}$=15	0.969	0.974	0.032	0.026	0.535	0.545
		n_{iT}=15;n_{iC}=30	0.928	0.977	0.073	0.023	0.463	0.506
	k_2=60	$n_{iT}=n_{iC}$=15	0.964	0.969	0.037	0.031	0.378	0.385
		n_{iT}=15;n_{iC}=30	0.927	0.969	0.073	0.032	0.328	0.358
	k_3=120	$n_{iT}=n_{iC}$=15	0.972	0.974	0.029	0.026	0.267	0.272
		n_{iT}=15;n_{iC}=30	0.942	0.978	0.058	0.022	0.232	0.253
			Complex contrast					
1:1	k_1=30	$n_{iT}=n_{iC}$=15	0.946	0.948	0.054	0.053	0.559	0.565
		n_{iT}=15;n_{iC}=30	0.956	0.958	0.044	0.042	0.483	0.486
	k_2=60	$n_{iT}=n_{iC}$=15	0.946	0.947	0.055	0.054	0.395	0.399
		n_{iT}=15;n_{iC}=30	0.947	0.947	0.053	0.054	0.341	0.343
	k_3=120	$n_{iT}=n_{iC}$=15	0.944	0.945	0.056	0.055	0.280	0.282
		n_{iT}=15;n_{iC}=30	0.948	0.950	0.052	0.050	0.241	0.243
3:1	k_1=30	$n_{iT}=n_{iC}$=15	0.965	0.968	0.035	0.033	0.568	0.578
		n_{iT}=15;n_{iC}=30	0.938	0.981	0.062	0.019	0.492	0.537
	k_2=60	$n_{iT}=n_{iC}$=15	0.966	0.971	0.035	0.030	0.401	0.408
		n_{iT}=15;n_{iC}=30	0.931	0.975	0.070	0.026	0.348	0.379
	k_3=120	$n_{iT}=n_{iC}$=15	0.964	0.967	0.036	0.033	0.284	0.289
		n_{iT}=15;n_{iC}=30	0.940	0.978	0.061	0.023	0.246	0.268

Table 12

Type I Error Rates of the FE and the Bonett Models (Chi-square df = 4)

			Average coverage		Type I error		Average width	
			FE	Bonett	FE	Bonett	FE	Bonett
$\sigma^2_T:\sigma^2_C$			Distribution: Chi-square df = 4					
			Simple contrast					
1:1	k_1=30	$n_{iT}=n_{iC}$=15	0.947	0.948	0.054	0.053	0.527	0.532
		n_{iT}=15;n_{iC}=30	0.943	0.943	0.057	0.057	0.455	0.459
	k_2=60	$n_{iT}=n_{iC}$=15	0.956	0.957	0.044	0.043	0.373	0.376
		n_{iT}=15;n_{iC}=30	0.952	0.951	0.048	0.050	0.322	0.324
	k_3=120	$n_{iT}=n_{iC}$=15	0.956	0.955	0.045	0.046	0.264	0.266
		n_{iT}=15;n_{iC}=30	0.947	0.947	0.054	0.053	0.228	0.229
3:1	k_1=30	$n_{iT}=n_{iC}$=15	0.969	0.974	0.032	0.027	0.544	0.556
		n_{iT}=15;n_{iC}=30	0.923	0.974	0.078	0.026	0.472	0.519
	k_2=60	$n_{iT}=n_{iC}$=15	0.972	0.976	0.028	0.024	0.384	0.393
		n_{iT}=15;n_{iC}=30	0.938	0.978	0.062	0.023	0.333	0.367
	k_3=120	$n_{iT}=n_{iC}$=15	0.973	0.975	0.027	0.025	0.272	0.278
		n_{iT}=15;n_{iC}=30	0.935	0.976	0.066	0.025	0.236	0.260
			Complex contrast					
1:1	k_1=30	$n_{iT}=n_{iC}$=15	0.951	0.951	0.050	0.050	0.558	0.564
		n_{iT}=15;n_{iC}=30	0.945	0.948	0.055	0.052	0.483	0.487
	k_2=60	$n_{iT}=n_{iC}$=15	0.951	0.951	0.049	0.049	0.395	0.399
		n_{iT}=15;n_{iC}=30	0.948	0.949	0.052	0.051	0.341	0.344
	k_3=120	$n_{iT}=n_{iC}$=15	0.954	0.953	0.046	0.048	0.279	0.282
		n_{iT}=15;n_{iC}=30	0.942	0.944	0.058	0.057	0.241	0.243
3:1	k_1=30	$n_{iT}=n_{iC}$=15	0.970	0.973	0.030	0.027	0.577	0.590
		n_{iT}=15;n_{iC}=30	0.933	0.974	0.068	0.027	0.500	0.551
	k_2=60	$n_{iT}=n_{iC}$=15	0.970	0.971	0.031	0.030	0.408	0.417
		n_{iT}=15;n_{iC}=30	0.939	0.985	0.061	0.016	0.354	0.390
	k_3=120	$n_{iT}=n_{iC}$=15	0.974	0.976	0.027	0.024	0.288	0.295
		n_{iT}=15;n_{iC}=30	0.929	0.976	0.072	0.024	0.250	0.275

Table 13

Type I Error Rates of the FE and the Bonett Models (Chi-square df = 8)

			Average coverage		Type I error		Average width	
			FE	Bonett	FE	Bonett	FE	Bonett
$\sigma^2_T:\sigma^2_C$			Distribution: Chi-square df = 8					
			<u>Simple contrast</u>					
1:1	k1=30	$n_{iT}=n_{iC}=15$	0.958	0.958	0.042	0.043	0.527	0.532
		$n_{iT}=15;n_{iC}=30$	0.952	0.955	0.048	0.046	0.455	0.459
	$k_2=60$	$n_{iT}=n_{iC}=15$	0.951	0.952	0.050	0.048	0.373	0.376
		$n_{iT}=15;n_{iC}=30$	0.950	0.953	0.050	0.048	0.322	0.324
	$k_3=120$	$n_{iT}=n_{iC}=15$	0.944	0.945	0.056	0.055	0.264	0.266
		$n_{iT}=15;n_{iC}=30$	0.961	0.960	0.040	0.040	0.228	0.229
3:1	$k_1=30$	$n_{iT}=n_{iC}=15$	0.977	0.978	0.023	0.022	0.561	0.579
		$n_{iT}=15;n_{iC}=30$	0.939	0.979	0.062	0.022	0.487	0.542
	$k_2=60$	$n_{iT}=n_{iC}=15$	0.953	0.971	0.048	0.029	0.324	0.409
		$n_{iT}=15;n_{iC}=30$	0.928	0.978	0.073	0.023	0.344	0.383
	$k_3=120$	$n_{iT}=n_{iC}=15$	0.961	0.965	0.040	0.036	0.280	0.289
		$n_{iT}=15;n_{iC}=30$	0.941	0.985	0.059	0.015	0.243	0.271
			<u>Complex contrast</u>					
1:1	$k_1=30$	$n_{iT}=n_{iC}=15$	0.956	0.957	0.045	0.044	0.557	0.564
		$n_{iT}=15;n_{iC}=30$	0.951	0.953	0.050	0.047	0.483	0.487
	$k_2=60$	$n_{iT}=n_{iC}=15$	0.954	0.955	0.046	0.046	0.396	0.399
		$n_{iT}=15;n_{iC}=30$	0.950	0.951	0.051	0.050	0.341	0.344
	$k_3=120$	$n_{iT}=n_{iC}=15$	0.951	0.950	0.050	0.050	0.280	0.282
		$n_{iT}=15;n_{iC}=30$	0.952	0.953	0.049	0.048	0.241	0.243
3:1	$k_1=30$	$n_{iT}=n_{iC}=15$	0.972	0.973	0.029	0.027	0.595	0.614
		$n_{iT}=15;n_{iC}=30$	0.940	0.975	0.060	0.025	0.517	0.575
	$k_2=60$	$n_{iT}=n_{iC}=15$	0.951	0.975	0.050	0.026	0.344	0.434
		$n_{iT}=15;n_{iC}=30$	0.937	0.979	0.063	0.022	0.365	0.407
	$k_3=120$	$n_{iT}=n_{iC}=15$	0.970	0.971	0.031	0.029	0.297	0.307
		$n_{iT}=15;n_{iC}=30$	0.945	0.980	0.055	0.020	0.258	0.288

Table 14

Type I Error Rates of the FE and the Bonett Models (Laplace)

