5-2018

Hierarchical Meta-Analysis: A Simulation Study Comparing Classical Random Effects and Fully Bayesian Methods

Nancy R. Andiloro
The Graduate Center, City University of New York

How does access to this work benefit you? Let us know!
Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

Part of the Educational Assessment, Evaluation, and Research Commons, Educational Psychology Commons, and the Social Statistics Commons

Recommended Citation
https://academicworks.cuny.edu/gc_etds/2579

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
HIERARCHICAL META-ANALYSIS: A SIMULATION STUDY COMPARING
CLASSICAL RANDOM EFFECTS AND FULLY BAYESIAN METHODS

by

Nancy R. Andiloro

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment
of the requirements for the degree of Doctor of Philosophy, The City University of New York

2018
HIERARCHICAL META-ANALYSIS: A SIMULATION STUDY COMPARING
EMPIRICAL BAYES AND FULLY BAYESIAN METHODS

by

Nancy R. Andiloro

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

________________    __________________________________________
Date      David Rindskopf
Chair of Examining Committee

________________    __________________________________________
Date      Bruce Homer
Executive Officer

Supervisory Committee

David Rindskopf
Hannah Rothstein
Philip Ramsey

THE CITY UNIVERSITY OF NEW YORK
ABSTRACT

HIERARCHICAL META-ANALYSIS: A SIMULATION STUDY COMPARING CLASSICAL RANDOM EFFECTS AND FULLY BAYESIAN METHODS

by

Nancy R. Andiloro

Advisor: David Rindskopf

Meta-analytic data have a natural hierarchical structure to them, where individuals are nested within studies, and have both within-and between-study variation to model. A random-effects hierarchical linear model is useful to conduct a meta-analysis because it allows one to appropriately parse out the two components of variation that exist within and across studies to determine an observed effect. Empirical Bayes estimation considers the reliability of variance estimates; when the reliability of the effect size estimate for a study is high, substantial weight is placed on that estimate. However, problems with estimation arise when the number of studies and their sample size is small. Although time-consuming to employ, fully Bayesian methods offer a solution, but few studies systematically compare random-effects to fully Bayesian methods. A simulation study was performed varying certain characteristics of meta-analyses, such as the number of studies, their sample size and level of heterogeneity across studies, to determine under which condition(s) a fully Bayesian method improves meta-analytic findings. Results are unexpectedly inconsistent, whereby certain scenarios in which the number of studies is small and level of heterogeneity large, show that empirical Bayes performs better than the fully Bayesian method. Despite this, bias and mean-squared error are lower, on average, among the fully Bayesian models, with a model specifying a Cauchy prior on $\tau$ performing best for the most favorable scenario. Implications and areas for future study are discussed.
# TABLE OF CONTENTS

Chapter I: Introduction........................................................................................................ 1

Chapter II: Literature Review .......................................................................................... 5
  Meta-analysis guidelines ............................................................................................... 5
    Choosing the best model ............................................................................................ 7
    Sources of variation in random-effects methods ...................................................... 13
    Methods for estimating $\tau^2$, the between-study variance ................................... 13
    Using weights in random-effects estimation ............................................................ 15
    Limitations of random-effects meta-analysis .......................................................... 16
  Fully Bayesian Methods ............................................................................................. 18
    Bayesian credible intervals ....................................................................................... 24
    Prior distributions in fully Bayesian random-effects meta-analysis ....................... 24
    Random Effects versus Fully Bayesian Approaches to Meta-analysis .................. 28

Chapter III: Simulation Design and Method................................................................... 34
  Aims and Objectives ................................................................................................. 34
  Simulation Procedures ............................................................................................. 34
  Software to perform simulations .............................................................................. 36
  Statistical Methods Evaluated .................................................................................. 36

Chapter IV: Results ....................................................................................................... 39
  Results for the estimate of $\tau$ ............................................................................... 40
  Results for the estimate of $\mu$ ................................................................................ 43

Chapter V: Discussion ................................................................................................... 45
  Tables and Figures .................................................................................................... 45

Appendix I: Syntax for simulation runs ......................................................................... 50

Appendix II: Raw results of 84 simulation runs – Fixed-effect and Random-effects .... 73

Appendix III: Raw results of 84 simulation runs – FB Pareto and FB Uniform ............. 77

Appendix IV: Raw results of 84 simulation runs – FB Half-Cauchy ......................... 81

References .................................................................................................................. 85
TABLES AND FIGURES

Figure 2. Variance of $\tau$ by number of studies across each simulation run – random effects and FB models. ........................................................................................................................................................................53

Figure 3. Mean of $\mu$ estimate by number of studies across each simulation run – fixed effect, random effects, and FB models........................................................................................................................................................................55

Figure 4. Variance of $\mu$ estimate by number of studies across each simulation run – fixed effect, random effects, and FB models........................................................................................................................................................................57

Figure 5. Lower (2.5%), median (50%), and upper (97.5%) quantiles of $\tau$ estimates by number of studies across each simulation run – random effects and FB models........................................................................................................................................................................59

Figure 6. Lower (2.5%), median (50%), and upper (97.5%) quantiles of $\mu$ estimates by number of studies across each simulation run – fixed effect, random effects, and FB models........................................................................................................................................................................63
Chapter I: Introduction

As part of a systematic review, meta-analysis “is the use of statistical techniques to integrate and summarize the results of included studies.” (Liberati et al., 2009, p. 2) Meta-analysis is also known as a study of studies (Glass, 1976) and is a technique that quantitatively summarizes results from a series of related studies to determine if there is consistency across their findings. A primary goal in meta-analysis is to examine whether a treatment in question has a real effect, and it typically consists of translating the results of multiple studies into one effect size estimate. Conducting a meta-analysis allows the synthetist to measure the magnitude of an effect with more power than any one study could accomplish, because it combines the effects of multiple studies.

Meta-analytic data have a natural hierarchical structure to them, where individuals are nested within studies, having both within and between-study variation. A hierarchical linear model is useful when conducting a meta-analysis because it allows one to appropriately parse out components of variation—due to, for example, differences in study design, methodology used, or measurement of the dependent variable—across studies to determine a real effect. An average effect size, variance of effect size estimates, and the residual variance of estimates may be considered in a hierarchical model.

Consider a meta-analysis model that aims to assess the effect of coaching (the treatment) versus no coaching (the control) on SAT scores. The effect size in study $i$ can be estimated as:

$$d_i = \frac{\bar{Y}_{Ti} - \bar{Y}_{Ci}}{s_i}$$

Where:

$\bar{Y}_{Ti} =$ the mean SAT score of the treatment group in study $i$

$\bar{Y}_{Ci} =$ the mean SAT score of the control group in study $i$
In a hierarchical meta-analysis, the model can be written as two equations, one at each level of aggregation. The level 1 equation—known as “the level of measurements” (Van Den Noortgate & Onghena, 2003, p. 2)—is:

\[ d_i = \delta_i + \epsilon_i, \quad i = 1, \ldots, k \]

Where:

- \( d_i \) = estimated standardized mean difference in study \( i \) between the coaching and no coaching groups
- \( \delta_i \) = true effect size
- \( \epsilon_i \) = sampling error, \( \epsilon_i \sim N(0, \sigma_i^2) \)

In the level 2 equation—known as the “the level of cases” (Van Den Noortgate & Onghena, 2003, p. 2)—the effect size estimates \( \delta_i \) can vary as a function of known characteristics of a study and random error:

\[ \delta_i = W_i \gamma + U_i \]

Where:

- \( W_i \) = a (q X 1) vector of constants representing known characteristics of studies
- \( \gamma \) = a (q X 1) vector of between-study parameters
- \( U_i \) = random error, \( U_i \sim N(0, \tau^2) \)

In a random effects (RE) analysis, studies with greater sample sizes are more reliably estimated and are therefore given greater weight. RE models do not assume that effect sizes are the same across studies—an assumption common with more traditional meta-analytic methods—but sampled from a distribution of true effect sizes. An empirical Bayes (EB) analysis involves estimating a prior distribution from the data (Raudenbush & Bryk, 1985). This contrasts with a
fully Bayesian (FB) analysis, in which a prior distribution is chosen before any data are observed.

There are disadvantages to an RE meta-analysis. One is that the estimate of the level 2 variance, $\tau^2$, is first estimated from the observed data and then treated as if it were known. This is a problem because the uncertainty of the estimate of $\tau^2$ is not considered and is coupled with the fact that the weighted estimates of the mean effect size and regression coefficients in a meta-analysis are dependent on this uncertainty (Raudenbush, 2009). Second, the number of studies and the sample size within each study, when small, may underestimate standard errors and compromise the validity of the estimates of the fixed effect, $\gamma$. Consequently, this exacerbates the problem made by the uncertainty in the estimate of $\tau^2$ just discussed.

A FB approach to random-effects analysis is characterized by its use of probability theory to express uncertainty about unknown parameters, as well as its inclusion of prior information that might be available about the parameters estimated in a model. In a FB hierarchical meta-analysis, parameters characterizing a study are treated not just as fixed unknown constants, but as random variables, having their own distribution, thereby more effectively modeling the uncertainty of combining potentially disparate studies for a pooled analysis. FB meta-analyses allow: (a) all parameter uncertainty to automatically be accounted for in the analysis; (b) evidence from a variety of sources, regarding a specific problem, to be considered within a single model; and (c) probability statements to be made directly regarding quantities of interest (Sutton & Abrams, 2001).

As with RE meta-analyses, there are also disadvantages to doing a FB meta-analysis, which include: (a) there are few guidelines available regarding the use of prior information; (b) the use of different priors leads to varying results, so a sensitivity analysis should always be
performed; and (c) FB meta-analyses are computationally complex and can be a daunting and time-consuming task to a researcher unfamiliar with this technique (Sutton & Abrams, 2001).

Given the computational challenges involved with a FB meta-analysis, the current study assesses the extent to which a FB method improves estimation over and above an EB method when performing a random-effects meta-analysis. A few studies (e.g., Smith, Spiegelhalter, & Thomas, 1995; Abrams & Sanso, 1998; Higgins, Thompson, & Spiegelhalter, 2009) have compared the results of meta-analyses for EB versus FB methods, with mixed findings. To date, there are no studies that have systematically varied characteristics important when considering meta-analysis, such as the number of studies combined and the level of heterogeneity across them, as well as the sample size within each study. In a simulation study, where the truth is known, a researcher can purposefully vary study characteristics to determine at what point, and for which combinations, does a method of analysis stray from the truth, so much so that a more computationally rigorous method of analysis, such as the FB method, should be employed. The current study aims to determine the conditions under which this occurs and potentially provide important guidance to not only researchers in the field of educational psychology, but to researchers in all the social sciences, who are becoming keener on employing meta-analyses for combining information on well-studied phenomena.
Chapter II: Literature Review

Meta-analysis guidelines

Meta-analyses and systematic reviews have become increasingly important, particularly in the medical field (Draper et al., 1993). A common phrase restated in the research literature is that findings predicting an outcome of interest are revealed to have “mixed results.” A systematic review of all studies testing this outcome is then necessary to determine the plausibility of a real effect. This effectively increases the need for consensus on meta-analytic reporting standards that emphasize high study quality and greater transparency.

Researchers in the social sciences have responded to this need by publishing a series of reporting standards to use as guidelines when conducting meta-analyses. The American Psychological Association (APA) Publications and Communications Board Working Group on Journal Article Reporting Standards (Applebaum et al., 2008) recommend including new data collection methods, evaluations of interventions having both random and nonrandom assignment, and meta-analyses (known as Meta-analysis Reporting Standards (MARS)) in every research article seeking publication (also see Cooper, 2010). When submitting a manuscript for publication, MARS provides a checklist for reporting inclusion and exclusion criteria, moderator and mediator analyses, search strategies, coding procedures, and statistical methods.

Moher, Liberati, Tetzlaff, and Altman (2009) focus specifically on meta-analysis research and aimed to address the “suboptimal reporting of meta-analyses” (p. 1). These PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines recommend reporting the identification, screening, eligibility requirements, and studies ultimately included in a meta-analysis because it will determine the type of statistical combination of effects (e.g., fixed-effects versus random-effects methods) that is appropriate. PRISMA also focuses on the
importance of assessing two key issues important for conducting a successful meta-analysis: (a) heterogeneity across studies and (b) risk of bias. Heterogeneity in meta-analysis deals with the degree of variation across studies that may be due to random error, but also may be due to systematic differences across studies not attributed to chance, including participant characteristics, variation in settings and interventions, study design factors, and variation in the method used to measure the outcome (Song, Sheldon, Sutton, Abrams, & Jones, 2001). Bias in meta-analyses can come in many forms, but meta-analysts should be concerned primarily with selection bias (Draper et al., 1993) which is bias caused by the inclusion of an unrepresentative sample of studies. There are three types of selection bias: (a) publication bias (Rosenthal, 1979), where studies with only significant results are published in journal articles, (b) reporting bias (Hedges, 1992), occurring among published studies, where particular analyses resulting in non-significant findings are not reported or, if they are reported, sufficient information required for the systematic review is not available and (c) retrieval bias (Rosenthal & Rubin, 1988), when one is unable to retrieve studies, both published and unpublished, that answer a particular research question. Publication bias is a difficult problem to address in a meta-analysis and is further exacerbated when considering other potential sources of bias including the variation in study quality (Song, Eastwood, Gilbody, Duley, & Sutton, 2000; Sutton, Song, Gilbody, & Abrams, 2000).

Methods to assess the extent of heterogeneity and bias in meta-analyses do exist, and include both graphical and statistical approaches (Cochran, 1954; Copas, 1999; Draper et al., 1993; Higgins & Piggott & Shepperd, 2013; Rosenthal, 1979; Song et al., 2001; Sutton et al., 2000; Thompson, 2002; Thompson, 1994; Thompson et al., 2010) and, while these methods are
important to consider before choosing the right meta-analytic approach, they are not the focus of this study.

Performing a meta-analysis involves making many decisions about the types of studies to include and the method in which to combine them. Choices for combining studies, particularly those with small samples, can seriously impact the findings and the conclusions that are made. Turner, Bird, and Higgins (2013) examined underpowered studies in Cochrane reviews and its impact on meta-analyses. The authors found that when at least two adequately-powered studies are included in the review, the underpowered studies do not contribute very much to the analyses. The underpowered studies, however, make up the bulk of the studies that were included in Cochrane reviews (70%) and in certain situations—for example, when the ensemble of studies is similar in size and number—can lead to great losses in the precision of the interpretation of the effect.

The results from a meta-analysis are said to be robust if they remain consistent across deviations from specific assumptions (Draper et al., 1993). This includes whether literature searches should involve the efforts of multiple researchers independently; findings are affected by modeling changes; the combined estimate being sensitive to the inclusion of one study; the studies being combined are a representative sample of studies addressing the research questions; and the extent to which publication bias affects the results of the systematic review.

**Choosing the best model**

*Fixed Effect (FE) models*

Despite their disadvantages, described below, many studies in education, psychology, and health have primarily utilized FE models when performing a systematic review (e.g., see Schmidt et al., 2009; Schmidt, 2008; Stangl & Berry, 2000). For the most part, these methods are
utilized without justification and oversimplified (Lau, Ioannidis, & Schmid, 1998; Shadish, Cook, & Campbell, 2002), especially given the plausibility that heterogeneity between studies does exist. However, FE models prevail because of their simplicity and ease of interpretation of findings (National Research Council, 1992).

The underlying assumption of the simple FE model considered here is that the true effect size, \( \delta \), is fixed and homogenous across all studies and that the only variability that needs to be considered is the sampling variability. This is the case in situations in which study characteristics are not considered. Therefore, the standard deviation between studies, \( r \), is equal to zero. The FE treatment effect, \( \hat{\mu} \), is estimated as a weighted average (Shadish & Haddock, 2009):

\[
\hat{\mu} = \frac{\sum_{i=1}^{k} w_i Y_i}{\sum_{i=1}^{k} w_i} = \frac{\sum_i w_i Y_i}{\sum_i w_i}
\]

\( Y_i \) is the observed effect size with an assigned weight, \( w_i \), calculated for each study as the inverse of the within-study error variance, \( w_i = \frac{1}{\hat{\sigma}_i^2} \). This is also inversely proportional to each study’s sample size.

The variance of the weighted mean effect size is the inverse of the sum of weights (Borenstein, Hedges, Higgins, & Rothstein, 2010):

\[
\nu_\mu = \frac{1}{\sum_{i=1}^{k} w_i}
\]

The square root of \( \nu_\mu \) is the standard error of the mean effect size. A 95% confidence interval is calculated as:

\[
Lower \ Limit = \hat{\mu} - 1.96\sqrt{\nu_\mu}
\]

\[
Upper \ Limit = \hat{\mu} + 1.96\sqrt{\nu_\mu}
\]

Note that the above equations do not include a term for the between-study error variance. The FE model assumes that this is equal to zero.
Advantages to the Fixed Effect model

The primary disadvantage of the FE model is its inability to model the variation between studies. It has been said that there will always be some variation between studies (National Research Council, 1992), especially given the inevitable differences in the study design, sample, and protocols of intervention across studies (Higgins & Thompson, 2002). The real question is not whether there is between-study variation but rather, what is the size of this variation. If large, the estimates calculated using the FE method will be seriously biased. This has been demonstrated by Higgins, Thompson, and Spiegelhalter (2009) among studies in medicine and the social sciences showing that studies included in meta-analyses vary a significant amount by the population, setting, treatment, and outcome, making application of the FE model ill-advised.

To further support this argument, Schmidt, Oh, and Hayes (2009) reanalyzed 68 meta-analytic studies using the FE model and published between 1988 and 2006. After employing the RE model, the authors found that the FE studies reported confidence intervals significantly narrower than those found with the RE model (on average, FE width was 56% of the width found with RE). When the variation found between studies is not accounted for, it results in an underestimation of the standard error of the mean effect size, leading to confidence intervals that are not wide enough, and an invalid estimate of the effect (National Research Council, 1992; Overton, 1998; DuMouchel, 1994; Kisamore & Brannick, 2008; Schmidt et al., 2009).

Random-Effects/Mixed-Effects meta-analysis

A random-effects (RE) meta-analytic model is a two-level multi-level model where observations are nested within studies. Implementing a RE model assumes that there are two error structures; the random error within studies and random error between studies (Viechtbauer, 2005). Therefore, the defining feature of the random-effects model is the assumption that the
estimation of the true effect size is a distribution of effect sizes with the combined studies representing this population of study effects. Utilization of the RE model results in a more thorough and complete explanation of variation and a more valid estimate of the mean effect size.

To reiterate, using slightly different notation, the RE model is expressed as (Raudenbush & Bryk, 2002):

\[ Y_i = \theta_i + \epsilon_i, \]

where \( Y_i \) is the observed effect size estimate for study \( i \), \( \theta_i \) represents the true effect size for each of the \( i=1,...,k \) studies, and \( \epsilon_i \) is the within-study sampling error, assumed IID (independently and identically distributed) and normally distributed with a mean of zero and known variance, \( \sigma_i^2 \).

The level-2 equation can include study-level covariates that may account for some of the study-level heterogeneity, and then the level-2 error term, \( \tau^2 \), is the remaining true variation in the pooled effect-size estimate that the covariates do not explain. Raudenbush (2009) specifies this as the following linear model:

\[ \theta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \delta_i. \]

Here \( \beta_0 \) is the estimate of the intercept across studies, the \( \beta_x \) represent the regression coefficients, \( X_{ip} \) are the study-level variables that may explain some of the variation, and \( \delta_i \) the random error of the \( i^{th} \) study, assumed iid and normally distributed with mean zero and variance, \( \tau^2 \). The equations for levels 1 and 2 can be combined into what is known as the hierarchical linear model:

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \delta_i + \epsilon_i. \]
Here, it is assumed that $\delta_i + \varepsilon_i \sim N(0, \nu_i^2)$ and $\nu_i^2 = \tau^2 + \sigma_i^2$, is the total variance of the observed effect sizes.

The underlying assumption of the RE model is that the estimate of the effect size is drawn from a distribution of the effect sizes and the true effect for each study is sampled from a distribution of true effects (Borenstein et al., 2010; Shadish et al., 2002).

Building from the FE model described above, the weights used to calculate the estimate of the mean effect uses both the within-and between-study variance (Borenstein et al., 2010):

$$w_i^* = \frac{1}{\sigma_i^2 + \tau^2}$$

The mean effect size, $\mu$, is calculated as:

$$\mu = \frac{\sum_{i=1}^{k} w_i^* Y_i}{\sum_{i=1}^{k} w_i^*} = \frac{1}{\sum_{i=1}^{k} w_i^*} \sum_{i=1}^{k} w_i^* Y_i,$$

and the variance of the mean, $V_{\mu^*}$ is estimated as:

$$V_{\mu^*} = \frac{1}{\sum_{i=1}^{k} w_i^*}$$

With the standard error taken from the square root of the variance of $\mu$, $V_{\mu^*}$, and the 95% confidence interval calculated the same as above, $\mu \pm 1.96(SE_{\mu^*})$ (Borenstein et al., 2010).

**Random effects versus fixed effect methods.**

While it has been said that the two main general methods for performing a meta-analysis are fixed-effects and random-effects, a random-effects model is usually the most suitable choice. Per Borenstein et al. (2010), this is due to several reasons. First, the random-effects model, which adds a between-studies variance parameter, $\tau^2$, is more likely to fit the sampling distribution of a meta-analysis, as the result of combining studies with different types of participants, study designs, and methodologies for measuring the treatment of interest is likely to produce effect sizes that differ across studies. Second, in statistical modeling, it is typically desirable to avoid
imposing unnecessary restrictions to the model. For a fixed-effect model, imposing a restriction of a common effect-size is just that. Third, if the studies do have a common effect, a random-effects model has the convenient property of reducing to the results of a fixed-effect model. Finally, if the combined studies use an “identical, narrowly defined population,” (p. 107), it would mean that the results would not be generalizable to a larger population, and so extrapolation to this population would not be advisable.

The focus of this paper is to test the benefits of a random-effects meta-analysis, but there are several mistakes to avoid when attempting to choose the best model (Borenstein et al., 2010). First, the test for heterogeneity should not be relied upon, due to its low power. This is particularly the case with the \( Q \) statistic (Cochran, 1954), which is a test for homogeneity having a \( \chi^2 \) distribution with \( k - 1 \) degrees of freedom. This test, as with all such tests, is problematic because it relies on \( k \), the number of studies, and therefore may have low power to detect between-study variation if a small number of studies are included in the systematic review.

The common approach to meta-analysis has been to start with a fixed-effect model, then test for heterogeneity and, if the heterogeneity test is significant, run a random-effects model. This strategy uses poor logic because it is the fixed-effect model that makes more stringent assumptions about a common effect size, so it can be considered a special case of the random-effects model. Borenstein et al. (2010) explain, “Rather than start with either model as the default, one should select the model based on their understanding about whether or not the studies share a common effect size and on their goals in performing the analysis.” (p.108) A random-effects model may be the model with which is most appropriate to begin, and it comes with its own set of assumptions and limitations to consider.
Sources of variation in random-effects methods.

In a random-effects model, there are two sources of variance: 1) between-study variation, referring to the functional differences between the studies included in a meta-analysis, and 2) within-study variation, referring to the idea that the observed effect size for a study is not the same as its true effect size (Borenstein et al., 2010). If the two sources of variance function independently, and we assume every study has the same population variance and sample size, then the variance of the combined effect is defined as:

\[ V_m = \frac{\sigma^2}{k \times n} + \frac{\tau^2}{k} \]

The first term is the same as that of the fixed-effect model and, with a large sample size, this term will approach zero (Borenstein et al., 2010). However, the second term, referring to the between-study variation, approaches zero only as \( k \) approaches infinity. Therefore, the extent to which between-study variation can be precisely estimated greatly depends on the number of studies combined.

Methods for estimating \( \tau^2 \), the between-study variance.

Several methods have been developed to estimate the between-study variance, \( \tau^2 \), in a RE model. One common method was introduced by DerSimonian and Laird (D-L) (1986). The D-L method is often used for point estimation in RE meta-analyses because it is a simple, non-iterative, moment-based approach to accounting for between-study variation in effect size estimation. For the D-L method, the equation used to calculate the point estimate of \( \tau^2 \) is:

\[ \tau^2 = \frac{Q - df}{C}, \]

where \( Q \) is calculated as:

\[ Q = \sum_{i=1}^{k} w_i Y_i^2 - \frac{(\sum_{i=1}^{k} w_i Y_i)^2}{\sum_{i=1}^{k} w_i}, \]
and $C$ is calculated as

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

(see Borenstein et al., 2010, p.106). While the D-L method—whereby a distributional assumption for the probability density of $\tau^2$ is not made—provides a consistent estimate of $\tau^2$, it lacks efficiency. Consequently, because of the uncertainty in the $\tau^2$ estimate, the result is an underestimation of the standard error of the overall effect size estimate.

Unlike the D-L method, the Restricted Maximum Likelihood (REML) method assumes a normal distribution for $\delta_i$ and is a maximum likelihood approach based on the likelihood of $\tau^2$, given the data (Raudenbush, 2009). REML balances unbiasedness and efficiency in the estimate of $\tau^2$, because it considers the uncertainty in the regression coefficients and improvement in the standard errors of the overall effect size estimate. Furthermore, because REML makes a distributional assumption for the $\delta_i$, it allows for the derivation of prediction intervals so that inferences can be made for new studies. REML also effectively addresses the estimation of study-specific effect sizes. Recent simulation studies have shown that REML is one of the preferred methods for estimating the between-study variance (Veroniki et al., 2015).

EB estimation is a special type of likelihood method for estimating $\tau^2$ that has not been utilized as often in meta-analysis. EB provides improved estimation of study-specific effects by combining the fixed effect with the estimate for the random effects (Raudenbush & Bryk, 1985). Raudenbush (2009) notes that EB estimates may be unconditional or conditional, depending on whether level-2 (i.e., study-level) covariates are included in the model. Based on the unconditional model, the average of $\theta_i$ is estimated using the equation: $\theta_i = \mu + \delta_i$ where the expectation of $\theta_i$ is expressed as

$$E(\theta_i|Y_i, \mu, \tau^2) = \theta_i^*(\tau) = \hat{\lambda}_i Y_i + (1 - \hat{\lambda}_i) \mu$$
and, when Level-2 covariates are included, with the conditional model being $\theta_i = x_i \beta + \delta_i$ is expressed as

$$E(\theta_i | Y_i, X, \mu, \tau^2) = \theta_i^*(\tau) = \hat{\lambda}_i Y_i + (1 - \hat{\lambda}_i) (X \beta).$$

EB estimates are also called shrinkage estimates because they shrink estimates toward the overall mean, especially for small values of $\tau$ (Morris, 1983; DuMouchel & Normand, 2000).

In the equations above $\hat{\lambda}_i$ represents the reliability of study $i$’s effect size and is computed as:

$$\hat{\lambda}_i = \tau^2 / (\tau^2 + \sigma^2).$$

EB estimates borrow strength in the form of reliability estimates from each of the studies. Studies having low reliability borrow the greatest strength from more reliable studies and, consequently, experience the greatest shrinkage towards the mean. The level of heterogeneity among the studies in a meta-analysis also affects EB estimation. Raudenbush (2009) explains that if the studies included were homogeneous—i.e., $\tau^2$ is close to 0—then the best estimate for $\theta_i$ is $\mu$ and the RE model would reduce to a FE model. Alternatively, if the studies are heterogeneous—i.e., when $\tau^2$ is large—they are not likely to borrow strength from the other studies and the estimate for $\theta_i$ will be closer to each study’s individual value.

**Using weights in random-effects estimation.**

Relative to the fixed-effect method, random-effects meta-analyses use different weights to calculate the mean effect size (Borenstein et al., 2010). In the fixed-effect model, the weights depend on the within-study variance alone, as the assumption is that there is no variation between studies, $W_i = 1/V_i$, where $V = \sigma_i^2 / n_i$. Therefore, it tends to ignore the information obtained from smaller studies, because it assumes the best information about the one effect size being estimated would come from a larger study. A random-effects method, however, is not
estimating one true effect, but a mean from a distribution of effects, so between-study variation must now be considered, $W_i = 1/(V_i + \tau^2)$. All studies included in the meta-analysis should be represented where smaller studies, although imprecise, provide information of a different effect size so they should not be ignored. The same idea applies to larger studies; they should not be given too much weight as they are only providing just one effect size, for each study, of a distribution of effect sizes. The study weights in a random-effects meta-analysis will be more similar than in a fixed-effect meta-analysis. However, this will depend on the ratio of within-to-between study variation; if the within-study variation is very large and the between-study variation very small, the weights will be driven largely by the sample size for each study. The reverse is true if the within-study variation is small and the between-study variation is large; the weights will be very similar across studies (Borenstein et al., 2010).

**Limitations of random-effects meta-analysis.**

Performing a random-effects meta-analysis comes with limitations. The first is encompassed in a statement from Borenstein et al. (2010): “While the random-effects model is often the appropriate model, there are cases where it cannot be implemented properly because there are too few studies to obtain an accurate estimate of the between-studies variance.” (p.109) If the number of studies included in a meta-analysis is small, the between-study variance estimate will be imprecise so, while it may be the right model, information is lacking to be carried out correctly (also see Bowater & Escarela, 2013; DuMouchel, 1994; Moreno et al., 2012; Raudenbush, 2009; and Sutton et al., 2000). Second, a departure from the assumption that the error variance is randomly distributed can lead to results that are misleading. If there is a correlation between effect-size estimates and the error variance, which happens if the effects vary by study size, can affect the validity of the findings. This typically occurs in the presence of
publication bias, making random-effects models more vulnerable to this type of bias. Additionally, RE models are vulnerable to publication bias because it gives larger weights to smaller studies, relative to the FE model (Sutton et al., 2000). It is for this reason that some meta-analysts have reported being wary of the idea of estimating a combined effect size that places too much emphasis on smaller studies and less emphasis on more precisely estimated larger studies. Consequently, researchers will place more trust in one large primary study when a meta-analysis includes studies with small sample sizes (Bowater & Escarela, 2013).

Some approaches exist to dealing with the inclusion of a small number of studies in meta-analyses, although there are caveats to each (Borenstein et al., 2009). One option is to estimate effects separately, refraining from estimating a pooled effect size, with the expectation that, once it is seen that the effect sizes are so different, it will be understood that combining disparate studies is inappropriate. The problem is that some researchers will attempt to combine studies anyway, resulting in conclusions that are misleading. A second option is to perform a simpler fixed-effect analysis, involving a descriptive analysis that would not allow for extrapolation of the findings to the larger population. The problem here is that because, generally, extrapolation is the very thing researchers wish to do, they may make these inferences anyway, however unwarranted.

The best option, according to Borenstein et al. (2010), is to perform a Bayesian meta-analysis, in which a model is built upon the notion of prior information—information that is available outside of the observed data gathered for the meta-analysis—that may be included to make estimates more precise. One issue preventing this method from being adopted is researchers’ unfamiliarity with the approach. Alternatively, some researchers are opposed to the underlying Bayesian philosophy.
Current research does not demonstrate—through simulation studies, specifically—the extent to which including a small number of studies for a meta-analysis affects statistical inference after implementing a random-effects model and, more importantly, whether proposed adjustments to the random-effects model improves its performance. The discussion now turns to the FB random-effects meta-analysis model which, despite the absence of a systematic comparison across studies, may have the potential to address many of the limitations just described.