			Average coverage		Type I error		Average width	
			FE	Bonett	FE	Bonett	FE	Bonett
$\sigma^2_T:\sigma^2_C$			Distribution: Laplace (0, 1)					
			Simple contrast					
1:1	k_1=30	$n_{iT}=n_{iC}$=15	0.953	0.953	0.047	0.048	0.527	0.532
		n_{iT}=15;n_{iC}=30	0.986	0.987	0.014	0.013	0.454	0.460
	k_2=60	$n_{iT}=n_{iC}$=15	0.949	0.951	0.052	0.050	0.373	0.376
		n_{iT}=15;n_{iC}=30	0.954	0.954	0.047	0.046	0.322	0.324
	k_3=120	$n_{iT}=n_{iC}$=15	0.950	0.949	0.051	0.051	0.264	0.266
		n_{iT}=15;n_{iC}=30	0.951	0.951	0.050	0.050	0.228	0.229
3:1	k_1=30	$n_{iT}=n_{iC}$=15	0.950	0.952	0.050	0.048	0.527	0.532
		n_{iT}=15;n_{iC}=30	0.976	0.992	0.024	0.009	0.455	0.497
	k_2=60	$n_{iT}=n_{iC}$=15	0.941	0.992	0.059	0.009	0.373	0.376
		n_{iT}=15;n_{iC}=30	0.906	0.953	0.095	0.048	0.323	0.351
	k_3=120	$n_{iT}=n_{iC}$=15	0.945	0.946	0.055	0.055	0.264	0.266
		n_{iT}=15;n_{iC}=30	0.910	0.947	0.090	0.053	0.228	0.248
			Complex contrast					
1:1	k_1=30	$n_{iT}=n_{iC}$=15	0.947	0.950	0.053	0.051	0.559	0.564
		n_{iT}=15;n_{iC}=30	0.982	0.985	0.018	0.015	0.482	0.488
	k_2=60	$n_{iT}=n_{iC}$=15	0.953	0.953	0.048	0.048	0.395	0.399
		n_{iT}=15;n_{iC}=30	0.954	0.953	0.047	0.047	0.341	0.344
	k_3=120	$n_{iT}=n_{iC}$=15	0.947	0.946	0.054	0.054	0.280	0.282
		n_{iT}=15;n_{iC}=30	0.951	0.951	0.050	0.049	0.241	0.243
3:1	k_1=30	$n_{iT}=n_{iC}$=15	0.947	0.951	0.053	0.049	0.559	0.565
		n_{iT}=15;n_{iC}=30	0.974	0.993	0.027	0.007	0.482	0.527
	k_2=60	$n_{iT}=n_{iC}$=15	0.950	0.951	0.051	0.049	0.395	0.399
		n_{iT}=15;n_{iC}=30	0.910	0.957	0.090	0.044	0.342	0.372
	k_3=120	$n_{iT}=n_{iC}$=15	0.948	0.948	0.053	0.052	0.280	0.282
		n_{iT}=15;n_{iC}=30	0.912	0.957	0.088	0.044	0.242	0.263

Table 11, Table 12 and Table 13 present the results for gamma and chi-square distributions with degrees of freedoms of 4 and 8, respectively. Similar outcomes were observed for these distributions. For the three skewed distributions, the Bonett model performed well with average coverage probabilities over or very close to 0.95, and Type I error rates less than or close to 0.055. The FE model produced similar results with the Bonett model in most conditions, except for when the within-study variances and within-study sample sizes were both unequal. When unequal within-study variances were paired with unequal within-study sample sizes, the FE model had coverage probabilities far below 0.95 across numbers of studies. To be more specific, the FE model could have an average coverage probability as low as 0.923 (see Table 12 when k_1=30), which was noticeably smaller than that of the Bonett model (0.974) under the same condition. Correspondingly, the Type I error rates of the FE model were inflated in these cases. The average widths of confidence intervals for the two methods were very similar, with the Bonett model giving slightly less precise contrast estimates by providing somewhat wider confidence intervals. The two models both provided narrower confidence intervals with increasing numbers of studies.

In summary, for gamma and chi-square distributions with degrees of freedom of 4 and 8, respectively, the Bonett model always performed well with Type I error rates that were stable and less than the α level as well as average coverage probabilities that were around 0.95. The FE model behaved worse than the Bonett model with inflated Type I error rates when both the within-study variances and sample sizes were unequal.

The results for a Laplace distribution are displayed in Table 14. Average coverage probabilities around 0.95 and good control of Type I error rates were observed for both of the FE and the Bonett models when the within-study variances were equal. Similar to results of the normal and other skewed distributions examined above, the behavior of the Bonett model was strong and remained unchanged across conditions.

However, the FE model, as observed again, had inflated Type I error rates when the within-study variances were paired with unequal within-study sample sizes. Under these conditions, a relatively large Type I error rate of approximately 0.1 was obtained for Laplace distribution, with corresponding average coverage probabilities markedly smaller than 0.95. Giving the consistent patterns of the two models under the 4 non-normal distributions examined here, these findings may be cautiously generalized to other skewed distributions.

Summarizing the results of the FE and the Bonett models across various raw data distributions, it was found that the Bonett model always performed well in contrast testing and provided contrast estimates as precise as those of the FE model. The FE model only performed as well as the Bonett model for studies with balanced designs or when the normality assumption was met. The combination of unequal within-study variances, unequal within-study sample sizes, and skewed data frequently produced inflated Type I error rates (as large as 0.1) for the FE model.

Comparing Power and Confidence Interval Widths

In this section, power and the average confidence interval widths of the FE and the Bonett models for the simple contrast are summarized for the five distributions across

the various data conditions examined. The results are reported in Table 15 through Table 19.

Table 15 presents the results of both simple and complex contrasts for normally distributed data. The raw data were simulated to allow the power value of 0.8 for the classic FE model under the condition of equal within-study sample sizes, equal within-study variances, and k = 30 studies. The empirical power under this particular condition was 0.796 which was close to 0.8 as expected. Power tended to increase with the increasing numbers of studies. Compared to unequal within-study sample size conditions, power was larger under equal within-study sample size conditions. In general, power under unequal within-study variance conditions was lower than that under conditions of equal within-study variances. Both the classic FE model and the Bonett model failed to achieve satisfactory power of 0.8 for a small number of studies (k=30) under unequal within-study variance conditions.

As shown in Table 15, when the within-study variances were equal, the classic FE model and the Bonett model had similar power and average confidence interval widths across different numbers of studies for either equal or unequal within-study sample sizes. However, for unequal within-study variances, the power and confidence interval widths of the two models were only similar when the within-study sample sizes were equal; the Bonett model showed noticeably larger power and smaller confidence interval widths when the within-study sample sizes were unequal. This pattern was consistently observed for k = 30, k = 60, and k = 120.

The complex contrast compared the mean effect sizes among the 3 levels of a moderator for k studies. The power value was smaller compared to that under the same study conditions for the simple contrast, especially when k was small (i.e. k = 30 or k = 60). Comparing the power between the classic FE and the Bonett models, the same patterns as that for the simple contrast were observed. For example, the Bonett model had somewhat larger power and smaller confidence interval widths when the within-study variances were unequal and the within-study sample sizes were unequal. The patterns of comparing the power between the FE and the Bonett models for complex contrasts were almost identical to that for simple contrasts for data from all distributions. Thus, the power results for complex contrasts are only reported for normal distribution as an example.

Table 16 shows the power values of the simple contrast for a gamma distribution. In general, under the same study conditions, power values for a gamma distribution were slightly higher than that under the normal distribution. Again, power increased when the number of studies increased. Power values were smaller under unequal within-study sample size conditions than these under equal within-study sample size conditions. Power values under unequal within-study variances conditions were lower than that under conditions of equal within-study variances. Both the classic FE and the Bonett models had power lower than 0.8 for k = 30 under unequal within-study variance conditions. With inflated Type I error rates, the FE model did not perform well when the within-study sample sizes were unequal and the within-study variances were unequal, especially for small sample sizes, in which case the power was low.

Table 17 and Table 18 present the power values of the simple contrast for chi-square distributions with 4 and 8 degrees of freedoms, respectively. Power values under the two distributions were low in general compared to those under a normal distribution. Using the between-level mean effect size difference obtained from normal distribution as the bench mark value, neither of the two chi-square distributions had satisfactory power (>0.8) across conditions studied. Comparing the power between the classic FE and the Bonett models, similar values were observed when the within-study variances were equal. When the within-study variances were unequal and the within-study sample sizes were unequal, the Bonett model had slightly higher power values; whereas the classic FE model had not only lower power, but also inflated Type I error rates.

The power values of the simple contrast under a Laplace distribution is shown in Table 19. When the within-study variances were equal, the classic FE and the Bonett models had comparable power values. They both achieved satisfactory power (> 0.8) when the number of studies was k = 60 and larger. When the within-study variances were unequal, both models had lower power values in general compared to that under equal within-study variances. For example, satisfactory power (> 0.8) was achieved only for k = 120. Consistent with the patterns observed under the other skewed distributions, when the within-study sample sizes were unequal, compared to the Bonett model, the classic FE model had lower power as well as inflated Type I error rates for k = 60 and k = 120.

In summary, under various conditions examined across different data distributions, for both classic FE and the Bonett models, power values showed the following patterns: 1) increased with the increasing sample sizes 2) were larger for

unequal within-study sample sizes than for equal within-study sample sizes 3) were generally larger for equal within-study variances than for unequal within-study variances.

The mean effect size difference between levels of a moderator, which allowed a simple contrast with power of 0.8 for normally distributed data, was used across all distributions. The power values were the largest for a gamma distribution, followed by a normal distribution, and the Laplace distribution. Chi-square distributions with 4 and 8 degrees of freedom, respectively, had the lowest power values.

When comparing between the FE and the Bonett models, similar power values were observed when the within-study variances were equal, across normal and skewed distributions. However, when the raw data was skewed, under the conditions of unequal within-study variances combined with unequal within-study sample sizes, the FE model had not only lower power but also inflated Type I error rates than the Bonett model. It is important to note that comparing the power values of two tests when one (or both) has an inflated Type I error rates (even for modest inflation) must be done quite cautiously, since a model with inflated Type I error rates is expected to have larger power values.