**Fully Bayesian Methods**

A fully Bayesian (FB) modelling approach is becoming more popular because it is intuitive and flexible, especially due to recent advances in computational methods. FB methods offer a more advantageous approach compared to the traditional, frequentist approach, especially given the limitations to the RE methods. Namely, the misestimation of $\tau^2$ when the number of studies included is small and, once estimated, treating it as if it is known when it is not. (Sutton, 2000; Sutton & Abrams, 2001).

A unique characteristic of Bayesian methods is the use of probability theory to express uncertainty about parameters in statistical models (Gelman, Carlin, Stern, & Rubin, 2004; Lynch, 2007). Most statistical methods view parameters as fixed unknown quantities, with probability theory applying only to data, given the parameters. For a Bayesian, inference is about refining beliefs about parameters using data. A Bayesian could, for example, correctly say that there is a 95% chance the parameter is in a certain interval as a summary of his or her beliefs about the parameter, or that the probability that a parameter is positive is (e.g.) 73%. Another unique characteristic of Bayesian models is the inclusion of any prior information that might be available about parameters. In the social sciences, for example, it would often be reasonable to
have a prior distribution on a correlation such that the correlation is likely to be positive (as are
most of the correlations between positively viewed traits) and likely not be too close to 1. In
other cases, little may be known about a situation, and then a prior is used that expresses this
uncertainty by specifying what is called an uninformative, or flat, prior.

Gelman et al. (2004) describe Bayesian data analysis as involving three basic steps. The
first step is to create a full probability model, which sets up a joint probability distribution for all
parameters in the model, both observable (e.g., data) and unobservable (e.g., including prior
information). This step requires mindful inclusion of knowledge of the scientific problem, as
well as design and methodology. The second step involves calculating a posterior distribution,
which is a conditional probability distribution \( p(\theta|y) \) of the parameters of interest, denoted \( \theta \),
given the observed data \( y \). The third step involves evaluating model fit, establishing whether the
conclusions make sense and whether they are consistent with the assumptions established in the
first step. The posterior distribution is the product of the likelihood of the observed
data, \( y \), times a distribution of \( \theta \) that summarizes beliefs about \( \theta \) before observing the data, the
prior distribution. The posterior distribution is computed using the following equation, known as
Bayes Rule:

\[
p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}
\]

\( p(y|\theta) \), a function of \( y \) for a fixed \( \theta \), is the likelihood function. \( p(\theta) \) is the prior information
about \( \theta \). The less prior information we know about the scientific question, the vaguer (or flatter)
is the prior applied to the parameters, which then provides only minimal information. In this
case, more emphasis is placed on the likelihood, which summarizes information in the observed
data. The researcher dictates the degree to which a prior is informative. All parameters included
in the model have their own prior probability distributions. In ML models, RE methods first get
point estimates of variances (random effects), and then find estimates of fixed effects, with the variances assumed fixed. If the variances are estimated from a large enough sample of studies this generally is not problematic but, with small samples, uncertainty in the variance estimates can make standard errors for fixed effects smaller than they should be. FB methods provide a better approach than RE in cases with small sample sizes because the standard errors are realistically larger.

Higgins et al. (2009) re-evaluate the role of meta-analysis in the presence of heterogeneity and note some limitations of commonly-used methods. The authors test how well five main objectives in random-effects meta-analysis (i.e., quantification of heterogeneity of findings; estimation of the underlying mean $\mu$; estimation of study-specific effects, $\theta_i$; prediction of effect in a new study, $\theta_{new}$; and testing of whether an effect exists in any study) can be met by: (a) assuming nothing about the distribution of random effects, (b) assuming a normal distribution for the random effects, and (c) allowing a more flexible distribution for the random effects. This is demonstrated using an example of studies on set-shifting ability in people with eating disorders. The authors conclude that the Bayesian approach has the advantage of naturally allowing for full uncertainty, especially for prediction. They note that “An advantage of a Bayesian approach to random-effects meta-analysis over a classical implementation of the same model is the allowance for all uncertainties, particularly in obtaining a predictive distribution for the true effect in a new study. This may be important when there are few studies in the meta-analysis.”

Performing a FB analysis is, therefore, well suited to meta-analysis because an FB model emphasizes both estimation and prediction of parameters alongside the degree of uncertainty in reporting these estimates (National Research Council, 1992). The FB approach, through its
estimation of parameters using probability distributions, takes into account the uncertainty involved with estimating all model parameters (Raudenbush & Bryk, 2002). FB is different than parameter estimation using the RE approach in which $\tau$ is assumed known, when it is not. Using RE models may lead to misestimating the treatment effect size and the incorrect identification of its significance (DuMouchel & Normand, 2000). This especially occurs when the number of studies included is small, as demonstrated in one comparative study (Spiegelhalter, Abrams, & Myles, 2004).

DuMouchel (1994) conceptualizes a FB hierarchical meta-analysis model, without covariates, using the following equation:

$$Y_i = \mu + \delta_i + \epsilon_i$$

Where:

- $Y_i$ = the observed effect size for the $i^{th}$ study, $Y_i \sim N(0, \sigma_i^2)$
- $\theta_i = \mu + \delta_i$, the study-specific parameter
- $Y_i | \theta_i \sim N(\theta_i, s_i^2)$
- $\delta_i$ = the random effects, $\delta_i \sim N(0, \tau^2)$
- $\epsilon_i$ = the sampling error associated with $Y_i$, $\epsilon_i \sim N(0, s_i^2)$

DuMouchel (1994) includes study-level covariates represented by the following equation.

$$Y_i = [X_i \beta + \delta_i] + \epsilon_i$$

$$\theta_i = X_i \beta + \delta_i$$

$X_i \beta$ replaces $\mu$ and represents a linear combination of between-study effects, or:

$$X_i \beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij}.$$ 

Therefore, a hierarchical FB meta-analysis contains three sources of variation (see DuMouchel, 1994): 1) $s_i$, the within-study random error, assumed to be known; b) $\beta$, the between-study fixed
effects and c) $\tau$, the between-study random error. The accurate estimation of $\tau$ is important for assessing the uncertainty surrounding $\mu$ and in the prediction of $\theta_i$.

Just as with the random-effects model, a FB model is a mixed model with two error terms, accounting for both the within-and-between-study variation. One feature that sets the FB method apart from a RE method is its use of prior information, and so the model is flexible if new evidence comes into play (Sutton & Abrams, 2001); an attractive approach when considering multiple studies for a research synthesis. When it comes to estimating unknown model parameters, a RE model aims to consider all sources of variation, but it treats the unknown parameter, $\tau^2$, as known, whereas a FB method treats all unknown parameters as such (DuMouchel, 1994). Concretely, probability distributions are assigned to all coefficients and the two error terms so that the uncertainty in these estimates can be properly assessed using Bayesian credible intervals. This is most important when the number of studies included in a meta-analysis is small (Higgins et al., 2009).

While a FB meta-analysis is a useful technique, it does come with a few limitations. Sutton and Abrams (2001) first discuss that the use of prior information in the model does not lend itself to an objective analysis, with very few guidelines currently available to help the Bayesian analyst make decisions. Second, the analysis can be computationally complex, resulting in a time-consuming effort on the part of the analyst who may not even be familiar with FB methods. Furthermore, while the mention of Bayesian analysis as a practical approach—particularly in the medical field (Burton, Altman, Royston, & Holder, 2006)—has increased, work is needed to make it more prevalent in the literature (Ashby, 2006).

To address the computational complexity of performing a FB model, Abrams and Sanso (1998) proposed using approximations of the first and second moments of the parameters of a
Bayesian meta-analysis, providing an efficient method for estimation from the posterior distribution. The authors test this by taking two published examples of meta-analyses which were previously performed using a fixed-effect model—one from a series of studies showing extreme between-study heterogeneity ($\chi^2 = 58, df = 21$) and the second not ($\chi^2 = 5.4, df = 8$)—and show that they lead to “sensible approximations” and “broad agreement” with the traditional Bayesian approach using Gibbs sampling. For studies that are small and unbalanced, however, the approximations over-shrink study effects and under-shrink the variance estimates.

Carlin (1992) develops and implements a FB approach to meta-analysis, in which uncertainty about effects in distinct but comparable studies is represented by an exchangeable prior distribution. Hierarchical normal models are used and illustrated with two examples; the first involving a dataset on the effect of beta-blockers after myocardial infarction, and the second based on a dataset comprising 14 case-control studies on the effects of smoking on lung cancer. Different conclusions from those previously obtained were found for both studies. Specifically, the Bayesian analysis produces different conclusions in terms of the width of the confidence intervals and some of the point estimates.

Bayesian methods allow the researcher to fully model all parameter uncertainty, make probability statements directly about quantities of interest, and incorporate prior knowledge into the analysis (Sutton & Abrams, 2001). These methods lead naturally into a decision-theoretic framework, considering cost and utilities when, for example, making important healthcare and policy decisions. Simulation studies comparing the performance of FB versus RE methods may allow a researcher to show these advantages when performing a meta-analysis. Furthermore, simulation studies can show the conditions under which performing a computationally complex procedure like a FB method is warranted when assessing a series of combined studies. To date,
studies comparing a FB method to a RE method (e.g., Browne & Draper, 2000; Browne & Draper, 2006) suggest that FB methods outperform the likelihood-based RE method—defined by evaluating the bias of the point estimates and nominal versus actual coverage in interval estimates—sometimes by a very wide margin. Studies comparing RE versus FB methods, however, are few, and most are strictly comparative studies that do not systematically vary study conditions to assess areas of concern (e.g., concerns surrounding the number of studies included).

**Bayesian credible intervals.**

In Bayesian theory, the range containing 95% probability is known as the posterior, or credible, interval. It is distinguished from the traditional, frequentist confidence interval in that it has a more straightforward interpretation where the range includes the value of the true intervention effect, $\theta$, with a 95% probability. This has been what is erroneously interpreted with the frequentist confidence interval, where the actual meaning is that if the same study were to be conducted 100 times, 95% of the time, the results will include the true value of the underlying effect (Spiegelhalter et al., 2004). Additionally, with the inclusion of the prior information, the resulting credible interval can be in a range that is more realistic (i.e., wider) because it is incorporating all possible sources of variation.

**Prior distributions in fully Bayesian random-effects meta-analysis.**

The Bayesian statistical method uses prior information, or beliefs about how the data are distributed, combined with information from observed data to make statistical inferences. In Bayesian statistics, all unknown parameters are assumed to have a probability distribution. For a Bayesian hierarchical meta-analysis, priors must be chosen for the individual study effect sizes ($\theta_i$), the weighted mean effect size estimate ($\mu$), and the variation for the random effects ($\tau^2$).
The choice of a prior distribution for each parameter affects the conclusion a researcher makes of results from the posterior distribution (Lambert, Sutton, Burton, Abrams, & Jones, 2005); prior distributions affect the credible interval estimates for both $\mu$ and $\theta$ as well as the amount of shrinkage imposed on $\hat{\theta}_i$ (Pauler & Wakefield, 2000). The prior distribution that is chosen for $\tau^2$ is especially important because it provides a technique for determining the similarity across studies and whether information can be borrowed from the group of studies included in the analysis (Greenhouse & Iyenger, 2009).

There are three classes of priors that may be chosen for a FB analysis. The first is the noninformative class, which is used when relatively little is known beyond the data that are available for analysis. This uncertainty is manifested using a flat prior and is expressed in a Bayesian analysis via the uniform distribution which, formally, means that the posterior distribution will have the same shape as the likelihood, making the results of a Bayesian analysis like those of a classical analysis (Spiegelhalter et al., 2004). A second class of priors is known as weakly informative, which Gelman (2006) describes as “…proper but set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available.” (p.517) A third class of priors is the strongly informative prior probability which, as the name suggests, requires that much must be known about the parameters of interest, such as from previous studies deemed relevant to the current data (Browne & Draper, 2006). Although not necessary to qualify it as such, a strongly informative prior is one where the information contained therein can dominate that provided by the current data being analyzed, so that the posterior distribution can get most of its information from the prior and therefore would look much the same.
Browne and Draper (2006) designed a simulation study comparing likelihood-based methods to Bayesian methods using two diffuse prior distributions—the uniform, $U(0, \frac{1}{\varepsilon})$, where $\varepsilon$ is a suitably small value, and inverse gamma, $\Gamma^{-1}(\varepsilon, \varepsilon)$—for the variance components of a multi-level model (see Browne & Draper, 2006, p. 9, for a discussion of the choice of $\varepsilon$). These were evaluated using the bias of point estimates and comparing the nominal versus actual coverage of confidence intervals during repeated sampling. For the two-level variance components model, both likelihood-based and Bayesian methods produce unbiased estimates but have a problem with nominal coverage for small samples (i.e., small number of studies). As for a three-level RELR model, quasi-likelihood methods perform poorly when estimating random-effects variances in terms of bias and interval coverage, while Bayesian methods, using diffuse prior information led to well-calibrated (i.e., a Bayesian performance measure for bias) point and interval estimates. This study demonstrates the usefulness of FB methods and the use of prior information in a hierarchical analysis.

Browne and Draper (2006) also explain that constructing diffuse priors for RE variances results in discrepancy in performance across models. Consider the case where hundreds or thousands of participants are nested within a much smaller number of studies. The likelihood information about between-study variation is limited. The choice of prior to be used in this case can seriously influence the results, so a sensitivity test should be performed to assess how results change as the prior changes. Another avenue for exploration is the consideration of the best prior for which to use in the event of small level-2 sample sizes.

In another simulation study (Lambert et al., 2005), the importance of the number of studies combined when choosing a prior distribution is demonstrated, where the effects of 13 different diffuse prior distributions on a random-effects meta-analysis were compared. After
varying the number of studies and values for the between-study variances, when there were just five studies, the estimates of the between-study variance changed depending on the prior distribution. The effect size estimates were not biased, but the precision with which it was estimated—again, when the number of studies is small—varied depending on the prior. The choice of prior was less crucial when the number of studies included was large (i.e., 10 or 30 studies).

Gelman (2006) makes a few recommendations when performing hierarchical Bayesian analyses. First, start with a non-informative prior, which will generally work for most situations, unless the number of i groups is small—for example, less than five—which could yield overestimates of the variance; an unavoidable circumstance to restricting the variance to be positive. If it is desired to restrict the variance even further—for instance, to prevent it from being too large—then the half-t family of priors is recommended because it is more flexible and performs better for values of τ near zero. The half-Cauchy family is a good starting point, with “a value that is high, but not off the scale.” (Gelman, 2006, p. 528) For his study, Gelman parameterizes the half-t in terms of scale A and degrees of freedom ν:

\[ p(\tau) \propto \left( 1 + \frac{1}{\nu} \left( \frac{\tau}{A} \right)^2 \right)^{-(\nu+1)/2} \]

The scale A was set to 25 and it was found that using the weakly-informative half-Cauchy on τ performed better when the sample size on level 2 is small (N=3) when compared to a non-informative uniform(0, ∞) prior.

Other commonly-used prior distributions are the weakly-informative half-normal and DuMouchel priors. The half-normal prior for τ is known as a subjective prior because researchers may choose the upper and lower bounds for the possible values for τ (Spiegelhalter et al., 2004). As the name suggests, the half-normal distribution represents the positive half of a normal
distribution with a mode at 0. Gelman (2006) recommends that if a researcher wishes to choose a
proper noninformative prior, it should be the half-normal with mean of 0 and a high value for the
standard deviation, such as 100.

DuMouchel's prior (DuMouchel, 1994)—also known as the Pareto prior (Pareto, 1895)—
represents a proper prior (i.e., distribution integrates to 1) for $\tau$ that satisfies requirements of a
true probability distribution for $\pi(\tau)$. Assuming a log-logistic prior distribution,

$$
\pi(\tau) = \frac{s_0}{(s_0 + \tau)^2}
$$

where

$$
s_0^2 = \frac{K}{\Sigma s_i^{-2}}
$$

$s_0$ is the median of the density and $s_0^2$ is the harmonic mean of K sampling variances, $s_i^2$. $s_i^{-2}$ is the precision for each study. DuMouchel (1994) explains that this prior is highly dispersed
because the expectations for both $\tau$ and $1/\tau$ are infinite, which is an advantage of using this prior.
It protects against values skewing toward large values of $\tau$ (Sutton & Abrams, 2001).
Additionally, because this prior has a maximum probability of the value for $\tau$ at 0, it allows for
more realistic values of $\tau$ to be more likely. Finally, $s_0$ tends to be weighted toward smaller
values of $s_i$.

**Random Effects versus Fully Bayesian Approaches to Meta-analysis**

A RE approach is an improved method over a FE analysis because it accounts for
variation between studies—that is most assuredly present—with the assumption that there is a
distribution of effect sizes and the mean of that distribution is estimated. While the FE model
provides the cleanest and simplest way to address the statistical problem to make inferences
(Morris & Normand, 1992), it is likely that the assumption of homogeneous true effects is
untearnable, and so results following this method could be misleading or overly optimistic.

Hierarchical (RE) models are recommended in this case (e.g., Raudenbush, 2009).

RE methods can, in the presence of heterogeneous true effects, do the following: (a) Calibrate and improve upon methods already used for meta-analysis, e.g., with small sample sizes; (b) Adapt existing methods and software for random effects models and for borrowing strength to the needs of meta-analysis; (c) Identify other models needed for meta-analysis and extend the collection for analyzing these models; and (d) Develop and utilize new perspectives for meta-analysis, such as using a collection of studies to strengthen estimation of the true effect of a particularly important study, or to determine whether that study has a significant true effect. Despite their value, however, RE models (especially using EB techniques) are underutilized.

The problem arises when a small number of studies are included in the meta-analysis; then $\tau^2$, the between-study variance is estimated with poor precision, along with other caveats summarized above and discussed in Borenstein et al. (2010). The FB approach, through its modelling of uncertainty of all model parameters, including $\tau^2$, may be the solution for estimating with greater accuracy the pooled effect size estimate. One issue with the FB approach, however, is the general unfamiliarity among researchers of this method. An important question that has not been systematically answered is the extent to which a FB method improves estimation over and above the RE method, the goal being to describe the conditions or characteristics of a meta-analysis in which it is recommended to take the time and effort to perform a FB analysis. Very few have compared RE to FB methods, and those that do are not conducted via a simulation study, in which the performance of statistical methods is systemically assessed and compared against a known truth (Burton et al., 2006). A summary of the comparative studies that have been done is discussed next.
Sutton and Abrams (2001) compare fixed classical and Bayesian as well as random classical and Bayesian meta-analytic methods using studies related to electronic fetal heart rate monitoring (EFM) systems and their relationship to perinatal mortality to illustrate the ability of such methods to consider the totality of evidence. Different study methodologies (randomized-controlled trial (RCT), comparative cohort, and before and after studies) were also discussed. After combining RCTs, the treatment effect for each individual study based on the posterior distributions for the $\theta$s using a Bayesian random effects model produce shrunken estimates due to strength being borrowed by other studies and CIs that are narrower than the observed estimates. Trials where the treatment differences were measured least precisely are shrunken most towards the overall pooled treatment difference. A meta-regression attempting to account for heterogeneity between studies shows that there is an association between the year of publication and risk difference, suggesting the benefit of EFM has increased over time. When 4 Bayesian models using a vague prior are compared using Bayes Factors, more complex models (random effects, fixed effects regression and random effects regression) are moderately better-fitting than a fixed effects model, with the BF between the random effects versus fixed effects regression being the largest and therefore best model.

Su and Po (1996) compared EB, FB, and classical methods for pooling event rates from separate epidemiological studies. Four datasets were evaluated; the first was a series of clinical trials measuring the effect of beta-blockers on death rates; the second deals with smoking and lung cancer in case-control studies; the third, the effect of an antihistamine compound called chlorpheniramine on drowsiness; and the fourth is the use of intravenous magnesium and its effect on suspected myocardial infarction. To test the methods further, randomly-chosen data points were made to be more extreme. The main outcome measures are the pooled estimates of
the effect expressed as odds ratios and their 95% confidence intervals. The study found that the Bayesian methods provided wider confidence intervals. Additionally, the point estimates for the individual studies were substantially different, especially for small studies. Overall, the Bayesian methods gave results that were consistent with the classical fixed-effect method and introduction of randomly extreme data points did not change these results. The authors do acknowledge that, at present, there are no studies to show, unequivocally, the benefits that have been purported from performing a more computationally complex FB meta-analysis.

Another study (Biggerstaff, Tweedie, and Mengerson, 1994) describes and compares current classical (Fisher’s exact, logit, and Mantel-Haenszel), EB, and FB methods, applying them to published studies of the relative risk of lung cancer associated with exposure to environmental tobacco smoke (ETS) in the workplace. The authors found that, although all methods give reasonably similar combined estimates of relative risk of lung cancer associated with exposure to ETS – which was not found to be greater than chance, in either the frequentist or Bayesian framework – the approximations arising from classical methods appear to be non-conservative and should be used with caution. The Bayesian methods, which account more explicitly for non-homogeneity in studies, give slightly lower estimates of relative risk and wider posterior credible intervals, indicating that inferences from the non-Bayesian approaches might be overly optimistic. The main concern here is modeling between-study variation, i.e., \( \tau^2 \). Also of concern is the confidence interval or Bayesian credible interval and how classical methods tend to reveal overly optimistic (i.e., narrow) ranges for the estimates. Results also suggest the need to find and use appropriate covariates explaining heterogeneity.

In another comparative study, Smith et al. (1995) attempt to show that a FB meta-analysis can naturally deal with a multitude of issues regarding the combining of information, including
the choice between fixed and random-effects models, the choice of population distribution in a random-effects analysis, the treatment of small studies and extreme results, and incorporation of study-specific covariates. The authors derive appropriate proper prior distributions and perform a sensitivity analysis testing a variety of prior assumptions. Results from main non-Bayes methods are summarized and compared to the FB estimates. After comparing various fixed versus random-effects methods, it was found that 95% CIs were narrower for the random effects compared to fixed effect for the individual study effects but they were wider for the pooled estimate. This is because some of the within study variation in the fixed-effect analysis is accounted for as between study variation in the random-effects analysis. After pooling the random-effects analysis, a greater treatment effect is estimated. This is because the RE analysis places less weight on larger studies than FE models. When comparing all FE methods for binary-response randomized clinical trials, all led to similar estimates, but the Peto method—a technique for estimating the conditional likelihood of $\theta_i$—is very conservative, logistic regression does not perform well for proportions at 0 or 1, and the fully Bayesian method produces similar results to Woolf’s technique which uses the maximum likelihood estimate of the log-odds ratios in each study and calculates a pooled effect by weighting each study’s precision (i.e., the precision of its variance). After comparing all RE methods, it was found that logistic regression underestimates uncertainty. The FB estimate was about one standard error more extreme than with any other technique, because it was strongly influenced by two studies with the most extreme values.

As discussed, these comparative studies do not systematically vary the number of studies, their sample sizes, the presence of covariates, and extreme values (e.g., proportions near 0 or 1),
and other general imbalances to see, under what conditions, whether the classical approach does as well as the FB approach, and under which conditions should the FB method be employed.
Chapter III: Simulation Design and Method

Aims and Objectives

Theoretical discussions justify the use of a FB method exist to account for all uncertainty involved with parameter estimation. However, because it is a method unfamiliar among researchers, and can be complex, it would add to the current knowledge in the field to determine under which circumstances an FB method is most appropriate when implementing a meta-analysis. This study proposed a simulation to systematically vary study heterogeneity (τ), number of studies, and number of participants per study for a simulated dataset having a continuous outcome, \( y_i \) (e.g., SAT scores) and a known effect size, \( \theta \), of an often-tested treatment (say coaching versus a non-coaching control). These conditions were tested using classical fixed-effect (FE) and random-effects (RE) methods of EB and FB described in chapters I and II to assess whether—and for which conditions—each method produces significantly different results. Specifically, the FE and RE methods were compared to the FB method by assessing parameter estimates, and their bias and variance calculations. It was expected that RE would perform better than the FE method, especially in the face of large between-study heterogeneity. It was also expected that the FB would outperform the RE method, particularly for scenarios in which the number of studies was small (\( N \leq 5 \)).

Simulation Procedures

Parameter estimates and their performance measures were assessed in a 7 (\( K_j \)) x 3 (\( N_j \)) x 4 (\( \tau_j \)) factorial design, whereby study sample size, number of studies, and between-study heterogeneity were varied, and each scenario was tested using different meta-analytic methods (i.e., FE, RE, and FB). The model described in Chapter I and copied below was specified as a meta-analysis.
that aims to assess the effect of coaching (versus no coaching) on SAT scores, with a combined
Level 1 and Level 2 equation:

\[ d_i = U_i + \varepsilon_i, \quad i = 1, \ldots, k \]

Where:

- \( d_i \) is the estimated standardized mean difference in study \( i \) between the coaching and no coaching groups.
- \( U_i \) is the random error, \( U_i \sim N(0, \tau^2) \)
- \( \varepsilon_i \) is the sampling error, \( \varepsilon_i \sim N(0, \sigma^2) \) # Note: \( \sigma^2 \) is also identified as \( \nu_i \) and \( s_i \) in previous chapters

Table 1 lists the scenarios tested for each meta-analytic method. Again, the goal was to
determine the conditions under which a method of analysis strays from the truth, and may serve
as a guide to inform which methods are more likely to lead to the greatest performance issues
(see Chapter II). These simulation conditions were adopted partially from those described by
Lambert et al. (2005), Mittlböck and Heinzl (2006), and Trikalinos, Hoaglin, and Schmid (2013)
in their studies assessing the performance of various FB models, models with binary outcomes,
and sample sizes for multilevel logistic regression models. The number of studies to be
combined is manipulated at 7 levels (3, 4, 5, 6, 9, 16, and 25) to simulate a statistical
combination of a range of “small” to “large” meta-analytic studies. The level of heterogeneity
between studies (\( \tau^2 \)) is manipulated at four levels, (.001, .2, .5, .8) representing near-zero, small,
medium, and large between-study variance. A between-study variance of .001 would signify that
almost no true heterogeneity exists, and a fixed effect would be appropriate (Lambert et al.,
2005). Varying the level of heterogeneity determined under which circumstances it is most
important to account for the uncertainty in the estimate of the between-studies variance. The
number of participants per study is evaluated at three levels by varying the heterogeneity within
a study (\( \sigma^2 \)) as a proxy (.05, .1, and .2). Assessment of the effect of the number of studies
included in a meta-analysis determined under which scenarios the reliability of combining information seriously influenced parameter estimation and the ability to make sound interpretations about a treatment’s real effect. In the case of the RE method, it determines how the weights are calculated, which also affects the effect size estimates. Varying the heterogeneity within studies as a proxy for the number of participants per study represents the change in the homogeneity within studies. That is, large sample sizes within studies produces smaller within-study variance, $\sigma_i^2$, which influences parameter estimation.

**Software to perform simulations**

Independent datasets were randomly generated for each of the 84 scenarios using the R statistical software program (R Core Team, 2016). A starting seed was chosen and remained fixed to allow for future replications of the simulation, if necessary. As mentioned and discussed in further detail below, three separate models were compared using the same simulated data.

If a simulated analysis fails to run due to, for example, convergence issues, the failure was recorded, sample discarded, and procedure repeated. The study design allows for the possibility that for some simulations that fail, post hoc changes of the protocol will be made and certain scenarios that do not run are omitted (Burton et al., 2006) so that the reliability of the simulation is not compromised. Summary statistics were calculated to ensure that a realistic dataset is being generated for the simulation runs.

**Statistical Methods Evaluated**

The fixed effects (FE), random effects (RE), and fully Bayesian (FB) techniques for meta-analysis are compared using the R statistical software program (R Core Team, 2016). For the FE and RE methods, the meta-analysis package *metafor* (Viechtbauer, 2010) was used. Within *metafor* the *rma* modeling function (*rma* stands for random effects meta-analysis) uses
the syntax `rma(yi, vi, method=, ...),` where `yi = {effect sizes}` for the K studies, `vi = {variances}` within and between the K studies, and `method = {type of model approach}`.

For the FE method, the `rma method = “FE”` is specified; for RE it is `method = “REML”` (an EB estimator). As discussed in Chapter II, REML is just one method for estimating $\tau^2$; the other common method being DerSimonian and Laird (D-L; Method of Moments). While one advantage to using the D-L method is that it does not require making distributional assumptions for the random effects, REML, being a likelihood-based approach, performs more efficiently (Thompson & Sharp, 1999). According to Raudenbush (2009), REML is based on the likelihood of the between-study variance, $\tau^2$, given the data, while integrating out the regression coefficients ($\beta$). REML offers an attractive balance between unbiasedness and efficiency (Viechtbauer, 2005) because of its accounting for the uncertainty in $\beta$—unlike the full maximum likelihood (FEML)—and providing estimates that are efficient—unlike the D-L Method of Moments.

The FB method was fit by the `hblm` package in R, a program written by DuMouchel (1994), with functions adapted from S-plus. The package uses the syntax `hblm(es, se, prior=list(…))`, where `es = estimate of effect size for each study`, `se = standard error for each study`, and `prior = specified for each parameter in the Bayesian model [in this case, only a prior for tau (between-study) variation will be specified]`

For purposes of this study and based on prior research, three types of priors for $\tau$—one non-informative and two weakly informative—are utilized for the current study and their performance will be tested under the different scenarios discussed above. The first is the noninformative uniform prior; the second and third are weakly informative half-Cauchy DuMouchel/Pareto priors.
For \( y_i \), a normal distribution, \( N(\mu, s_i^2) \), is specified with hyperparameter \( \mu \) specified as \( \mu \sim N(0, \tau^2) \) and testing three alternatives for \( \tau \) (described in Chapter II). A Bayesian analysis assigns prior distributions to \( \mu, \tau, \) and—if there are study-level covariates—\( \beta \). The hblm package has default priors for \( \mu \) and \( \tau \). However, because one aim for this study is to test alternative priors for \( \tau \), a nearly uniform prior will be specified using the `prior=list(tau=function(tau) 100/(100+tau)^2)` command—a variation of DuMouchel's prior—within the hblm package. Additionally, an approximation of the half-Cauchy distribution will be specified as another prior for \( \tau \) using the `prior=list(tau=function(tau) (1/(pi*(1+tau^2))) * (tau>0)` command within the hblm package. The scale for the half-Cauchy is ignored here because values for \( \tau \) will not be larger than 1. The third and final prior for \( \tau \) is the DuMouchel, or Pareto, prior, \( \pi(\tau) = s_0/(s_0 + \tau)^2 \), the default prior within the hblm package. The priors have been constructed so that the value of \( \tau \) was of a restricted range (i.e., \( \tau > 0 \)), because the real values of \( \tau \) are small. This was attempted because (a) it was initially done to be exploratory within the previously determined hblm function code and (b) because of the substantial number of iterations, it was believed that the estimate would have sufficient time to converge to the small values of \( \tau \) by the end of the simulation runs.