Summary

To compare the performance of the classic FE and the Bonett model, this study examined the Type I error rates and power of the two models under various conditions in meta-analysis across different data distributions.

To sum up, the Bonett model had good control of Type I error rates, while the classic FE model tended to have inflated Type I error rates for unequal within-study sample sizes and unequal within-study variances, when the raw data distribution deviated

from normal. Under the conditions of skewed raw data in studies with unbalanced designs, the classic FE model also showed noticeably lower power. The comparable widths of confidence intervals for contrast estimates suggested similar precision of estimation for both models.

Integrating the results from Type I error rates and power analysis leads to the conclusion that the Bonett model seems to outperform the classic FE model, showing robust Type I error rates and larger statistical power even for non-normally distributed data. The Bonett model was also robust to the violation of normality for unbalanced designs and was clearly superior to the FE model under these conditions.

Table 15

Power Comparison of the FE and the Bonett Models (Normal)

Simple contrast			Power		Average width	
			FE	Bonett	FE	Bonett
σ^2_T:σ^2_C						
1:1	k=30	nt=nc=15	0.796	0.795	0.529	0.534
		nt=15;nc=30	0.899	0.897	0.457	0.462
	k=60	nt=nc=15	0.977	0.976	0.374	0.378
		nt=15;nc=30	0.994	0.994	0.323	0.326
	k=120	nt=nc=15	1.000	1.000	0.265	0.267
		nt=15;nc=30	1.000	1.000	0.229	0.231
3:1	k=30	nt=nc=15	0.501	0.508	0.528	0.534
		nt=15;nc=30	0.588	0.699	0.455	0.420
	k=60	nt=nc=15	0.802	0.801	0.374	0.378
		nt=15;nc=30	0.867	0.935	0.362	0.297
	k=120	nt=nc=15	0.978	0.978	0.264	0.267
		nt=15;nc=30	0.995	0.997	0.228	0.210
Complex contrast			Power		Average width	
			FE	Bonett	FE	Bonett
σ^2_T:σ^2_C						
1:1	k=30	nt=nc=15	0.519	0.511	0.560	0.565
		nt=15;nc=30	0.602	0.601	0.484	0.488
	k=60	nt=nc=15	0.802	0.802	0.396	0.400
		nt=15;nc=30	0.901	0.898	0.342	0.345
	k=120	nt=nc=15	0.970	0.969	0.280	0.283
		nt=15;nc=30	0.992	0.992	0.242	0.244
3:1	k=30	nt=nc=15	0.292	0.289	0.560	0.565
		nt=15;nc=30	0.287	0.417	0.482	0.445
	k=60	nt=nc=15	0.491	0.493	0.396	0.400
		nt=15;nc=30	0.578	0.703	0.341	0.315
	k=120	nt=nc=15	0.788	0.787	0.280	0.283
		nt=15;nc=30	0.881	0.936	0.241	0.222

Table 16

Power Comparison of the FE and the Bonett Models (Gamma (1, 1))

$\sigma^2_T:\sigma^2_C$			Power		Average width	
			FE	Bonett	FE	Bonett
1:1	k=30	nt=nc=15	0.850	0.851	0.530	0.536
		nt=15;nc=30	0.906	0.902	0.457	0.460
	k=60	nt=nc=15	0.983	0.982	0.375	0.379
		nt=15;nc=30	0.997+	0.996+	0.323	0.325
	k=120	nt=nc=15	1.000	1.000	0.265	0.168
		nt=15;nc=30	1.000	1.000	0.229	0.230
3:1	k=30	nt=nc=15	0.553	0.526	0.541	0.553
		nt=15;nc=30	0.617+	0.716	0.464	0.434
	k=60	nt=nc=15	0.846	0.834	0.382	0.391
		nt=15;nc=30	0.925+	0.952	0.328	0.307
	k=120	nt=nc=15	0.998	0.996	0.271	0.277
		nt=15;nc=30	0.999+	1.000	0.232	0.217

Note. + suggests inflated Type I error rate (average coverage < 0.945) under this condition

Table 17

Power Comparison of the FE and the Bonett Models (Chi-square df = 4)

$\sigma^2_T:\sigma^2_C$			Power		Average width	
			FE	Bonett	FE	Bonett
1:1	k=30	nt=nc=15	0.168	0.163	0.527	0.532
		nt=15;nc=30	0.235+	0.239+	0.455	0.459
	k=60	nt=nc=15	0.317	0.317	0.373	0.376
		nt=15;nc=30	0.382	0.379	0.322	0.324
	k=120	nt=nc=15	0.520	0.515	0.264	0.266
		nt=15;nc=30	0.625	0.624	0.228	0.229
3:1	k=30	nt=nc=15	0.084	0.077	0.546	0.560
		nt=15;nc=30	0.071+	0.126	0.467	0.438
	k=60	nt=nc=15	0.155	0.142	0.386	0.396
		nt=15;nc=30	0.130+	0.224	0.330	0.310
	k=120	nt=nc=15	0.265	0.248	0.273	0.280
		nt=15;nc=30	0.290+	0.400	0.234	0.219

Note. + suggests inflated Type I error rate (average coverage < 0.945) under this condition

Table 18

Power Comparison of the FE and the Bonett Models (Chi-square df = 8)

$\sigma^2_T:\sigma^2_C$			Power		Average width	
			FE	Bonett	FE	Bonett
1:1	k=30	nt=nc=15	0.099	0.096	0.527	0.532
		nt=15;nc=30	0.121	0.119	0.455	0.459
	k=60	nt=nc=15	0.172	0.168	0.373	0.376
		nt=15;nc=30	0.202	0.203	0.322	0.325
	k=120	nt=nc=15	0.294	0.293	0.264	0.266
		nt=15;nc=30	0.357+	0.360	0.228	0.230
3:1	k=30	nt=nc=15	0.052	0.050	0.563	0.582
		nt=15;nc=30	0.033+	0.072	0.479	0.453
	k=60	nt=nc=15	0.080	0.073	0.398	0.411
		nt=15;nc=30	0.054+	0.106	0.338	0.321
	k=120	nt=nc=15	0.136	0.125	0.281	0.291
		nt=15;nc=30	0.107+	0.182	0.239	0.227

Note. + suggests inflated Type I error rate (average coverage < 0.945) under this condition

Table 19

Power Comparison of the FE and the Bonett Models (Laplace)

$\sigma^2_T:\sigma^2_C$			Power		Average width	
			FE	Bonett	FE	Bonett
1:1	k=30	nt=nc=15	0.513	0.512	0.528	0.533
		nt=15;nc=30	0.644	0.648	0.456	0.459
	k=60	nt=nc=15	0.819	0.817	0.374	0.377
		nt=15;nc=30	0.904	0.907	0.323	0.325
	k=120	nt=nc=15	0.980	0.982	0.264	0.267
		nt=15;nc=30	0.998	0.998	0.228	0.230
3:1	k=30	nt=nc=15	0.284	0.281	0.528	0.533
		nt=15;nc=30	0.311+	0.434	0.455	0.420
	k=60	nt=nc=15	0.527	0.530	0.373	0.377
		nt=15;nc=30	0.591+	0.717	0.322	0.297
	k=120	nt=nc=15	0.820	0.817	0.264	0.267
		nt=15;nc=30	0.904+	0.949	0.227	0.210

Note. + suggests inflated Type I error rate (average coverage < 0.945) under this condition

Discussion

Conclusion

Meta-analysis is one of the most widely used research synthesis methods in the social sciences. Within meta-analysis, one of the most popular models to combine and compare effect sizes is the traditional fixed-effects model developed by Hedges and Olkin (1985), although some of its assumptions have received criticism.

In response to the criticisms, Bonett (2009) proposed a modified fixed effects model which does not require the key assumptions of the traditional fixed-effects model such as homogeneity of variance within studies and equality of population effect sizes across studies. Bonett provided evidence of the strong performance of his model under a limited set of conditions for normally distributed data.

In attempting to identify the model to be recommended, this simulation study empirically examined the behavior of the Bonett model and the traditional fixed-effect model in estimating and testing contrasts among levels of a moderator in meta-analysis under a broader set of realistic data conditions. The conditions considered include number of studies, sample sizes, raw data distributions, and within-study variances. The average coverage probabilities, Type I error rates, average coverage widths, and power of the two models were compared under each condition.

Depending on the condition, the two models performed differently. With respect to the Type I error rates, both the traditional fixed-effects and Bonett models behaved similarly for normally distributed data. The Bonett model had somewhat better control of Type I error rates when unequal within-study sample sizes were combined with unequal

within-study variances. The contrast values estimated from both models were comparable in precision, with minor advantages for the traditional fixed-effects model.

The Bonett model, however, showed significant advantages over the traditional fixed-effects model in controlling for Type I error rates in contrast testing while providing precise estimation for non-normal data and studies with variance heterogeneity and unequal sample sizes. Disappointingly, under the same conditions the traditional fixed-effect model could have inflated Type I error rates as large as 0.095.

With respect to power, the mean effect size difference between levels of a moderator, for a simple contrast with a power of 0.8 for normally distributed data, was used across all distributions. Both the Bonett and the fixed-effects models had the largest power values for the gamma distribution. Next in line were the normal and Laplace distributions. The lowest power values were observed for Chi-square distributions with 4 and 8 degrees of freedom.