Performance of the estimation methods were evaluated via assessment of bias and overall accuracy (Burton et al., 2006; Walther & Moore, 2005). Estimation bias, also known as systematic error, is defined as the deviation in the average of repeated estimates from its true value:

\[
\delta = \hat{\beta} - \beta
\]

Where:
\( \hat{\beta} \) = the estimate of interest (here it is for \( \mu \) and \( \tau \))

\( \beta \) = the true value (i.e., what has been simulated).
In a typical research study, the true value of the bias is never known; this is one reason conducting a simulation study is advantageous. The overall accuracy, or precision, is assessed by calculating the variance of both $\tau$ and $\mu$ estimates. Where helpful and informative, the mean squared error (MSE) using $(\hat{\beta} - \beta)^2 + (SE(\hat{\beta}))^2$, is reported, but it is not the primary focus when reporting results. The MSE is defined as the distance between the estimates and the true value (Walther & Moore, 2005). It is a way of combining both the bias and precision to get the overall performance of an estimation method in one statistic. A third set of measures, lower (2.5%), middle (50%), and upper (97.5%) quantiles of the frequency distributions of each simulation’s results are reported. This takes the proportion of the distribution that lies below the $i^{th}$ observation. If the quantile is equal to $q$, in a series of $n$ ordered observations, value $i$ is calculated as $i = q(n + 1)$. Plotting quantiles is an easy and direct method for summarizing the frequency of a distribution of results (Bland, 2000); in this case, the frequency of models estimating $\tau$ and $\mu$ across simulation runs.

Each of these performance measures will be discussed, in turn, for both the $\mu$ and $\tau$ estimates stored among the 84 simulation runs of 1000 iterations.

Chapter IV: Results

A total 84 scenarios were run—simulating a meta-analysis varying the number of studies, degree of heterogeneity between studies ($\tau^2$), and sample size within each study ($\sigma_j^2$). Note that, before this chapter, we’ve been discussing the variance between and within studies; $\sigma_j^2$ and $\tau^2$, respectively. Depending on the model specified within the R software (rma vs. hblm), either the variances ($\tau^2, \sigma^2$) or standard deviations ($\tau, \sigma$) are simulated and/or estimated. For consistency in reporting, only the standard deviations are reported below.
The results of each simulation run by estimation method—fixed (FE), random (RE), fully Bayesian (FB) using the default prior, Pareto, followed by specification of an approximate half-Cauchy and uniform prior on \( \tau_j \)—are reported in the Appendix. As proposed in Chapter III, the estimates of \( \mu \) and \( \tau \) were stored and their bias, variance, and quantile intervals reported. As will be reported below, some unexpected mean results were recorded; therefore, the median of the simulation results was monitored, and median squared errors were calculated. They were found to be identical to the mean calculations, so they are not included here, but it indicates that something both unexpected and unorthodox was occurring with the simulations, although the cause was never determined. See Figures 1-6 for a review of the mean, variance, and quantiles for the estimates of \( \mu \) and \( \tau \). They are reported by the number of \( K \) studies and across each simulation run. A summary of the results is discussed next.

**Results for the estimate of \( \tau \)**

Before reporting the bias and variance calculation for the mean of the \( \tau \) estimates, a general note: the FE method does not estimate a between-study error variance—these models assume that \( \tau = 0 \) (because the FE model assumes that there is one fixed value for the true effect size and not a distribution of effect sizes)—so the bias is equal to the value of \( \tau \). The results for FE, therefore, will not be reported in the summary below and are excluded from Figures 1-2 and Figure 5.

Figure 1 shows the mean estimates of \( \tau \) by \( K \) across each simulation run. RE underestimates \( \tau \), particularly when \( K \) is small, however the trend remains consistently negatively biased across simulation runs and for all values of \( \tau \). When values of \( \tau \) are lowest (i.e., .03 and .45), mean estimates across runs are closest to zero. When \( \tau = .71 \), mean values are between .25 and .45; lowest when \( \sigma = .45 \). Mean values range from .6 to about .7 when \( \tau \) is at its highest
value, .89, and are closest to its true value, not surprisingly, when $K$ is greatest. When $\tau=.89$ and $\sigma=.45$, the mean estimates for $\tau$ hover around .5 for lower values of $K$, almost half of its real value. The variation around these mean values (see Figure 2) is greatest for lower values of $K$ (i.e., $K < 16$), with a maximum MSE = .65. However, variation is at or near zero for all values of $K$ when $\tau$ is equal to .03 and .45. Figure 5 shows the lower, middle, and upper quantiles of $\tau$ estimates across the 1,000 iterations for each simulation run, reporting the full range of $\tau$ values across $K$. It again shows how, for lower true values of $\tau$, estimates for $\tau$ remain at or near zero. When $\tau = .71$, the range is from 0 at the lower quantile, .2 in the middle, and about 1.0 at the higher end. For $\tau = .89$ the range is zero at the lower end, to about .6 in the middle, and 1.5 at the highest. For all runs, the range is widest for smaller values of $K$.

As for the FB models, those specifying half-Cauchy and Pareto priors show comparable results where, for small values of $K$, $\tau$ is underestimated when the actual value is large, yet overestimated for smaller values. However, as $K$ increases, $\tau$ estimates either reach or nearly approach the actual value of $\tau$. It is for smaller values of $\tau$ (i.e., $\tau = .03$) that the half-Cauchy and Pareto priors have the most trouble. However, as Figure 1 shows, the half-Cauchy prior reports results for mean of $\tau$ approximately equal to the true value of $\tau$, especially when $K$ is large, and $\tau = .45, .71, \text{and} .89$. When $\tau = .03$, the mean of $\tau$ across runs is positively biased, particularly when $K$ is small, and this was consistently the case, as the variance and quantiles of the $\tau$ estimates show (see Figures 2 and 5). There is some variation in the $\tau$ estimation. For small true values of $\tau$, values fall from about .2 at the lower quantile, .1 to .4 at the middle quantile, and .1 to about .9 at the upper quantile. For larger true values of $\tau$, the range is .2 to .6 at the low end, .5 to about .8 in the middle, and 1.0 to about 1.7 at the higher end of the distribution (see Figure 5).
The FB model specifying the uniform prior revealed starkly different results than with the half-Cauchy and Pareto priors where, for small values of $K$, $\tau$ estimates were more biased than those found with all other models, and are highly positively biased, as opposed to the negatively biased results of the RE, half-Cauchy, and Pareto methods. Results become less biased for $K$ values $\geq 9$. Again, as for the half-Cauchy and Pareto priors, the bias persists when $\tau = .03$, with a mean value about 13 times greater than the true value when $K$ is small. However, even when $K$ is large, at $K = 25$, the mean value for $\tau$ is about 7 times greater than the true value of .03. Both Figures 2 and 5 reveal that the variation in the $\tau$ estimates is greatest with the uniform prior on $\tau$, where the variance reaches as high as .35 for small values of $K$ (greatest MSE = .94), but becomes at or near zero for $K = 16$ and 25. The quantiles plotted in Figure 5 show that values for $\tau$ range from .1 at the lower, to 1.2 in the middle, and 3.0 at the upper quantile, when the true value for $\tau$ is highest at .89; about 70% greater than the true value.

As expected, the $\tau$ estimates are more biased, and less accurate, when $K$ is small, but both the FB models with half-Cauchy and Pareto priors specified for $\tau$ seem to perform better overall, producing less biased results, especially when $\tau$ is greater than .45. The FB models perform better when $K \geq 9$ compared to RE, except when $\tau = .03$ and $\sigma = .45$. It seems that RE estimates $\tau$ best when values are smallest; the FB models assume that the variation between studies is larger and results show that it will assume that $\tau$ is larger than zero, but the priors on $\tau$ won’t allow it to get too large. The variation in the estimates is greater for FB with uniform prior, followed by RE, then FB with half-Cauchy and Pareto priors. However, the RE estimates show the greatest variance overall when $\tau = .89$. 
Results for the estimate of $\mu$

The means for the estimate of $\mu$ are similar for the FE and RE models and identical for the FB models. This is because $\sigma$ was simulated to be equal across the “studies” included in each meta-analysis, so that weights are applied in the same way, no matter which method is used. Additionally, all models were run beginning with the same starting seed. Consequently, the estimates stored are identical, or very nearly so, for each pair of estimation methods.

Figures 3 and 4 summarize the mean and variance for the estimates of $\mu$. For the most part, $\mu$ estimates do not stray too far from the true value (i.e., intervention has no effect and $\mu = 0$). However, as expected, all models show the greatest bias when $\tau$ is large, although overall values of bias, both negative and positive, are higher for more runs of the FB models, compared to FE/RE (see Figure 3).

The FE method reports consistently greater bias when $\tau$ is largest ($\tau = .89$; see Figure 3) and when $K$ is smallest. They also seem to bounce around from being negatively biased to positively biased, with no discernible pattern. However, the runs with the greatest bias, both positive and negative, occurs when $\tau$ is highest at .89. As for the variation across estimates (see Figure 4), values are greatest when $K$ is small, with greatest variance consistently across runs where $\tau = .89$. The ranges reported in the quantile plots start from -1.0 to 0.8 at the upper end (see Figure 6).

The mean values of $\mu$ estimates for the RE method hover more around the true zero value than the FE method, where the means range from -.02 to .02. Some more extreme outliers, however, are shown when $\tau = .71$. The variation across estimates is significantly greater compared to FE, particularly for large $\tau$ and $K = 3$ and 4. The lower, mid-, and upper-quantile values show similar ranges as FE, from -1.0 to 0.8 (see Figure 6).
For FB methods, like the FE/RE methods, the mean value for $\mu$ bounce around the value, with some positive and negative bias, and no discernible pattern across simulation runs. Also, like FE/RE, the most extreme bias, and variation of the $\mu$ estimates are greatest when $\tau$ is large. The lower, mid-, and upper-quantile plot ranges from -1.0 to 1.0; wider than the FE/RE methods.

Overall, though, the accuracy—as defined as the distance from the true value of the effect size—of the $\mu$ estimates for the remaining scenarios is slightly better for FE/RE. The RE model reports greater variation across iterations compared to FE or all FB models, although all models perform consistently when $K \geq 16.$
Chapter V: Discussion

In Chapter I, it was noted that performing a FB meta-analysis has its challenges. Not so much technical challenges, but with setting up the model with proper priors; decisions for which may influence the results of the analysis. Therefore, it is worthwhile to undergo an assessment to determine the extent to which a FB method improves estimation over and above a RE method for meta-analysis. No studies have yet systematically varied characteristics in a simulation to help answer this question. Comparative analyses have been conducted, with mixed findings. In a simulation study, a researcher can purposefully vary study characteristics to determine at what point, and for which combinations, does one method of analysis stray from the “truth,” so much so that a more complex method of analysis, such as the FB method, should be employed. This question is especially pertinent today, during a time where publications of meta-analyses is at an all-time high (Sutton & Higgins, 2008).

Previous comparative studies show that FE models are the most simple and straightforward to employ but assume conditions that are not realistic—namely that there is one true effect to estimate—and results are overly optimistic—that is, the standard error of the effect size estimate is underestimated, resulting in a narrow confidence interval and false claims that in intervention has a statistically significant effect when/if it does not. The RE models (including empirical Bayes methods) provide an improvement on the FE method, assuming heterogenous true effects and calibrating estimations by borrowing strength from all studies included. However, the RE models are run estimating the variation between studies, $\tau$, from the observed data and effectively assumes it to be known, so it tends to underestimate the uncertainty in the effect size estimate. When the number of studies included is small, the measure of the between-study variation will be imprecise (Borenstein et al., 2010) and it has been recommended that
Bayesian approaches be utilized in these scenarios. Studies systematically demonstrating the advantages of the fully Bayesian approach have not been conducted.

When running a fully Bayesian model for meta-analysis, the choice for a prior distribution on $\tau$ should be carefully made. One comparative study (Gelman, 2006) found that using a weakly-informative half-Cauchy distribution performs better than a noninformative uniform prior.

This study primarily aimed to answer the question of whether the FB method performs better than the RE approach, particularly in situations where the number of studies included ($K$) is small. Results from the simulations show that when $K$ is small and $\tau$ and $\sigma$ are large, RE and FB estimation methods have trouble estimating both $\mu$, the true effect size, and $\tau$, the between-study variation. However, as expected, the RE model was more optimistic—producing more narrow confidence intervals of the estimate of the true effect—when estimating $\tau$, as shown by the consistently underestimated, negatively biased results. FB methods perform better than RE when estimating $\tau$ when $K$ is small and, except when $\tau = .03$, estimates are most accurate when $K \geq 16$. When $K$ is at its highest value ($K = 25$) there seems to be the greatest differences between the RE and FB methods, in favor of the FB models.

Across FB methods, results are consistent with past comparative studies that have shown that the half-Cauchy performs better than the uniform prior on $\tau$. Gelman (2006) started with a non-informative prior, which works unless the number of $K$ groups is small (Gelman notes problems arise for $K < 5$) because they yield overestimates for $\tau$, the consequence when restricting the variance to be positive. Results suggest that problems persist for $K \leq 9$. The uniform prior produced estimates for $\tau$ that are most biased and less precise than any of the models run. The FB model specifying the Pareto prior on $\tau$ performs second-best to the half-
Cauchy in estimating the most realistic, and accurate, estimates of $\tau$, except when $\tau = .03$. Results also show that these two prior distributions are best when estimating $\tau$ when its value is neither too small nor too big, which may make sense given that their likelihoods allow $\tau$ to be larger, but not unrealistically so.

Among the $\mu$ estimates, RE tends to be less biased and more accurate for all variations of $\tau$ and $\sigma$, and as $K$ becomes larger. However, bias and precision, are lowest for simulation runs where the $\tau$ and $\sigma$ are highest (i.e., $\tau = .89, \sigma = .45$).

There were two especially unexpected findings reported. The first were the highly negatively biased estimates for $\tau$ when running the RE model and $\tau = .45$. This was consistent across all $K$ groups and did not improve significantly as $K$ increased. It’s unknown why the model failed to accurately estimate $\tau$, even when $K$ is large, and may be an area for future investigation. Conversely, the second unexpected finding was the extremely biased and inaccurate estimates among the FB methods in which $\tau$ and $\sigma$ are small (i.e., $\tau=.03, \sigma=.22$), even as the number of $K$ studies (i.e., $K = 25$). As with the RE model, it’s not clear why the model failed to eventually converge to the approximate value of $\tau$—and may be an opportunity for future analysis—but may be due to the incongruity of the study characteristics; the unlikely scenario where between-study variation is close to zero and within-study variation is small and larger than between-study variation.

Overall, findings are consistent with comparative studies that have been conducted. Results have shown that estimates are similar across methods, but when estimating uncertainty, RE methods tend to underestimate the heterogeneity between studies ($\tau$). This has been the justification for implementing the FB model, arguing that it can model the uncertainty with greater accuracy and is, therefore, the model of choice. Results of this study show that RE
models do tend to underestimate the value of $\tau$ by up to 41% more than with the FB models. However, when values for $K$ are small, the $\tau$ estimates tend to be more biased and less precise. The half-Cauchy and Pareto priors perform better than RE, while the uniform prior is highly positively biased. For larger values of $K$, estimates for $\mu$ and $\tau$ are only slightly improved.

The results from these simulations show that performing a meta-analysis including fewer than nine studies is generally not advisable, because even relatively small levels of heterogeneity will lead to biased estimations of $\tau$, no matter the statistical method. This will impact whether a significant treatment effect is found. FB methods perform best, although not without some imprecision. This is particularly relevant for studies in the medicinal sciences, where meta-analyses typically include five or fewer studies (Sutton & Higgins, 2008). For meta-analyses including more than nine studies, FB estimation methods generally perform well, except in situations where the heterogeneity between studies is very small (i.e., nearly zero). In the event a researcher knows this to be the case, perhaps the RE method should be employed. Overall, it does seem that the effort put into performing a FB meta-analysis may be worthwhile, but in some situations, RE performs just as well. If FB is employed, then it’s better to use the weakly-informative half-Cauchy distribution as a prior on $\tau$, where the posterior distribution more accurately estimates its value.

It’s arguable that this study raises more questions than it answers. Some results were unexpected, and it is unclear why, for scenarios in which $\tau$ and $\sigma$ are small, do the FB models perform so poorly. Future studies should replicate and explain these results. Furthermore, future studies should address, by systematically varying the sample size within each study, how weights are applied; especially because, unlike FE, RE models give greater weight to smaller studies. They should also investigate what happens when other issues such as publication bias are present.
because, as with other types of bias, it can lead to assumption violations where the true effect correlates with the error variance. While the present study showed that the half-Cauchy prior on $\tau$ performs well, future studies should explore alternative options for priors on $\tau$. Finally, a random effects models using an empirical Bayes approach to statistical inference, were discussed in Chapter II, but not directly evaluated in this study apart from using the REML technique, which is an EB estimator. An area of future study could be to compare multiple EB approaches and their results when running RE models in meta-analysis.
### Table 1. Criteria for scenarios of proposed simulation study.

<table>
<thead>
<tr>
<th>Study Characteristics</th>
<th>Description</th>
<th>Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_i$</td>
<td>Number of studies</td>
<td>3, 4, 5, 6, 9, 16, 25</td>
</tr>
<tr>
<td>$\sigma_i^2$</td>
<td>Within-study variance</td>
<td>.05, .10, .20</td>
</tr>
<tr>
<td>$\sigma_i$</td>
<td>Within-study standard deviation</td>
<td>.22, .32, .45</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>Between-study variance</td>
<td>.001, .20, .50, .80</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Between-study standard deviation</td>
<td>.030, .45, .71, .89</td>
</tr>
</tbody>
</table>

Note: The within-study variance for each study is assumed to be equal.
Figure 1. Mean of $\tau$ by number of studies across each simulation run – random effects and FB models

Note: For RE model, the variance, rather than the standard deviation, must be specified, therefore the square root of $\tau^2$ estimates are recorded here.
Figure 1. Cont.
Figure 2. Variance of $\tau$ by number of studies across each simulation run – random effects and FB models.

Note: There is some overlap in estimates across simulation runs. Therefore, some values are not shown.
Figure 2. Cont.

**Uniform Prior**

- tau: 0.89, 0.71, 0.45, 0.03
- sigma: 0.22, 0.32, 0.45

**Pareto Prior**

- tau: 0.89, 0.71, 0.45, 0.03
- sigma: 0.22, 0.32, 0.45

The graphs illustrate the effect of different tau and sigma values on the performance measure for K values ranging from 3 to 25.
Figure 3. Mean of $\mu$ estimate by number of studies across each simulation run – fixed effect, random effects, and FB models.
Figure 3. Cont.
Figure 4. Variance of μ estimate by number of studies across each simulation run – fixed effect, random effects, and FB models.

Note: There is some overlap in estimates across simulation runs. Therefore, some values are not shown.
Figure 4. Cont.
Figure 5. Lower (2.5%), median (50%), and upper (97.5%) quantiles of $\tau$ estimates by number of studies across each simulation run – random effects and FB models.

**Random Effects**

Note: For RE model, the variance, rather than the standard deviation, must be specified, therefore the square root of $\tau^2$ estimates are recorded here.
Figure 5. Cont.

Half-Cauchy
Figure 5. Cont.

Uniform Prior
Figure 5. Cont.

Pareto Prior
Figure 6. Lower (2.5%), median (50%), and upper (97.5%) quantiles of μ estimates by number of studies across each simulation run – fixed effect, random effects, and FB models.

Fixed Effect
Random Effects

Figure 6. Cont.
Figure 6. Cont.

Half-Cauchy Prior
Figure 6. Cont.

Uniform Prior
Figure 6. Cont.

Pareto Prior

![Graphs showing Pareto Prior with different plots for K values.](image-url)
Appendix I: Syntax for simulation runs

```r
#summary.1 - cauchy
#summary.2 - uniform
#summary.3 - pareto
#summary.4 - fixed
#summary.5 - random

set.seed(2091786)

summary.5 <- NULL

nobs <- c(3,4,5,6,9,16,25)
tau <- c(.001,.2,.5,.8)  #Note tau-squared simulated here since RE package requires the variance to run model
sigma <- c(.05,.1,.2)

for (i in 1:length(nobs))
{  for (j in 1:length(tau))
    { for (k in 1:length(sigma))
    {
      results <- NULL
      for (l in 1:1000)
      {
        es <- rnorm(nobs[i],0,tau[j]) + rnorm(nobs[i],0,sigma[k])
        se <- rep(sigma[k], nobs[i])
        #   fit <- rma(yi=es, vi=se, method="FE") # Fixed-effect
        fit <- rma(yi=es, vi=se, method="REML") # random effects
        #   fit <- hblm(es ~ 1, se)      #  Pareto prior
        #   fit <- hblm(es ~ 1, se, prior=list(tau=
        #     function(tau) 100/(100+tau)^2) )  # Uniform
        #   fit <- hblm(es ~ 1, se, prior=list(tau=
        #     function(tau) (1/(pi*(1+tau^2))) * (tau>0) )) # Cauchy
        
        #results <- rbind(results, c(fit$b,nobs[i],tau[j],sigma[k])) #For Fixed
        results <- rbind(results, c(fit$tau2,fit$b,nobs[i],tau[j],sigma[k])) #For Random
        #results <- rbind(results, c(fit$tau, fit$coef.s.p[1],nobs[i],tau[j],sigma[k])) #For Bayesian
      }
      write.table(results,file="random_sims", sep=" ", col.names=TRUE)
    }
    mean.tau <- mean(results[,1])
    var.tau <- var(results[,1])
    q.tau <- quantile(results[,1],c(.025,.5,.975))

    mean.mu <- mean(results[,2])
    var.mu <- var(results[,2])
    q.mu <- quantile(results[,2],c(.025,.5,.975))

    summary.5 <- rbind(summary.5, c(nobs[i],tau[j],sigma[k],mean.tau, var.tau, q.tau,
```
mean.mu, var.mu, q.mu))

#summary.4 <- rbind(summary.4, c(nobs[i], tau[j], sigma[k], mean.mu, var.mu, q.mu))

colnames(summary.5) <- c("nobs", "tau", "sigma", "mean.tau", "var.tau",
"q025.tau", "q500.tau", "q975.tau", "mean.mu", "var.mu", "q025.mu", "q500.mu", "q975.mu")

# colnames(summary.4) <- c("nobs", "tau", "sigma", "mean.mu", "var.mu", "q025.mu", "q500.mu", "q975.mu")

})

summary.5
write.table(summary.5, file="Random", sep=" ", col.names=TRUE)

plot(sqrt(mean.tau) ~ nobs, log="x", data=summary.5, # Note the square-root of mean.tau is calculated
col=as.factor(tau), xaxt="n",
pch=c(16,17,18)[as.factor(sigma)],
main="Random Effects",
ylim=c(0,1.2))
axis(1, labels=nobs,
at=c(3,4,5,6,9,16,25))
abline(h=.03,lty=3)
abline(h=.45,lty=3,col="red")
abline(h=.71,lty=3,col="green")
abline(h=.89,lty=3,col="blue")
legend(10,.7,title="tau",
legend=c(".89",".71",".45",".03"),
col=c("blue","green","red","black"),
pch=c(15))
legend(17,.7,title="sigma",
legend=c(".22",".32",".45"),
col=c("black"),
pch=c(16,17,18))

plot(var.tau ~ nobs, log="x", data=summary.5,
col=as.factor(tau), xaxt="n",
pch=c(16,17,18)[as.factor(sigma)],
main="Random Effects")
axis(1, labels=nobs,
at=c(3,4,5,6,9,16,25))
abline(h=0,lty=3)
legend(10,.4, title="tau",
legend=c(".89", ".71", ".45", ".03"),
col=c("blue", "green", "red", "black"),
pch=c(15))
legend(17,.4, title="sigma",
legend=c(".22", ",.32", ",.45"),
col=c("black"),
pch=c(16,17,18))
plot(sqrt(q975.tau) ~ nobs, log="x", data=summary.5, #Note square-root of all quantiles of tau
        col=as.factor(tau), xaxt="n",
        pch=c(16,17,18)[as.factor(sigma)],
        main="Random Effects",
        ylim=c(0,2.0))
axis(1,labels=nobs,
at=c(3,4,5,6,9,16,25))
abline(h=.03,lty=3)
abline(h=.45,lty=3,col="red")
abline(h=.71,lty=3,col="green")
abline(h=.89,lty=3,col="blue")
legend(10,2.0,title="tau",
       legend=c(".89",".71",".45",".03"),
       col=c("blue","green","red","black"),
       pch=c(15))
legend(17,2.0,title="sigma",
       legend=c(".22",".32",".45"),
       col=c("black"),
       pch=c(16,17,18))

plot(sqrt(q500.tau) ~ nobs, log="x", data=summary.5,
        col=as.factor(tau), xaxt="n",
        pch=c(16,17,18)[as.factor(sigma)],
        main="Random Effects",
        ylim=c(0,.9))
axis(1,labels=nobs,
at=c(3,4,5,6,9,16,25))
abline(h=.03,lty=3)
abline(h=.45,lty=3,col="red")
abline(h=.71,lty=3,col="green")
abline(h=.89,lty=3,col="blue")
legend(10,8,title="tau",
       legend=c(".89",".71",".45",".03"),
       col=c("blue","green","red","black"),
       pch=c(15))
legend(17,8,title="sigma",
       legend=c(".22",".32",".45"),
       col=c("black"),
       pch=c(16,17,18))

plot(sqrt(q025.tau) ~ nobs, log="x", data=summary.5,
        col=as.factor(tau), xaxt="n",
        pch=c(16,17,18)[as.factor(sigma)],
        main="Random Effects",
        ylim=c(0,.9))
axis(1,labels=nobs,
at=c(3,4,5,6,9,16,25))
abline(h=.03,lty=3)
abline(h=.45,lty=3,col="red")
abline(h=.71,lty=3,col="green")
abline(h=.89,lty=3,col="blue")
legend(4,.7,title="tau",
    legend=c(".89",".71",".45",".03"),
    col=c("blue","green","red","black"),
    pch=c(15))
legend(6,.7,title="sigma",
    legend=c(".22",".32",".45"),
    col=c("black"),
    pch=c(16,17,18))

plot(mean.mu ~ nobs, log="x", data=summary.5,
     col=as.factor(tau), xaxt="n",
     pch=c(16,17,18)[as.factor(sigma)],
     main="Random Effects",
     ylim=c(-.03,.02))
axis(1,labels=nobs,
     at=c(3,4,5,6,9,16,25))
abline(h=0,lty=3)
legend(10,-.01,title="tau",
    legend=c(".89",".71",".45",".03"),
    col=c("blue","green","red","black"),
    pch=c(15))
legend(17,-.01,title="sigma",
    legend=c(".22",".32",".45"),
    col=c("black"),
    pch=c(16,17,18))

plot(var.mu ~ nobs, log="x", data=summary.5,
     col=as.factor(tau), xaxt="n",
     pch=c(16,17,18)[as.factor(sigma)],
     main="Random Effects")
axis(1,labels=nobs,
     at=c(3,4,5,6,9,16,25))
abline(h=0,lty=3)
legend(10,.3,title="tau",
    legend=c(".89",".71",".45",".03"),
    col=c("blue","green","red","black"),
    pch=c(15))
legend(17,.3,title="sigma",
    legend=c(".22",".32",".45"),
    col=c("black"),
    pch=c(16,17,18))