The Bonett model worked equally well as the FE model with similar power values when the within-study variances were equal, across normal and skewed distributions. It outperformed the FE model for skewed raw data by showing larger power under the conditions of unequal within-study variances combined with unequal within-study sample sizes.

An important finding was that the non-normal data had a significant impact on the performance of the classic FE model. Under the condition of skewed raw data combined with unequal within-study sample sizes and unequal within-study variances, the classic FE model tended to have inflated Type I error rates as well as low power.

These results lead to the conclusion that the Bonett model has significant advantages over the traditional fixed-effects model in controlling for Type I error rates and providing larger power in contrast testing while providing precise estimation even for non-normal data and studies with variance heterogeneity and unequal sample sizes. Disappointingly, under the same conditions the traditional fixed-effects model not only produced inflated Type I error, a pattern which was amplified by skewed raw data, but also lower power values.

The good performance of the Bonett model for skewed data and unequal variances was consistent with Chen and Peng's (2012) findings. In their study on the Bonett model for a single standardized linear contrast of means in one-way ANOVA design, a major discovery was that the coverage probability of Bonett's model was not impacted by moderate violation of the assumptions of normally distributed data and equal variances.

The impact of skewed data on the behavior of the traditional FE model was also consistent with previous findings of studies which examined meta-analytic models. For example, when evaluating several techniques to estimate a meta-analytic moderator, Steel et al. (2002) concluded that under the condition of skewed data and heterogeneity of effect sizes, these techniques failed to converge on the true moderator effect and substantially underestimated the effect.

Another example is Harwell's (1997) simulation study on the performance of the Q test for the fixed-effects meta-analytical model. This study discovered that Q test had Type I error rates close to the nominal value and adequate power only for ideal data

condition, such as normally distributed and homoscedastic data and equal within-study sample sizes. However, skewed data combined with unequal variances tended to produce inflated Type I error rates for the Q test, with the magnitude of inflation increasing with the skew of the distribution and variance heterogeneity. Low power values when unequal variances were paired with unequal sample sizes. The results of the current study reaffirm the importance of the raw data distribution in meta-analytic model performance.

Recommendation

This study supported Bonett's (2009) conclusions and showed the excellent and stable performance of his modified fixed-effects model across various realistic data conditions in testing and estimating contrasts in meta-analysis. The simulation results provided strong evidence of the robustness of the Bonett model in the presence of unequal within-study sample sizes, unequal within-study variances, and non-normal raw data.

The Bonett model stands out with several good characteristics compared to the classic fixed-effects model. First, it does not assume effect-size homogeneity. In meta-analytic practice, data collected across studies published at different times often show certain degrees of heterogeneity. When the homogeneity assumption cannot be satisfied the Bonett model can be used with confidence, unlike the traditional fixed-effects model.

Second, the Bonett model provides better control of Type I error rates and larger statistical power, especially for non-normal data with unbalanced designs and unequal within-study variances. Examples of skewed data are often present in meta-analytic research (Sanchez-Meca and Marin-Martinez, 1998a). Because the Bonett model

performs better when data are skewed, results generated from this model seem to be more trustworthy.

Third, the Bonett model results in estimates as precise as the fixed-effects model. Confidence intervals provide readers with more information to interpret the size of the effect. It is highly recommended that point estimates as well as confidence intervals of effect sizes should both be provided in meta-analytic research (Bonett, 2009). The confidence intervals generated by the Bonett model have comparable confidence interval widths with the traditional fixed-effects model, suggesting their similar abilities to estimate effect sizes precisely.

Fourth, the Bonett model is easy to use and has fewer computational demands. As shown in the literature review, the formulas to compute effect sizes and obtain contrasts in meta-analysis are simpler for the Bonett model than the traditional fixed-effects model.

In sum, the Bonett model is as good as, and appears to be better than, the traditional fixed-effects model and thus serves as a good alternative. With these good features, the Bonett model surpasses the traditional fixed-effect model, which only performs satisfactorily when the data are normally distributed with both equal within-study variances and sample sizes. As a result, researchers and policy makers whose work relies on meta-analytic results should consider choosing the Bonett model rather than the traditional fixed-effects model.

Limitations and Future Research

This simulation study compared the performance of the Bonett model and the classic fixed-effect model under data conditions chosen by reviewing a set of published

meta-analyses. However, they may not be representative of all meta-analyses. Technically, the results are restricted to the factors examined and thus have limited generalizability to other data conditions that arise in meta-analytic practice. In the future, the Bonett model should be further examined under expanded conditions with other factors.

References

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research,* 58(2), 151-179.

Aguinis, H., Sturman, M.C., Pierce, C.A., (2008).Comparison of Three Meta-Analytic Procedures for Estimating Moderating Effects of Categorical Variables. *Organizational Research*, 11(1), 9-34.

Banda, D. R. & Therrien, W. J. (2008). A teachers' guide to meta-analysis. *Teaching Exceptional Children,* 41, 66-71.

Berkey, C. S., Hoaglin,D.C., Antczak-Bouckoms, A., Mosteller, F., Colditz, G.A.(1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, 17(22), 2537–2550.

Bohning, D., Malzahn, U., Dietz, E., Schlattmann, P. (2002). Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. *Biostatistics*, 3, 4, 445–457

Bonett, D.G (2008a). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13(2)*, 99–109.

Bonett, D.G (2008b). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods, 13(3)*, 173–181.

Bonett, D.G (2009a). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods, 14 (3)*, 225–238.

Bonett, D.G (2009b). Estimating standardized linear contrasts of means with desired precision. *Psychological Methods, 14 (1),* 1–5.

Bonett, D.G (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15 (4)*, 368–385.

Brannick, M.T., Yang.L., Cafri,G. (2010). Comparison of Weights for Meta-Analysis of r and d Under Realistic Conditions. *Organizational Research Methods.* 000(00) 1-21.

Brockwell, S. E., & Gordon, R. I. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine, 20,* 825–840.

Cafri, G, Kromrey, J.D., Brannick, M.T. (2010). A Meta-Meta-Analysis: Empirical Review of Statistical Power, Type I Error Rates, Effect Sizes, and Model Selection

of Meta-Analyses Published in Psychology. *Multivariate Behavioral Research,* 45:239–270

Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. *Evaluation &the Health Profession*. 2002; 25:12-37.

Chang, L. (1992). A power analysis of the test of homogeneity in effect-size meta-analysis. Unpublished doctoral dissertation, Michigan State University

Li-Ting Chen & C.Y. Joanne Peng (2012) *Confidence Interval Estimations for Standardized Linear Contrasts of Means: The One Way Fixed Effects Between Subjects Univariate Case*. Paper presented at the annual meeting of American Educational Research Association. Vancouver, British Columbia, Canada

Cooper, H. (1997). Some finer points in meta-analysis. In M. Hunt (ed.), *How Science Takes Stock: The Story of Meta-analysis*. New York: Russell Sage Foundation, pp. 169-81.

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

Cooper, H. & Hedges, L. V., (2009). The *Handbook of Research Synthesis*, 2nd Edition. New York: Russel Sage Foundation.

Cornwell, J. M., & Ladd, R. T. (1993). Power and accuracy of the Schmidt and Hunter meta-analytic procedures. *Educational and Psychological Measurement*, 53, 877-895.

Cornwell, J. M. (1993). Monte Carlo comparison of three tests for homogeneity of independent correlations. *Educational & Psychological Measurement, 53,* 605–618.

DeCoster,J. (2004). Meta-analysis notes. Available from: http://www.stat-help.com/notes.html

Durlak, J.A., & Lipsey, M.W. (1991). A practitioner's guide to meta-analysis .*American Journal of Community Psychology, 19(3),* 291-332.

Dunst, C.J., Hamby, D.W., & Trivette, C.M. (2004). Guidelines for calculating effectsizes for practice-based research synthesis. *Centerscope: Evidence-Based Approaches to Early Childhood Development,* 3,1-10.

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

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver and Boyd.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed and random-effects methods. *Psychological Methods*, 6, 161–180.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population effect sizes vary? *Psychological Methods*, 10, 444-467.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality Indicators for group experimental and quasi-experimental research in special education. *Exceptional Children,* 71, 149-164.

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

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

Hall, S.M & Brannick, M.T. (2002). Comparison of Two Random-Effects Methods of Meta-Analysis. *Journal of Applied Psychology*, 87 (2), 377-389.

Hardy,R.J,Thompson, S.G (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine.* 17,841-856.

Hartley, H. O., & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika, 54, 93–108.

Harwell, M. (1997). An empirical study of Hedge's homogeneity test. *Psychological Methods, 2,* 219–231.

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

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

Hedges, L. V. (1982a).Fitting categorical models to effect sizes from a series of experiments, *Journal of Educational and Behavioral Statistics*, 7(2), 119-137.

Hedges, L. V. (1982b). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, 92, 490–499.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin,* 93, 388–395.

Hedges, L. V. & Valentine, J. C. (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 279-293). New York: Russell Sage Foundation.

Higgins, J. P. T., and S. G. Thompson. 2004. Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine,* 23, 1663–1682.

Huedo-Medina TB & Sanchez-Meca J. (2006). Assessing heterogeneity in metaanalysis: Q statistic or I squared index? *Psychol Methods*,1, 193- 206.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Mela-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Beverly Hills, CA: Sage

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,* 275–292.

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

Hyde, J.S., Fennema, E., Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin.* 107(2), 139-155.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools,* 44, 483-493.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analysis approaches. *Journal of Applied Psychology. 80,* 94-106.

Kisamore, J. L. (2003). Validity generalization and transportability: An investigation of distributional assumptions of random-effects meta-analytic methods. Unpublished doctoral dissertation, University of South Florida.

Konstantopoulos, S. & Hedges, L.V. (2009). Analyzing effect sizes: Fixed effects models. In *The handbook of research synthesis and meta-analysis* (2nd ed., H.

Cooper, L.V. Hedges, & J.C. Valentine, Eds., pp. 279-293). New York: Sage Foundation.

Lipsey, M.W. and Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment confirmation from meta-analysis. *American Psychologist.* 48(12).1181-1209.

Lipsey, M.W. and Wilson, D.B. (2001). Practical meta-analysis. Applied social research methods, 49. Thousand Oaks, CA: Sage Publications, Inc.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Miller, N., & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 457–483). New York: Russell Sage Foundation.

Mullen, B., Salas, E., & Miller, N. (1991). Using meta-analysis to test theoretical hypotheses in social psychology. *Personality and Social Psychology Bulletin, 17,* 258–264.

Murlow, C. D. (1995). Rationale for systematic reviews. In I. Chalmers & D. G. Altman (Eds.), *Systematic reviews* (pp. 1–8). London: BMJ Publishing Group.

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

Oswald, F.L. & Johnson, J.W. (1998). On the Robustness, Bias, and Stability of Statistics From Meta-Analysis of Correlation Coefficients: Some Initial Monte Carlo Findings. *Journal of Applied Psychology*, 83(2), 164-178.

Oswald, F. L. (1999). On deriving validity generalization and situational specificity from meta-analysis: A conceptual review and some empirical findings (Doctoral dissertation, University of Minnesota). *Dissertation Abstracts International, 60*, 399.

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*, 354–379.

Petticrew, M. (2001). Systematic reviews from astronomy to zoology: myths and misconceptions. *British Medical Journal*. 322 (13) 98–101.

Quintana, S.M., & Minami, T. (2006). Guidelines for meta-analyses of counseling psychology research. *The Counseling Psychologist,* 34, 839-877.

Raudenbush, S.W. (2009). Analyzing effect sizes: Random-effects models. In *The handbook of research synthesis and meta-analysis* (2nd ed., H. Cooper, L.V. Hedges, & J.C. Valentine, Eds., pp. 295-316). New York: Sage Foundation.

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

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin,* 118(2), 183-192.

Rosenthal, D.A., Hoyt, W.T., Ferrin, J.M., Miller, S., Cohen. N.D. (2006). ―Advanced Methods in Meta-Analytic Research: Applications and Implications for Rehabilitation Counseling Research. *Rehabilitation Counseling Bulletin.* 49(4) (Summer): 234–46.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, 71, 302–310.

Sagie, A., & Koslowsky, M. (1993). Detecting moderators with meta-analysis: An evaluation and comparison of techniques. *Personnel Psychology,* 46, 629-640.

Sanchez-Meca,J & Marin-Martinez,F. (1998a). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Jounal of Mathematical and Staitistical Psychology*. 51, 311-326.

Sanchez-Meca,J & Marin-Martinez,F. (1998b). Weighting by Inverse Variance or by Sample Size in Meta-Analysis: A Simulation Study. *Educational and Psychological Measurement*. 58, 211.

Schmidt, F.L., Oh, I, Hayes, T.L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62(1),* 97-128.

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

Schulze, R. (2004). Meta-analysis: a comparison of approaches. Cambridge, MA: Hogrefe & Huber.

Shuster, J.J (2009). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine, 29(12),* 1259-1265.

Sidik K. & Jonkman J.N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, 26, 1964–1981.

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology*, 72, 3-9.

Stanovich, K. E. (2004). Metarepresentation and the great cognitive divide. *Journal of Clinical Psychology*, 60, 1263-1264.

Steel, P.D., Kammeyer-Mueller, J.D. (2002). Comparing Meta-Analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*. 87(1), 96–111

Sutton, A.J, Song, F, Gilbody, S.M, & Abrams, K.R (2000) Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research*, 9, 421

Suri H. A critique of contemporary methods of research synthesis. *Post-Script* 2000;1:49-55. http://www.edfac.unimelb.edu.au/student/insight/postscriptfiles/vol1/suri.pdf

Takkouche B, Cadarso-Suarez C. & Spiegelman D (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Jounal of Epidemiology*, 150:206-215.

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. Biometrics, 59, 254–262.

Van den Noortgate W, Onghena P. (2003a). Estimating the mean effect size in meta-analysis: Bias, precision, and mean squared error of different weighting methods. *Behavior Research Methods, Instruments & Computers*, 35, 504-511.

Van den Noortgate, W. & Onghena,P (2003b). Multilevel Meta-Analysis: A Comparison with Traditional Meta-Analytical Procedures. *Educational and Psychological Measurement.* 63(5) 765-790.

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *60*, 29–60.

Appendix Sample R codes

```
# Sample 1
# Type 1 error;
# Normal distribution N (0, 1);
# Number of study k = 30;
# Equal within-study sample sizes: nc = nt =15;
# Equal within-study variance 1:1.

   #==================================;
   #         The Hedges & Olkin FE model
   #==================================;
   set.seed(100)
   nsim<-2000; nsamc<-15; nsamt<-15; nstudy<-30;
   ut<-matrix(0,nsim,nsamt); uc<-matrix(0,nsim,nsamc);
   sigc<-matrix(0,nsim,nsamc);sigt<-matrix(0,nsim,nsamt);
   si<-matrix(0,nsim,nsamc);
   di<-matrix(0,nsim,nstudy);
   vi<-matrix(0,nsim,nstudy);

   for (i in 1:2000)
      {
         ## compute effect size by data
         for (j in 1:30)
         {
            nc=15
            nt=15
        ut<-rnorm(nt,0,1)
        uc<-rnorm(nc,0,1)
       sigc<-var(uc)
       sigt<-var(ut)
       si<- ((nc-1)*sigc+(nt-1)*sigt)/(nc+nt-2)

# effect size in study i
         di[i,j]<- (mean(ut)-mean(uc))/sqrt(si)
   vi[i,j]<-(nt+nc)/(nt*nc)+di[i,j]*di[i,j]/(2*(nt+nc))
   wi<-1/vi
         }
      }
   # combine effect size from multiple studies k
   # di.1 is the (combined) effect size for group 1, with variance vd.1;

   di.1<-di[,1:15];di.2<-di[,16:30];
```

```
di.3<-di[,1:10];di.4<-di[,11:20]; di.5<-di[,21:30];
wi.1<-wi[,1:15];wi.2<-wi[,16:30];
wi.3<-wi[,1:10];wi.4<-wi[,11:20]; wi.5<-wi[,21:30];

d.1<-rowSums(wi.1*di.1)/rowSums(wi.1)
d.2<-rowSums(wi.2*di.2)/rowSums(wi.2)
d.3<-rowSums(wi.3*di.3)/rowSums(wi.3)
d.4<-rowSums(wi.4*di.4)/rowSums(wi.4)
d.5<-rowSums(wi.5*di.5)/rowSums(wi.5)

vd.1<-1/rowSums(wi.1)
vd.2<-1/rowSums(wi.2)
vd.3<-1/rowSums(wi.3)
vd.4<-1/rowSums(wi.4)
vd.5<-1/rowSums(wi.5)

# simple contrast: (d.1-d.2)
# 95% CI
ci.su<-d.1-d.2+1.96*sqrt(vd.1+vd.2)
ci.sl<-d.1-d.2-1.96*sqrt(vd.1+vd.2)
wid.s<-abs(ci.su-ci.sl)
ave.wds<-ave(wid.s)
t1.s<-sum(ci.sl>=0 & ci.su>=0)+sum(ci.sl<=0 & ci.su<=0)

# complex contrast: (d.3-(d.4+d.5)/2)
# 95% CI:
ci.cu<-(d.3-(d.4+d.5)/2)+1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
ci.cl<-(d.3-(d.4+d.5)/2)-1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
wid.c<-abs(ci.cu-ci.cl)
ave.wdc<-ave(wid.c)
t1.c<-sum(ci.cl>=0 & ci.cu>=0)+sum(ci.cl<=0 & ci.cu<=0)


#==================================;
#        The Bonett's model
#==================================;
set.seed(100)
nsim<-2000; nsamc<-15; nsamt<-15; nstudy<-30;
ut<-matrix(0,nsim,nsamt); uc<-matrix(0,nsim,nsamc);
df.t<-matrix(0,nsim,nsamt); df.c<- matrix (0,nsim,nsamc);
sigc<-matrix(0,nsim,nsamc);sigt<-matrix(0,nsim,nsamt);
si<-matrix(0,nsim,nsamc);
di<-matrix(0,nsim,nstudy);
```

```
vi<-matrix(0,nsim,nstudy);
bi<-matrix(0,nsim,nstudy);

# data simulation
for (i in 1:2000)
   {
      ##compute effect size by data
      for (j in 1:30)
      {
         nc=15
         nt=15
      ut<-rnorm(nt,0,1)
      uc<-rnorm(nc,0,1)
     sigc=var(uc)
     sigt=var(ut)
     si=(sigc+sigt)/2
 df.t<-nt-1
df.c<-nc-1
 #effect size in study j
       di[i,j]=(mean(ut)-mean(uc))/sqrt(si)

vi[i,j]=di[i,j]*di[i,j]*(sigt*sigt/df.t+sigc*sigc/df.c)/(8*si*si)+(sigt/df.t+sigc/df.c)/si
      bi[i,j] <-1-3/(4*(nc+nt)-9)
       }
   }
# combine effect sizes from multiple studies k
ng.1<-15; ng.2<-15;
ng.3<-10; ng.4<-10; ng.5<-10;
di.1<-di[,1:15];di.2<-di[,16:30];di.3<-di[,1:10];di.4<-di[,11:20]; di.5<-di[,21:30];
bi.1<-bi[,1:15];bi.2<-bi[,16:30];bi.3<-bi[,1:10];bi.4<-bi[,11:20]; bi.5<-bi[,21:30];
vi.1<-vi[,1:15];vi.2<-vi[,16:30];vi.3<-vi[,1:10];vi.4<-vi[,11:20]; vi.5<-vi[,21:30];

d.1<-rowSums(di.1*bi.1)/ng.1
d.2<-rowSums(di.2*bi.2)/ng.2
d.3<-rowSums(di.3*bi.3)/ng.3
d.4<-rowSums(di.4*bi.4)/ng.4
d.5<-rowSums(di.5*bi.5)/ng.5

vd.1<-rowSums(bi.1*bi.1*vi.1)/(ng.1*ng.1)
vd.2<-rowSums(bi.2*bi.2*vi.2)/(ng.2*ng.2)
vd.3<-rowSums(bi.3*bi.3*vi.3)/(ng.3*ng.3)
vd.4<-rowSums(bi.4*bi.4*vi.4)/(ng.4*ng.4)
vd.5<-rowSums(bi.5*bi.5*vi.5)/(ng.5*ng.5)
```

```
   # simple contrast: (d.1-d.2)
   # 95% CI
   ci.su<-d.1-d.2+1.96*sqrt(vd.1+vd.2)
   ci.sl<-d.1-d.2-1.96*sqrt(vd.1+vd.2)
   wid.s<-abs(ci.su-ci.sl)
   ave.wds<-ave(wid.s)
   t1.s<-sum(ci.sl>=0 & ci.su>=0)+sum(ci.sl<=0 & ci.su<=0)

   #complex contrast: (d.3-(d.4+d.5)/2)
   #95% CI:
   ci.cu<-(d.3-(d.4+d.5)/2)+1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
   ci.cl<-(d.3-(d.4+d.5)/2)-1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
   wid.c<-abs(ci.cu-ci.cl)
   ave.wdc<-ave(wid.c)
   t1.c<-sum(ci.cl>=0 & ci.cu>=0)+sum(ci.cl<=0 & ci.cu<=0)
```

 # Sample 2

```
# Type 1 error;
# Normal distribution N (0, 1);
# Number of study k = 30;
# Equal within-study sample sizes: nc = nt = 15;
# Unequal within-study variance 1:3.

   #==================================;
   #         The Hedges & Olkin FE model
   #==================================;
   set.seed(100)
   nsim<-2000; nsamc<-15; nsamt<-30; nstudy<-30;
   ut<-matrix(0,nsim,nsamt); uc<-matrix(0,nsim,nsamc);
   sigc<-matrix(0,nsim,nsamc);sigt<-matrix(0,nsim,nsamt);
   si<-matrix(0,nsim,nsamc);
   di<-matrix(0,nsim,nstudy);
   vi<-matrix(0,nsim,nstudy);

   for (i in 2:1000)
      {
         ## compute effect size by data
         for (j in 1:30)
         {
            nc=15
            nt=30
         ut<-rnorm(nt,0,sqrt(3))
         uc<-rnorm(nc,0,1)
```

```
      sigc<-var(uc)
      sigt<-var(ut)
      si<- ((nc-1)*sigc+(nt-1)*sigt)/(nc+nt-2)
# effect size in study i
       di[i,j]<- (mean(ut)-mean(uc))/sqrt(si)
vi[i,j]<-(nt+nc)/(nt*nc)+di[i,j]*di[i,j]/(2*(nt+nc))
wi<-1/vi
       }
   }
# combine effect sizes from multiple studies k
di.1<-di[,1:15];di.2<-di[,16:30];
di.3<-di[,1:10];di.4<-di[,11:20]; di.5<-di[,21:30];
wi.1<-wi[,1:15];wi.2<-wi[,16:30];
wi.3<-wi[,1:10];wi.4<-wi[,11:20]; wi.5<-wi[,21:30];

d.1<-rowSums(wi.1*di.1)/rowSums(wi.1)
d.2<-rowSums(wi.2*di.2)/rowSums(wi.2)
d.3<-rowSums(wi.3*di.3)/rowSums(wi.3)
d.4<-rowSums(wi.4*di.4)/rowSums(wi.4)
d.5<-rowSums(wi.5*di.5)/rowSums(wi.5)

vd.1<-1/rowSums(wi.1)
vd.2<-1/rowSums(wi.2)
vd.3<-1/rowSums(wi.3)
vd.4<-1/rowSums(wi.4)
vd.5<-1/rowSums(wi.5)

# simple contrast: (d.1-d.2)
# 95% CI
ci.su<-d.1-d.2+1.96*sqrt(vd.1+vd.2)
ci.sl<-d.1-d.2-1.96*sqrt(vd.1+vd.2)
wid.s<-abs(ci.su-ci.sl)
ave.wds<-ave(wid.s)
t1.s<-sum(ci.sl>=0 & ci.su>=0)+sum(ci.sl<=0 & ci.su<=0)

# complex contrast: (d.3-(d.4+d.5)/2)
# 95% CI:
ci.cu<-(d.3-(d.4+d.5)/2)+1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
ci.cl<-(d.3-(d.4+d.5)/2)-1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
wid.c<-abs(ci.cu-ci.cl)
ave.wdc<-ave(wid.c)
t1.c<-sum(ci.cl>=0 & ci.cu>=0)+sum(ci.cl<=0 & ci.cu<=0)
```

```
#==================================;
#       The Bonett's model
#==================================;
set.seed(100)
nsim<-2000; nsamc<-15; nsamt<-30; nstudy<-30;
ut<-matrix(0,nsim,nsamt); uc<-matrix(0,nsim,nsamc);
df.t<-matrix(0,nsim,nsamt); df.c<- matrix (0,nsim,nsamc);
sigc<-matrix(0,nsim,nsamc);sigt<-matrix(0,nsim,nsamt);
si<-matrix(0,nsim,nsamc);
di<-matrix(0,nsim,nstudy);
vi<-matrix(0,nsim,nstudy);
bi<-matrix(0,nsim,nstudy);

# data simulation
for (i in 1:2000)
   {
      ## compute effect size by data
      for (j in 1:30)
      {
         nc=15
         nt=30
      ut<-rnorm(nt,0,sqrt(3))
      uc<-rnorm(nc,0,1)
      sigc=var(uc)
      sigt=var(ut)
      si=(sigc+sigt)/2
   df.t<-nt-1
  df.c<-nc-1
   # effect size in study j
        di[i,j]=(mean(ut)-mean(uc))/sqrt(si)

vi[i,j]=di[i,j]*di[i,j]*(sigt*sigt/df.t+sigc*sigc/df.c)/(8*si*si)+(sigt/df.t+sigc/df.c)/si
     bi[i,j] <-1-3/(4*(nc+nt)-9)
        }
    }

  # combine effect sizes from multiple studies k
    ng.1<-15; ng.2<-15;
    ng.3<-10; ng.4<-10; ng.5<-10;
    di.1<-di[,1:15];di.2<-di[,16:30];di.3<-di[,1:10];di.4<-di[,11:20]; di.5<-di[,21:30];
    bi.1<-bi[,1:15];bi.2<-bi[,16:30];bi.3<-bi[,1:10];bi.4<-bi[,11:20]; bi.5<-bi[,21:30];
    vi.1<-vi[,1:15];vi.2<-vi[,16:30];vi.3<-vi[,1:10];vi.4<-vi[,11:20]; vi.5<-vi[,21:30];
```

```
d.1<-rowSums(di.1*bi.1)/ng.1
d.2<-rowSums(di.2*bi.2)/ng.2
d.3<-rowSums(di.3*bi.3)/ng.3
d.4<-rowSums(di.4*bi.4)/ng.4
d.5<-rowSums(di.5*bi.5)/ng.5

vd.1<-rowSums(bi.1*bi.1*vi.1)/(ng.1*ng.1)
vd.2<-rowSums(bi.2*bi.2*vi.2)/(ng.2*ng.2)
vd.3<-rowSums(bi.3*bi.3*vi.3)/(ng.3*ng.3)
vd.4<-rowSums(bi.4*bi.4*vi.4)/(ng.4*ng.4)
vd.5<-rowSums(bi.5*bi.5*vi.5)/(ng.5*ng.5)

# simple contrast: (d.1-d.2)
# 95% CI
ci.su<-d.1-d.2+1.96*sqrt(vd.1+vd.2)
ci.sl<-d.1-d.2-1.96*sqrt(vd.1+vd.2)
wid.s<-abs(ci.su-ci.sl)
ave.wds<-ave(wid.s)
t1.s<-sum(ci.sl>=0 & ci.su>=0)+sum(ci.sl<=0 & ci.su<=0)

# complex contrast: (d.3-(d.4+d.5)/2)
# 95% CI:
ci.cu<-(d.3-(d.4+d.5)/2)+1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
ci.cl<-(d.3-(d.4+d.5)/2)-1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
wid.c<-abs(ci.cu-ci.cl)
ave.wdc<-ave(wid.c)
t1.c<-sum(ci.cl>=0 & ci.cu>=0)+sum(ci.cl<=0 & ci.cu<=0)
```

 # Sample 3;

```
# Power;
# Chi-square distribution df = 8;
# Number of study k = 120;
# Equal within-study sample sizes: nc = nt = 15;
# Equal within-study variance 1:1.

#==================================;
#         The Hedges & Olkin FE model
#==================================;
set.seed(100)
nsim<-2000; nsamc<-15; nsamt<-15; nstudy<-120;
ut<-matrix(0,nsim,nsamt); uc<-matrix(0,nsim,nsamc);
sigc<-matrix(0,nsim,nsamc);sigt<-matrix(0,nsim,nsamt);
si<-matrix(0,nsim,nsamc);
```

```
di.1<-matrix(0,nsim,nstudy);
vi.1<-matrix(0,nsim,nstudy);
di.2<-matrix(0,nsim,nstudy);
vi.2<-matrix(0,nsim,nstudy);

for (i in 1:2000)
   {
       ## compute effect size by data
       ## study group 1, nstudy1=60
       for (j in 1:60)
       {
          nc=15
          nt=15
       ut<-rchisq(nt,8)
       uc<-rchisq(nc,8)
      sigc<-var(uc)
      sigt<-var(ut)
      si<- ((nc-1)*sigc+(nt-1)*sigt)/(nc+nt-2)

   # effect size in study i
      di.1[i,j]<- (mean(ut)-mean(uc))/sqrt(si)
vi.1[i,j]<-(nt+nc)/(nt*nc)+di.1[i,j]*di.1[i,j]/(2*(nt+nc))
wi.1<-1/vi.1
       }

   ## study group 2, nstudy2=60
       for (j in 1:60)
       {
          nc=15
          nt=15
       ut<-rchisq(nt,8)+0.3752
       uc<-rchisq(nc,8)
      sigc<-var(uc)
      sigt<-var(ut)
      si<- ((nc-1)*sigc+(nt-1)*sigt)/(nc+nt-2)

  # effect size in study i
      di.2[i,j]<- (mean(ut)-mean(uc))/sqrt(si)
vi.2[i,j]<-(nt+nc)/(nt*nc)+di.2[i,j]*di.2[i,j]/(2*(nt+nc))
wi.2<-1/vi.2
       }
   }
# combine effect sizes from multiple studies k
```

```
di<- cbind(di.1,di.2)
wi<- cbind(wi.1,wi.2)
di.3<-di[,1:40];di.4<-di[,41:80]; di.5<-di[,81:120];
wi.1<-wi[,1:60];wi.2<-wi[,61:120];
wi.3<-wi[,1:40];wi.4<-wi[,41:80]; wi.5<-wi[,81:120];

d.1<-rowSums(wi.1*di.1)/rowSums(wi.1)
d.2<-rowSums(wi.2*di.2)/rowSums(wi.2)
d.3<-rowSums(wi.3*di.3)/rowSums(wi.3)
d.4<-rowSums(wi.4*di.4)/rowSums(wi.4)
d.5<-rowSums(wi.5*di.5)/rowSums(wi.5)

vd.1<-1/rowSums(wi.1)
vd.2<-1/rowSums(wi.2)
vd.3<-1/rowSums(wi.3)
vd.4<-1/rowSums(wi.4)
vd.5<-1/rowSums(wi.5)

# simple contrast: (d.1-d.2)
# 95% CI
ci.su<-d.1-d.2+1.96*sqrt(vd.1+vd.2)
ci.sl<-d.1-d.2-1.96*sqrt(vd.1+vd.2)
wid.s<-abs(ci.su-ci.sl)
ave.wds<-ave(wid.s)[1]
t1.s<-sum(ci.sl>=0 & ci.su>=0)+sum(ci.sl<=0 & ci.su<=0)
power.s <- t1.s/nsim

# complex contrast: (d.3-(d.4+d.5)/2)
# 95% CI:
ci.cu<-(d.3-(d.4+d.5)/2)+1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
ci.cl<-(d.3-(d.4+d.5)/2)-1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
wid.c<-abs(ci.cu-ci.cl)
ave.wdc<-ave(wid.c)[1]
t1.c<-sum(ci.cl>=0 & ci.cu>=0)+sum(ci.cl<=0 & ci.cu<=0)
power.c <- t1.c/nsim


#==================================;
#      The Bonett's model
#==================================;
set.seed(100)
nsim<-2000; nsamc<-15; nsamt<-15; nstudy<-120;
```

```
ut<-matrix(0,nsim,nsamt); uc<-matrix(0,nsim,nsamc);
df.t<-matrix(0,nsim,nsamt); df.c<- matrix (0,nsim,nsamc);
sigc<-matrix(0,nsim,nsamc);sigt<-matrix(0,nsim,nsamt);
si<-matrix(0,nsim,nsamc);
di.1<-matrix(0,nsim,nstudy);
vi.1<-matrix(0,nsim,nstudy);
bi.1<-matrix(0,nsim,nstudy);
di.2<-matrix(0,nsim,nstudy);
vi.2<-matrix(0,nsim,nstudy);
bi.2<-matrix(0,nsim,nstudy);

# data simulation
for (i in 1:2000)
   {
      ## compute effect size by data
      ## study group 1, nstudy1 = 60
      for (j in 1:60)
      {
         nc=15
         nt=15
      ut<-rchisq(nt,8)
       uc<-rchisq(nc,8)
     sigc=var(uc)
     sigt=var(ut)
     si=(sigc+sigt)/2
df.t<-nt-1
df.c<-nc-1

# effect size in study j
      di.1[i,j]=(mean(ut)-mean(uc))/sqrt(si)

vi.1[i,j]=di.1[i,j]*di.1[i,j]*(sigt*sigt/df.t+sigc*sigc/df.c)/(8*si*si)+(sigt/df.t+sigc/df.c)/si
      bi.1[i,j] <-1-3/(4*(nc+nt)-9)
      }

      ## study group 2, nstudy2 = 60
      for (j in 1:60)
      {
         nc=15
         nt=15
      ut<-rchisq(nt,8)+0.3752
       uc<-rchisq(nc,8)
     sigc=var(uc)
```

```
        sigt=var(ut)
        si=(sigc+sigt)/2
      df.t<-nt-1
     df.c<-nc-1

      # effect size in study j
         di.2[i,j]=(mean(ut)-mean(uc))/sqrt(si)

vi.2[i,j]=di.2[i,j]*di.2[i,j]*(sigt*sigt/df.t+sigc*sigc/df.c)/(8*si*si)+(sigt/df.t+sigc/df.c)/si
      bi.2[i,j] <-1-3/(4*(nc+nt)-9)
          }
        }
      # combine effect sizes from multiple studies k
      ng.1<-60; ng.2<-60;
      ng.3<-40; ng.4<-40; ng.5<-40;

      di<- cbind(di.1,di.2)
      bi<- cbind(bi.1,bi.2)
      vi<- cbind(vi.1,vi.2)
      di.3<-di[,1:40];di.4<-di[,41:80]; di.5<-di[,81:120];
      bi.3<-bi[,1:40];bi.4<-bi[,41:80]; bi.5<-bi[,81:120];
      vi.3<-vi[,1:40];vi.4<-vi[,41:80]; vi.5<-vi[,81:120];

      d.1<-rowSums(di.1*bi.1)/ng.1
      d.2<-rowSums(di.2*bi.2)/ng.2
      d.3<-rowSums(di.3*bi.3)/ng.3
      d.4<-rowSums(di.4*bi.4)/ng.4
      d.5<-rowSums(di.5*bi.5)/ng.5

      vd.1<-rowSums(bi.1*bi.1*vi.1)/(ng.1*ng.1)
      vd.2<-rowSums(bi.2*bi.2*vi.2)/(ng.2*ng.2)
      vd.3<-rowSums(bi.3*bi.3*vi.3)/(ng.3*ng.3)
      vd.4<-rowSums(bi.4*bi.4*vi.4)/(ng.4*ng.4)
      vd.5<-rowSums(bi.5*bi.5*vi.5)/(ng.5*ng.5)

      # simple contrast: (d.1-d.2)
      # 95% CI
      ci.su<-d.1-d.2+1.96*sqrt(vd.1+vd.2)
      ci.sl<-d.1-d.2-1.96*sqrt(vd.1+vd.2)
      wid.s<-abs(ci.su-ci.sl)
      ave.wds<-ave(wid.s)[1][1]
      t1.s<-sum(ci.sl>=0 & ci.su>=0)+sum(ci.sl<=0 & ci.su<=0)
      power.s <- t1.s/nsim
```

```
   # complex contrast: (d.3-(d.4+d.5)/2)
   # 95% CI:
   ci.cu<-(d.3-(d.4+d.5)/2)+1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
   ci.cl<-(d.3-(d.4+d.5)/2)-1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
   wid.c<-abs(ci.cu-ci.cl)
   ave.wdc<-ave(wid.c)[1]
   t1.c<-sum(ci.cl>=0 & ci.cu>=0)+sum(ci.cl<=0 & ci.cu<=0)
  power.c <- t1.c/nsim
```

 # Sample 4;

```
# Power;
# Chi-square distribution df = 8;
# Number of study k = 120;
# Equal within-study sample sizes: nc = nt = 15;
# Unequal within-study variance 1:3.

   #==================================;
   #       The Hedges & Olkin FE model
   #==================================;
   set.seed(100)
   nsim<-2000; nsamc<-15; nsamt<-15; nstudy<-120;
   ut<-matrix(0,nsim,nsamt); uc<-matrix(0,nsim,nsamc);
   sigc<-matrix(0,nsim,nsamc);sigt<-matrix(0,nsim,nsamt);
   si<-matrix(0,nsim,nsamc);
   di.1<-matrix(0,nsim,nstudy);
   vi.1<-matrix(0,nsim,nstudy);
   di.2<-matrix(0,nsim,nstudy);
   vi.2<-matrix(0,nsim,nstudy);

   for (i in 1:2000)
      {
         ## compute effect size by data
         ## study group 1, nstudy1=60
         for (j in 1:60)
        {
           nc=15
           nt=15
         ut1<-rchisq(nt,8)
        ut <-ut1*sqrt(3)
         uc<-rchisq(nc,8)
       sigc<-var(uc)
       sigt<-var(ut)
       si<- ((nc-1)*sigc+(nt-1)*sigt)/(nc+nt-2)
```

```
# effect size in study i
       di.1[i,j]<- (mean(ut)-mean(uc))/sqrt(si)
vi.1[i,j]<-(nt+nc)/(nt*nc)+di.1[i,j]*di.1[i,j]/(2*(nt+nc))
wi.1<-1/vi.1
       }


   ## study group 2, nstudy2=60
        for (j in 1:60)
       {
          nc=15
          nt=15
       ut1<-rchisq(nt,8)
       ut <-ut1*sqrt(3)+0.3752
        uc<-rchisq(nc,8)
      sigc<-var(uc)
      sigt<-var(ut)
      si<- ((nc-1)*sigc+(nt-1)*sigt)/(nc+nt-2)
# effect size in study i
       di.2[i,j]<- (mean(ut)-mean(uc))/sqrt(si)
vi.2[i,j]<-(nt+nc)/(nt*nc)+di.2[i,j]*di.2[i,j]/(2*(nt+nc))
wi.2<-1/vi.2
       }
   }
# combine effect sizes from multiple studies k
di<- cbind(di.1,di.2)
wi<- cbind(wi.1,wi.2)
di.3<-di[,1:40];di.4<-di[,41:80]; di.5<-di[,81:120];
wi.1<-wi[,1:60];wi.2<-wi[,61:120];
wi.3<-wi[,1:40];wi.4<-wi[,41:80]; wi.5<-wi[,81:120];

d.1<-rowSums(wi.1*di.1)/rowSums(wi.1)
d.2<-rowSums(wi.2*di.2)/rowSums(wi.2)
d.3<-rowSums(wi.3*di.3)/rowSums(wi.3)
d.4<-rowSums(wi.4*di.4)/rowSums(wi.4)
d.5<-rowSums(wi.5*di.5)/rowSums(wi.5)

vd.1<-1/rowSums(wi.1)
vd.2<-1/rowSums(wi.2)
vd.3<-1/rowSums(wi.3)
vd.4<-1/rowSums(wi.4)
vd.5<-1/rowSums(wi.5)
```

```
# simple contrast: (d.1-d.2)
# 95% CI
ci.su<-d.1-d.2+1.96*sqrt(vd.1+vd.2)
ci.sl<-d.1-d.2-1.96*sqrt(vd.1+vd.2)
wid.s<-abs(ci.su-ci.sl)
ave.wds<-ave(wid.s)[1][1]
t1.s<-sum(ci.sl>=0 & ci.su>=0)+sum(ci.sl<=0 & ci.su<=0)
power.s <- t1.s/nsim

# complex contrast: (d.3-(d.4+d.5)/2)
# 95% CI:
ci.cu<-(d.3-(d.4+d.5)/2)+1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
ci.cl<-(d.3-(d.4+d.5)/2)-1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
wid.c<-abs(ci.cu-ci.cl)
ave.wdc<-ave(wid.c)[1]
t1.c<-sum(ci.cl>=0 & ci.cu>=0)+sum(ci.cl<=0 & ci.cu<=0)
power.c <- t1.c/nsim

#==================================;
#        The Bonett's model
#==================================;
set.seed(100)
nsim<-2000; nsamc<-15; nsamt<-15; nstudy<-120;
ut<-matrix(0,nsim,nsamt); uc<-matrix(0,nsim,nsamc);
df.t<-matrix(0,nsim,nsamt); df.c<- matrix (0,nsim,nsamc);
sigc<-matrix(0,nsim,nsamc);sigt<-matrix(0,nsim,nsamt);
si<-matrix(0,nsim,nsamc);
di.1<-matrix(0,nsim,nstudy);
vi.1<-matrix(0,nsim,nstudy);
bi.1<-matrix(0,nsim,nstudy);
di.2<-matrix(0,nsim,nstudy);
vi.2<-matrix(0,nsim,nstudy);
bi.2<-matrix(0,nsim,nstudy);

# data simulation
for (i in 1:2000)
   {
      ## compute effect size by data
      ## study group 1, nstudy1 = 60
      for (j in 1:60)
      {
         nc=15
         nt=15
```

```
            ut1<-rchisq(nt,8)
           ut <-ut1*sqrt(3)
            uc<-rchisq(nc,8)
          sigc=var(uc)
          sigt=var(ut)
          si=(sigc+sigt)/2
        df.t<-nt-1
       df.c<-nc-1
        # effect size in study j
            di.1[i,j]=(mean(ut)-mean(uc))/sqrt(si)

vi.1[i,j]=di.1[i,j]*di.1[i,j]*(sigt*sigt/df.t+sigc*sigc/df.c)/(8*si*si)+(sigt/df.t+sigc/df.c)/si
        bi.1[i,j] <-1-3/(4*(nc+nt)-9)
            }

            ## study group 2, nstudy2 = 60
            for (j in 1:60)
            {
               nc=15
               nt=15
           ut1<-rchisq(nt,8)
           ut <-ut1*sqrt(3)+0.3752
            uc<-rchisq(nc,8)
          sigc=var(uc)
          sigt=var(ut)
          si=(sigc+sigt)/2
        df.t<-nt-1
       df.c<-nc-1
        # effect size in study j
            di.2[i,j]=(mean(ut)-mean(uc))/sqrt(si)

vi.2[i,j]=di.2[i,j]*di.2[i,j]*(sigt*sigt/df.t+sigc*sigc/df.c)/(8*si*si)+(sigt/df.t+sigc/df.c)/si
        bi.2[i,j] <-1-3/(4*(nc+nt)-9)
            }
         }
        # combine effect sizes from multiple studies k

        ng.1<-60; ng.2<-60;
        ng.3<-40; ng.4<-40; ng.5<-40;

        di<- cbind(di.1,di.2)
        bi<- cbind(bi.1,bi.2)
        vi<- cbind(vi.1,vi.2)
```

```
di.3<-di[,1:40];di.4<-di[,41:80]; di.5<-di[,81:120];
bi.3<-bi[,1:40];bi.4<-bi[,41:80]; bi.5<-bi[,81:120];
vi.3<-vi[,1:40];vi.4<-vi[,41:80]; vi.5<-vi[,81:120];

d.1<-rowSums(di.1*bi.1)/ng.1
d.2<-rowSums(di.2*bi.2)/ng.2
d.3<-rowSums(di.3*bi.3)/ng.3
d.4<-rowSums(di.4*bi.4)/ng.4
d.5<-rowSums(di.5*bi.5)/ng.5

vd.1<-rowSums(bi.1*bi.1*vi.1)/(ng.1*ng.1)
vd.2<-rowSums(bi.2*bi.2*vi.2)/(ng.2*ng.2)
vd.3<-rowSums(bi.3*bi.3*vi.3)/(ng.3*ng.3)
vd.4<-rowSums(bi.4*bi.4*vi.4)/(ng.4*ng.4)
vd.5<-rowSums(bi.5*bi.5*vi.5)/(ng.5*ng.5)

# simple contrast: (d.1-d.2)
# 95% CI
ci.su<-d.1-d.2+1.96*sqrt(vd.1+vd.2)
ci.sl<-d.1-d.2-1.96*sqrt(vd.1+vd.2)
wid.s<-abs(ci.su-ci.sl)
ave.wds<-ave(wid.s)[1]
t1.s<-sum(ci.sl>=0 & ci.su>=0)+sum(ci.sl<=0 & ci.su<=0)
power.s <- t1.s/nsim

# complex contrast: (d.3-(d.4+d.5)/2)
# 95% CI:
ci.cu<-(d.3-(d.4+d.5)/2)+1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
ci.cl<-(d.3-(d.4+d.5)/2)-1.96*sqrt(vd.3+0.25*vd.4+0.25*vd.5)
wid.c<-abs(ci.cu-ci.cl)
ave.wdc<-ave(wid.c)[1]
t1.c<-sum(ci.cl>=0 & ci.cu>=0)+sum(ci.cl<=0 & ci.cu<=0)
power.c <- t1.c/nsim
```