plot(q975.mu ~ nobs, log="x", data=summary.5,
Appendix II: Raw results of 84 simulation runs – Fixed-effect and Random-effects.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fixed-Effect</th>
<th>Random-Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>K</td>
<td>(\tau)</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.8</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>0.8</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.8</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>0.2</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>0.2</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>0.2</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>0.2</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>0.5</td>
</tr>
</tbody>
</table>
## Appendix II: Cont.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fixed-Effect</th>
<th>Random-Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mu.est</td>
<td>tau.est</td>
</tr>
<tr>
<td>21</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>22</td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>23</td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>24</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>25</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>26</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>27</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>28</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>29</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>30</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>31</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>32</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>33</td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>34</td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>35</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>36</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>37</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>38</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>39</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>40</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>41</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>42</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>43</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>44</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>
### Appendix II: Cont.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fixed-Effect</th>
<th>Random-Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>K</td>
<td>$\tau^2$</td>
</tr>
<tr>
<td>45</td>
<td>6</td>
<td>0.5</td>
</tr>
<tr>
<td>46</td>
<td>6</td>
<td>0.8</td>
</tr>
<tr>
<td>47</td>
<td>6</td>
<td>0.8</td>
</tr>
<tr>
<td>48</td>
<td>6</td>
<td>0.8</td>
</tr>
<tr>
<td>49</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>51</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>52</td>
<td>9</td>
<td>0.2</td>
</tr>
<tr>
<td>53</td>
<td>9</td>
<td>0.2</td>
</tr>
<tr>
<td>54</td>
<td>9</td>
<td>0.2</td>
</tr>
<tr>
<td>55</td>
<td>9</td>
<td>0.5</td>
</tr>
<tr>
<td>56</td>
<td>9</td>
<td>0.5</td>
</tr>
<tr>
<td>57</td>
<td>9</td>
<td>0.5</td>
</tr>
<tr>
<td>58</td>
<td>9</td>
<td>0.8</td>
</tr>
<tr>
<td>59</td>
<td>9</td>
<td>0.8</td>
</tr>
<tr>
<td>60</td>
<td>9</td>
<td>0.8</td>
</tr>
<tr>
<td>61</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>62</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>63</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>16</td>
<td>0.2</td>
</tr>
</tbody>
</table>
### Appendix II: Cont.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fixed-Effect</th>
<th>Random-Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td># K t^2 z^2 s^2</td>
<td>mu.est tau.est sd.mu var.mu sd.tau var.tau bias.mu bias.tau mse.mu mse.tau</td>
<td>mu.est tau.est sd.mu var.mu sd.tau var.tau bias.mu bias.tau mse.mu mse.tau</td>
</tr>
<tr>
<td>65 16 0.2 0.45 0.32</td>
<td>0.0131 0.0113 0.0001</td>
<td>0.0131 -0.5623 0.0003 0.1362</td>
</tr>
<tr>
<td>66 16 0.2 0.2 0.45 0.45</td>
<td>-0.1168 0.0827 0.0068</td>
<td>-0.1168 -0.6687 0.0205 0.4472</td>
</tr>
<tr>
<td>67 16 0.5 0.71 0.22</td>
<td>-0.0494 0.0125 0.0002</td>
<td>-0.0494 -0.4729 0.0026 0.2236</td>
</tr>
<tr>
<td>68 16 0.5 0.71 0.32</td>
<td>0.0360 0.1431 0.0205</td>
<td>0.0360 -0.5623 0.0218 0.3162</td>
</tr>
<tr>
<td>69 16 0.5 0.2 0.71</td>
<td>0.1553 0.0368 0.0014</td>
<td>0.1553 -0.6687 0.0255 0.4472</td>
</tr>
<tr>
<td>70 16 0.8 0.1 0.89</td>
<td>-0.1170 0.1990 0.0396</td>
<td>-0.1170 -0.4729 0.0533 0.2236</td>
</tr>
<tr>
<td>71 16 0.8 0.1 0.89</td>
<td>-0.1017 0.1400 0.0196</td>
<td>-0.1017 -0.5623 0.0300 0.3162</td>
</tr>
<tr>
<td>72 16 0.8 0.2 0.89</td>
<td>-0.1596 0.0250 0.006</td>
<td>-0.1596 -0.6687 0.0261 0.4472</td>
</tr>
<tr>
<td>73 25 0 0.1 0.03</td>
<td>-0.0090 0.0084 0.0001</td>
<td>-0.0090 -0.4729 0.0002 0.2236</td>
</tr>
<tr>
<td>74 25 0 0.1 0.03</td>
<td>-0.0157 0.0212 0.0004</td>
<td>-0.0157 -0.5623 0.0007 0.3162</td>
</tr>
<tr>
<td>75 25 0 0.2 0.03</td>
<td>0.0036 0.0259 0.0007</td>
<td>0.0036 -0.6687 0.0007 0.4472</td>
</tr>
<tr>
<td>76 25 0 0.2 0.45</td>
<td>-0.0160 0.0146 0.0002</td>
<td>-0.0160 -0.4729 0.0005 0.2236</td>
</tr>
<tr>
<td>77 25 0 0.2 0.45</td>
<td>0.0030 0.1035 0.0107</td>
<td>0.0030 -0.5623 0.0107 0.3162</td>
</tr>
<tr>
<td>78 25 0 0.2 0.45</td>
<td>-0.0311 0.0497 0.0025</td>
<td>-0.0311 -0.6687 0.0034 0.4472</td>
</tr>
<tr>
<td>79 25 0.5 0.1 0.71</td>
<td>0.0303 0.1644 0.0270</td>
<td>0.0303 -0.4729 0.0279 0.2236</td>
</tr>
<tr>
<td>80 25 0.5 0.1 0.71</td>
<td>0.0148 0.0480 0.0023</td>
<td>0.0148 -0.5623 0.0025 0.3162</td>
</tr>
<tr>
<td>81 25 0.5 0.2 0.71</td>
<td>-0.0516 0.0004 0.0000</td>
<td>-0.0516 -0.6687 0.0027 0.4472</td>
</tr>
<tr>
<td>82 25 0.8 0.1 0.89</td>
<td>-0.2579 0.0451 0.0020</td>
<td>-0.2579 -0.4729 0.0686 0.2236</td>
</tr>
<tr>
<td>83 25 0.8 0.1 0.89</td>
<td>-0.0057 0.1394 0.0194</td>
<td>-0.0057 -0.5623 0.0195 0.3162</td>
</tr>
<tr>
<td>84 25 0.8 0.2 0.89</td>
<td>0.0758 0.1074 0.0115</td>
<td>-0.0758 -0.3016 0.0173 0.0010</td>
</tr>
</tbody>
</table>
### Appendix III: Raw results of 84 simulation runs – FB Pareto and FB Uniform.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fully Bayesian - Pareto</th>
<th>Fully Bayesian - Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td># K t_\mu t_\tau \sigma</td>
<td>mu.est tau.est sd.mu</td>
<td>var.mu sd.tau</td>
</tr>
<tr>
<td>1 3 0.00 0.05 0.03 0.22</td>
<td>-0.0609 0.1005 0.2283</td>
<td>0.0521 0.0060 0.0000</td>
</tr>
<tr>
<td>2 3 0.00 0.10 0.03 0.32</td>
<td>0.0986 0.1676 0.0528</td>
<td>0.0028 0.0777 0.0060</td>
</tr>
<tr>
<td>3 3 0.00 0.20 0.03 0.45</td>
<td>0.1591 0.2183 0.0174</td>
<td>0.0003 0.0109 0.0001</td>
</tr>
<tr>
<td>4 3 0.20 0.05 0.45 0.22</td>
<td>0.1785 0.2118 0.2790</td>
<td>0.0778 0.0184 0.0003</td>
</tr>
<tr>
<td>5 3 0.20 0.10 0.45 0.32</td>
<td>0.2521 0.3871 0.0651</td>
<td>0.0042 0.3420 0.1169</td>
</tr>
<tr>
<td>6 3 0.20 0.20 0.45 0.45</td>
<td>0.0241 0.1927 0.1091</td>
<td>0.0119 0.2868 0.0008</td>
</tr>
<tr>
<td>7 3 0.50 0.05 0.71 0.22</td>
<td>0.0851 0.3662 0.4536</td>
<td>0.2058 0.0048 0.0000</td>
</tr>
<tr>
<td>8 3 0.50 0.10 0.71 0.32</td>
<td>-0.3286 0.8468 0.2105</td>
<td>0.0443 0.0222 0.0005</td>
</tr>
<tr>
<td>9 3 0.50 0.20 0.71 0.45</td>
<td>-0.1402 0.8424 0.6199</td>
<td>0.3843 0.8640 0.7465</td>
</tr>
<tr>
<td>10 4 0.00 0.05 0.03 0.22</td>
<td>-0.4892 0.9497 0.4858</td>
<td>0.2360 0.4608 0.2123</td>
</tr>
<tr>
<td>11 4 0.00 0.10 0.03 0.32</td>
<td>-0.1742 0.7539 0.5298</td>
<td>0.2807 0.8614 0.7420</td>
</tr>
<tr>
<td>12 4 0.00 0.20 0.03 0.45</td>
<td>-0.6744 0.5669 0.4709</td>
<td>0.2217 0.4832 0.2334</td>
</tr>
<tr>
<td>13 4 0.00 0.05 0.03 0.22</td>
<td>-0.2364 0.2386 0.1224</td>
<td>0.0150 0.2152 0.0463</td>
</tr>
<tr>
<td>14 4 0.00 0.10 0.03 0.32</td>
<td>0.1647 0.1215 0.1277</td>
<td>0.0163 0.0536 0.0004</td>
</tr>
<tr>
<td>15 4 0.00 0.20 0.03 0.45</td>
<td>-0.0135 0.1621 0.1742</td>
<td>0.0303 0.0143 0.0002</td>
</tr>
<tr>
<td>16 4 0.20 0.05 0.03 0.22</td>
<td>-0.4938 0.1895 0.2661</td>
<td>0.0708 0.0264 0.0007</td>
</tr>
<tr>
<td>17 4 0.20 0.10 0.03 0.32</td>
<td>-0.6600 0.1715 0.1064</td>
<td>0.0113 0.0395 0.0016</td>
</tr>
<tr>
<td>18 4 0.20 0.20 0.03 0.45</td>
<td>0.0060 0.3539 0.0065</td>
<td>0.0000 0.2502 0.0626</td>
</tr>
<tr>
<td>19 4 0.50 0.05 0.03 0.22</td>
<td>-0.4940 0.5749 0.0997</td>
<td>0.0090 0.5282 0.2790</td>
</tr>
<tr>
<td>20 4 0.50 0.10 0.03 0.32</td>
<td>-0.1126 0.5026 0.0029</td>
<td>0.0000 0.2051 0.0421</td>
</tr>
</tbody>
</table>
### Appendix III: Cont.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fully Bayesian - Pareto</th>
<th>Fully Bayesian - Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>K</td>
<td>t^2</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>0.50</td>
</tr>
<tr>
<td>22</td>
<td>4</td>
<td>0.80</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>0.80</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>0.80</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>27</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>5</td>
<td>0.20</td>
</tr>
<tr>
<td>29</td>
<td>5</td>
<td>0.20</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>0.20</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
<td>0.50</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td>0.50</td>
</tr>
<tr>
<td>33</td>
<td>5</td>
<td>0.50</td>
</tr>
<tr>
<td>34</td>
<td>5</td>
<td>0.80</td>
</tr>
<tr>
<td>35</td>
<td>5</td>
<td>0.80</td>
</tr>
<tr>
<td>36</td>
<td>5</td>
<td>0.80</td>
</tr>
<tr>
<td>37</td>
<td>6</td>
<td>0.00</td>
</tr>
<tr>
<td>38</td>
<td>6</td>
<td>0.00</td>
</tr>
<tr>
<td>39</td>
<td>6</td>
<td>0.00</td>
</tr>
<tr>
<td>40</td>
<td>6</td>
<td>0.20</td>
</tr>
<tr>
<td>41</td>
<td>6</td>
<td>0.20</td>
</tr>
<tr>
<td>42</td>
<td>6</td>
<td>0.20</td>
</tr>
<tr>
<td>43</td>
<td>6</td>
<td>0.50</td>
</tr>
<tr>
<td>44</td>
<td>6</td>
<td>0.50</td>
</tr>
</tbody>
</table>

**Simulation Runs**

- | 78 |

**Appendix III: Cont.**

- | 78 |
### Appendix III: Cont.

<table>
<thead>
<tr>
<th>#</th>
<th>K</th>
<th>$t_0^*$</th>
<th>$t_1^*$</th>
<th>$t_2^*$</th>
<th>$t_3^*$</th>
<th>mu.est</th>
<th>tau.est</th>
<th>sd.mu</th>
<th>var.mu</th>
<th>sd.tau</th>
<th>var.tau</th>
<th>bias.mu</th>
<th>bias.tau</th>
<th>mse.mu</th>
<th>mse.tau</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>6</td>
<td>0.50</td>
<td>0.20</td>
<td>0.71</td>
<td>0.45</td>
<td>-0.0693</td>
<td>0.4537</td>
<td>0.2552</td>
<td>0.0651</td>
<td>0.0425</td>
<td>0.0018</td>
<td>-0.0693</td>
<td>-0.2534</td>
<td>0.0699</td>
<td>0.0660</td>
</tr>
<tr>
<td>46</td>
<td>6</td>
<td>0.80</td>
<td>0.05</td>
<td>0.89</td>
<td>0.22</td>
<td>-1.254</td>
<td>0.8589</td>
<td>0.6290</td>
<td>0.3957</td>
<td>0.1746</td>
<td>0.0305</td>
<td>-1.254</td>
<td>-0.0356</td>
<td>0.4114</td>
<td>0.0317</td>
</tr>
<tr>
<td>47</td>
<td>6</td>
<td>0.80</td>
<td>0.10</td>
<td>0.89</td>
<td>0.32</td>
<td>0.5253</td>
<td>1.0746</td>
<td>0.1811</td>
<td>0.0328</td>
<td>0.2126</td>
<td>0.0452</td>
<td>0.5253</td>
<td>0.1802</td>
<td>0.3088</td>
<td>0.0776</td>
</tr>
<tr>
<td>48</td>
<td>6</td>
<td>0.80</td>
<td>0.20</td>
<td>0.89</td>
<td>0.45</td>
<td>-0.3206</td>
<td>0.7558</td>
<td>0.0724</td>
<td>0.0052</td>
<td>0.1404</td>
<td>0.0197</td>
<td>-0.3206</td>
<td>-0.1387</td>
<td>0.1080</td>
<td>0.0389</td>
</tr>
<tr>
<td>49</td>
<td>9</td>
<td>0.00</td>
<td>0.05</td>
<td>0.03</td>
<td>0.22</td>
<td>-0.4206</td>
<td>0.1226</td>
<td>0.4296</td>
<td>0.1846</td>
<td>0.0289</td>
<td>0.0008</td>
<td>-0.4206</td>
<td>0.0910</td>
<td>0.3614</td>
<td>0.0091</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>0.00</td>
<td>0.10</td>
<td>0.03</td>
<td>0.32</td>
<td>0.0572</td>
<td>0.1116</td>
<td>0.0192</td>
<td>0.0004</td>
<td>0.0527</td>
<td>0.0028</td>
<td>0.0572</td>
<td>0.0799</td>
<td>0.0036</td>
<td>0.0092</td>
</tr>
<tr>
<td>51</td>
<td>9</td>
<td>0.00</td>
<td>0.20</td>
<td>0.03</td>
<td>0.45</td>
<td>-0.0348</td>
<td>0.1443</td>
<td>0.0785</td>
<td>0.0062</td>
<td>0.0505</td>
<td>0.0026</td>
<td>-0.0348</td>
<td>0.1127</td>
<td>0.0074</td>
<td>0.0153</td>
</tr>
<tr>
<td>52</td>
<td>9</td>
<td>0.20</td>
<td>0.05</td>
<td>0.03</td>
<td>0.22</td>
<td>-0.2741</td>
<td>0.2381</td>
<td>0.0868</td>
<td>0.0075</td>
<td>0.1354</td>
<td>0.0183</td>
<td>-0.2741</td>
<td>-0.2091</td>
<td>0.0826</td>
<td>0.0621</td>
</tr>
<tr>
<td>53</td>
<td>9</td>
<td>0.20</td>
<td>0.10</td>
<td>0.03</td>
<td>0.32</td>
<td>-0.1204</td>
<td>0.2532</td>
<td>0.0446</td>
<td>0.0020</td>
<td>0.1171</td>
<td>0.0137</td>
<td>-0.1204</td>
<td>-0.1940</td>
<td>0.0165</td>
<td>0.0514</td>
</tr>
<tr>
<td>54</td>
<td>9</td>
<td>0.20</td>
<td>0.20</td>
<td>0.03</td>
<td>0.45</td>
<td>0.0803</td>
<td>0.4283</td>
<td>0.4255</td>
<td>0.1811</td>
<td>0.0132</td>
<td>0.0002</td>
<td>0.0803</td>
<td>-0.0189</td>
<td>0.1875</td>
<td>0.0065</td>
</tr>
<tr>
<td>55</td>
<td>9</td>
<td>0.50</td>
<td>0.05</td>
<td>0.03</td>
<td>0.22</td>
<td>0.0236</td>
<td>0.2816</td>
<td>0.2053</td>
<td>0.0422</td>
<td>0.1613</td>
<td>0.0260</td>
<td>0.0236</td>
<td>-0.4255</td>
<td>0.0427</td>
<td>0.2071</td>
</tr>
<tr>
<td>56</td>
<td>9</td>
<td>0.50</td>
<td>0.10</td>
<td>0.03</td>
<td>0.32</td>
<td>-0.0990</td>
<td>0.7155</td>
<td>0.0321</td>
<td>0.0010</td>
<td>0.4325</td>
<td>0.1871</td>
<td>-0.0990</td>
<td>0.0084</td>
<td>0.0108</td>
<td>0.1871</td>
</tr>
<tr>
<td>57</td>
<td>9</td>
<td>0.50</td>
<td>0.20</td>
<td>0.03</td>
<td>0.45</td>
<td>-0.1645</td>
<td>0.4648</td>
<td>0.1552</td>
<td>0.0241</td>
<td>0.0807</td>
<td>0.0065</td>
<td>-0.1645</td>
<td>-0.2423</td>
<td>0.0511</td>
<td>0.0652</td>
</tr>
<tr>
<td>58</td>
<td>9</td>
<td>0.80</td>
<td>0.05</td>
<td>0.89</td>
<td>0.22</td>
<td>0.2194</td>
<td>0.6285</td>
<td>0.0122</td>
<td>0.0001</td>
<td>0.3452</td>
<td>0.1871</td>
<td>-0.1645</td>
<td>-0.2423</td>
<td>0.0511</td>
<td>0.0652</td>
</tr>
<tr>
<td>59</td>
<td>9</td>
<td>0.80</td>
<td>0.10</td>
<td>0.89</td>
<td>0.32</td>
<td>-0.2561</td>
<td>0.9813</td>
<td>0.1456</td>
<td>0.0212</td>
<td>0.2667</td>
<td>0.0711</td>
<td>-0.2561</td>
<td>0.0869</td>
<td>0.0868</td>
<td>0.0787</td>
</tr>
<tr>
<td>60</td>
<td>9</td>
<td>0.80</td>
<td>0.20</td>
<td>0.89</td>
<td>0.45</td>
<td>-0.0616</td>
<td>0.7888</td>
<td>0.1120</td>
<td>0.0125</td>
<td>0.2179</td>
<td>0.0475</td>
<td>-0.0616</td>
<td>-0.1056</td>
<td>0.0163</td>
<td>0.0586</td>
</tr>
<tr>
<td>61</td>
<td>16</td>
<td>0.00</td>
<td>0.05</td>
<td>0.03</td>
<td>0.22</td>
<td>-0.3198</td>
<td>0.5983</td>
<td>0.5296</td>
<td>0.2805</td>
<td>0.5780</td>
<td>0.5746</td>
<td>-0.3198</td>
<td>0.5667</td>
<td>0.3828</td>
<td>0.8957</td>
</tr>
<tr>
<td>62</td>
<td>16</td>
<td>0.00</td>
<td>0.10</td>
<td>0.03</td>
<td>0.32</td>
<td>-0.1398</td>
<td>0.0848</td>
<td>0.0047</td>
<td>0.0025</td>
<td>0.0255</td>
<td>0.0005</td>
<td>-0.1398</td>
<td>0.0531</td>
<td>0.0221</td>
<td>0.0035</td>
</tr>
<tr>
<td>63</td>
<td>16</td>
<td>0.00</td>
<td>0.20</td>
<td>0.03</td>
<td>0.45</td>
<td>-0.0739</td>
<td>0.1400</td>
<td>0.2289</td>
<td>0.0524</td>
<td>0.0708</td>
<td>0.0050</td>
<td>-0.0739</td>
<td>0.1086</td>
<td>0.0579</td>
<td>0.0168</td>
</tr>
<tr>
<td>64</td>
<td>16</td>
<td>0.20</td>
<td>0.05</td>
<td>0.45</td>
<td>0.22</td>
<td>-0.0557</td>
<td>0.4020</td>
<td>0.1183</td>
<td>0.0140</td>
<td>0.2993</td>
<td>0.0896</td>
<td>-0.0557</td>
<td>-0.0452</td>
<td>0.0171</td>
<td>0.0916</td>
</tr>
</tbody>
</table>

### Simulation Runs
- Fully Bayesian - Pareto
- Fully Bayesian - Uniform
### Appendix III: Cont.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fully Bayesian - Pareto</th>
<th>Fully Bayesian - Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mu.est</td>
<td>tau.est</td>
</tr>
<tr>
<td>65 16 0.20 0.10 0.35 0.32</td>
<td>0.0145</td>
<td>0.3942</td>
</tr>
<tr>
<td>66 16 0.20 0.05 0.45 0.45</td>
<td>0.0740</td>
<td>0.2390</td>
</tr>
<tr>
<td>67 16 0.50 0.05 0.71 0.22</td>
<td>-0.0201</td>
<td>0.1624</td>
</tr>
<tr>
<td>68 16 0.50 0.10 0.71 0.32</td>
<td>0.0266</td>
<td>0.7210</td>
</tr>
<tr>
<td>69 16 0.50 0.20 0.71 0.45</td>
<td>0.0837</td>
<td>0.5646</td>
</tr>
<tr>
<td>70 16 0.50 0.05 0.89 0.22</td>
<td>0.2016</td>
<td>0.8057</td>
</tr>
<tr>
<td>71 16 0.50 0.10 0.89 0.32</td>
<td>0.1413</td>
<td>0.9769</td>
</tr>
<tr>
<td>72 16 0.80 0.05 0.89 0.32</td>
<td>-0.2638</td>
<td>0.8057</td>
</tr>
<tr>
<td>73 25 0.00 0.05 0.03 0.22</td>
<td>-0.1276</td>
<td>0.3705</td>
</tr>
<tr>
<td>74 25 0.00 0.10 0.03 0.32</td>
<td>-0.0142</td>
<td>0.0857</td>
</tr>
<tr>
<td>75 25 0.00 0.20 0.03 0.45</td>
<td>-0.0570</td>
<td>0.1000</td>
</tr>
<tr>
<td>76 25 0.20 0.05 0.45 0.22</td>
<td>-0.0022</td>
<td>0.2702</td>
</tr>
<tr>
<td>77 25 0.20 0.10 0.45 0.32</td>
<td>-0.0908</td>
<td>0.5896</td>
</tr>
<tr>
<td>78 25 0.20 0.20 0.45 0.32</td>
<td>-0.0295</td>
<td>0.4077</td>
</tr>
<tr>
<td>79 25 0.50 0.05 0.71 0.22</td>
<td>-0.1534</td>
<td>0.4862</td>
</tr>
<tr>
<td>80 25 0.50 0.10 0.71 0.32</td>
<td>0.0692</td>
<td>0.7365</td>
</tr>
<tr>
<td>81 25 0.50 0.20 0.71 0.45</td>
<td>-0.0440</td>
<td>0.6996</td>
</tr>
<tr>
<td>82 25 0.80 0.05 0.89 0.22</td>
<td>-0.1328</td>
<td>0.7321</td>
</tr>
<tr>
<td>83 25 0.80 0.10 0.89 0.32</td>
<td>-0.0524</td>
<td>0.8824</td>
</tr>
<tr>
<td>84 25 0.80 0.20 0.89 0.45</td>
<td>-0.0044</td>
<td>0.9467</td>
</tr>
</tbody>
</table>
Appendix IV: Raw results of 84 simulation runs – FB Half-Cauchy.

<table>
<thead>
<tr>
<th>#</th>
<th>K</th>
<th>$\tau_1^2$</th>
<th>$\sigma_1^2$</th>
<th>$\tau_2^2$</th>
<th>$\sigma_2^2$</th>
<th>mu.est</th>
<th>tau.est</th>
<th>sd.mu</th>
<th>var.mu</th>
<th>sd.tau</th>
<th>var.tau</th>
<th>bias.mu</th>
<th>bias.tau</th>
<th>mse.mu</th>
<th>mse.tau</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0.00</td>
<td>0.05</td>
<td>0.03</td>
<td>0.22</td>
<td>-0.0609</td>
<td>0.1828</td>
<td>0.1614</td>
<td>0.0261</td>
<td>0.0094</td>
<td>0.0001</td>
<td>-0.0609</td>
<td>0.1512</td>
<td>0.0298</td>
<td>0.0229</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.00</td>
<td>0.10</td>
<td>0.03</td>
<td>0.32</td>
<td>0.0986</td>
<td>0.2926</td>
<td>0.0373</td>
<td>0.0014</td>
<td>0.0835</td>
<td>0.0070</td>
<td>0.0986</td>
<td>0.2610</td>
<td>0.0111</td>
<td>0.0751</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.00</td>
<td>0.20</td>
<td>0.03</td>
<td>0.45</td>
<td>0.1591</td>
<td>0.3520</td>
<td>0.0123</td>
<td>0.0002</td>
<td>0.0297</td>
<td>0.0009</td>
<td>0.1591</td>
<td>0.3204</td>
<td>0.0255</td>
<td>0.1035</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.20</td>
<td>0.05</td>
<td>0.45</td>
<td>0.22</td>
<td>0.1785</td>
<td>0.3447</td>
<td>0.1973</td>
<td>0.0389</td>
<td>0.0417</td>
<td>0.0017</td>
<td>0.1785</td>
<td>-0.1025</td>
<td>0.0708</td>
<td>0.0122</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.20</td>
<td>0.10</td>
<td>0.45</td>
<td>0.32</td>
<td>0.2521</td>
<td>0.5289</td>
<td>0.0460</td>
<td>0.0021</td>
<td>0.2563</td>
<td>0.0657</td>
<td>0.2521</td>
<td>0.0817</td>
<td>0.0657</td>
<td>0.0724</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.20</td>
<td>0.20</td>
<td>0.45</td>
<td>0.45</td>
<td>0.0241</td>
<td>0.3098</td>
<td>0.0772</td>
<td>0.0060</td>
<td>0.0164</td>
<td>0.0003</td>
<td>0.0241</td>
<td>-0.1374</td>
<td>0.0065</td>
<td>0.0191</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.50</td>
<td>0.05</td>
<td>0.71</td>
<td>0.22</td>
<td>0.0851</td>
<td>0.5358</td>
<td>0.3208</td>
<td>0.1029</td>
<td>0.0015</td>
<td>0.0000</td>
<td>0.0851</td>
<td>-0.1713</td>
<td>0.1101</td>
<td>0.0293</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.50</td>
<td>0.10</td>
<td>0.71</td>
<td>0.32</td>
<td>-0.3286</td>
<td>0.9767</td>
<td>0.1488</td>
<td>0.0222</td>
<td>0.0162</td>
<td>0.0003</td>
<td>-0.3286</td>
<td>0.2696</td>
<td>0.1301</td>
<td>0.0729</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.50</td>
<td>0.20</td>
<td>0.71</td>
<td>0.45</td>
<td>-0.1402</td>
<td>0.9445</td>
<td>0.4383</td>
<td>0.1921</td>
<td>0.5887</td>
<td>0.3466</td>
<td>-0.1402</td>
<td>0.2374</td>
<td>0.2118</td>
<td>0.4030</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.80</td>
<td>0.05</td>
<td>0.89</td>
<td>0.22</td>
<td>-0.4892</td>
<td>1.0584</td>
<td>0.3435</td>
<td>0.1180</td>
<td>0.2917</td>
<td>0.0851</td>
<td>-0.4892</td>
<td>0.1640</td>
<td>0.3573</td>
<td>0.1119</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>0.80</td>
<td>0.10</td>
<td>0.89</td>
<td>0.32</td>
<td>-0.1742</td>
<td>0.8600</td>
<td>0.3746</td>
<td>0.1403</td>
<td>0.5884</td>
<td>0.3463</td>
<td>-0.1742</td>
<td>-0.0344</td>
<td>0.1707</td>
<td>0.3474</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.80</td>
<td>0.20</td>
<td>0.89</td>
<td>0.45</td>
<td>-0.6744</td>
<td>0.6878</td>
<td>0.3330</td>
<td>0.1109</td>
<td>0.3418</td>
<td>0.1168</td>
<td>-0.6744</td>
<td>-0.2067</td>
<td>0.5656</td>
<td>0.1596</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>0.00</td>
<td>0.05</td>
<td>0.03</td>
<td>0.22</td>
<td>-0.2246</td>
<td>0.3556</td>
<td>0.0865</td>
<td>0.0075</td>
<td>0.2139</td>
<td>0.0457</td>
<td>-0.2246</td>
<td>0.3240</td>
<td>0.0579</td>
<td>0.1507</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>0.00</td>
<td>0.10</td>
<td>0.03</td>
<td>0.32</td>
<td>0.1647</td>
<td>0.1973</td>
<td>0.0903</td>
<td>0.0082</td>
<td>0.0385</td>
<td>0.0015</td>
<td>0.1647</td>
<td>0.1657</td>
<td>0.0353</td>
<td>0.0289</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>0.00</td>
<td>0.20</td>
<td>0.03</td>
<td>0.45</td>
<td>-0.0135</td>
<td>0.2468</td>
<td>0.1232</td>
<td>0.0152</td>
<td>0.0041</td>
<td>0.0000</td>
<td>-0.0135</td>
<td>0.2151</td>
<td>0.0154</td>
<td>0.0463</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>0.20</td>
<td>0.05</td>
<td>0.45</td>
<td>0.22</td>
<td>-0.4938</td>
<td>0.2890</td>
<td>0.1882</td>
<td>0.0354</td>
<td>0.0403</td>
<td>0.0016</td>
<td>-0.4938</td>
<td>-0.1582</td>
<td>0.2793</td>
<td>0.0266</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>0.20</td>
<td>0.10</td>
<td>0.45</td>
<td>0.32</td>
<td>-0.0660</td>
<td>0.2736</td>
<td>0.0752</td>
<td>0.0057</td>
<td>0.0456</td>
<td>0.0021</td>
<td>-0.0660</td>
<td>-0.1736</td>
<td>0.0100</td>
<td>0.0322</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>0.20</td>
<td>0.20</td>
<td>0.45</td>
<td>0.45</td>
<td>0.0609</td>
<td>0.4792</td>
<td>0.0046</td>
<td>0.0000</td>
<td>0.1948</td>
<td>0.0379</td>
<td>0.0609</td>
<td>0.0320</td>
<td>0.0037</td>
<td>0.0390</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>0.50</td>
<td>0.05</td>
<td>0.71</td>
<td>0.22</td>
<td>-0.4949</td>
<td>0.6697</td>
<td>0.0705</td>
<td>0.0050</td>
<td>0.3720</td>
<td>0.1384</td>
<td>-0.4949</td>
<td>-0.0374</td>
<td>0.2499</td>
<td>0.1398</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>0.50</td>
<td>0.10</td>
<td>0.71</td>
<td>0.32</td>
<td>-0.1126</td>
<td>0.6252</td>
<td>0.0020</td>
<td>0.0000</td>
<td>0.1252</td>
<td>0.0157</td>
<td>-0.1126</td>
<td>-0.0819</td>
<td>0.0127</td>
<td>0.0224</td>
</tr>
</tbody>
</table>
## Appendix IV: Cont.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fully Bayesian - Half-Cauchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>K</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>29</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>33</td>
<td>5</td>
</tr>
<tr>
<td>34</td>
<td>5</td>
</tr>
<tr>
<td>35</td>
<td>5</td>
</tr>
<tr>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>37</td>
<td>6</td>
</tr>
<tr>
<td>38</td>
<td>6</td>
</tr>
<tr>
<td>39</td>
<td>6</td>
</tr>
<tr>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td>41</td>
<td>6</td>
</tr>
<tr>
<td>42</td>
<td>6</td>
</tr>
<tr>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td>44</td>
<td>6</td>
</tr>
</tbody>
</table>
### Appendix IV: Cont.

<table>
<thead>
<tr>
<th>Simulation Runs</th>
<th>Fully Bayesian - Half-Cauchy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mu.est</td>
</tr>
<tr>
<td>#</td>
<td>K</td>
</tr>
<tr>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>46</td>
<td>6</td>
</tr>
<tr>
<td>47</td>
<td>6</td>
</tr>
<tr>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td>49</td>
<td>9</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
</tr>
<tr>
<td>51</td>
<td>9</td>
</tr>
<tr>
<td>52</td>
<td>9</td>
</tr>
<tr>
<td>53</td>
<td>9</td>
</tr>
<tr>
<td>54</td>
<td>9</td>
</tr>
<tr>
<td>55</td>
<td>9</td>
</tr>
<tr>
<td>56</td>
<td>9</td>
</tr>
<tr>
<td>57</td>
<td>9</td>
</tr>
<tr>
<td>58</td>
<td>9</td>
</tr>
<tr>
<td>59</td>
<td>9</td>
</tr>
<tr>
<td>60</td>
<td>9</td>
</tr>
<tr>
<td>61</td>
<td>16</td>
</tr>
<tr>
<td>62</td>
<td>16</td>
</tr>
<tr>
<td>63</td>
<td>16</td>
</tr>
<tr>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td>Simulation Runs</td>
<td>Fully Bayesian - Half-Cauchy</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>#</td>
<td>K</td>
</tr>
<tr>
<td>65</td>
<td>16</td>
</tr>
<tr>
<td>66</td>
<td>16</td>
</tr>
<tr>
<td>67</td>
<td>16</td>
</tr>
<tr>
<td>68</td>
<td>16</td>
</tr>
<tr>
<td>69</td>
<td>16</td>
</tr>
<tr>
<td>70</td>
<td>16</td>
</tr>
<tr>
<td>71</td>
<td>16</td>
</tr>
<tr>
<td>72</td>
<td>16</td>
</tr>
<tr>
<td>73</td>
<td>25</td>
</tr>
<tr>
<td>74</td>
<td>25</td>
</tr>
<tr>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>76</td>
<td>25</td>
</tr>
<tr>
<td>77</td>
<td>25</td>
</tr>
<tr>
<td>78</td>
<td>25</td>
</tr>
<tr>
<td>79</td>
<td>25</td>
</tr>
<tr>
<td>80</td>
<td>25</td>
</tr>
<tr>
<td>81</td>
<td>25</td>
</tr>
<tr>
<td>82</td>
<td>25</td>
</tr>
<tr>
<td>83</td>
<td>25</td>
</tr>
<tr>
<td>84</td>
<td>25</td>
</tr>
</tbody>
</table>
References


Greenhouse & Iyengar (2009)

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine, 17*(8), 841-856.


Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of


