UC Merced

UC Merced Electronic Theses and Dissertations

Title

A parsimonious weight function for modeling publication bias

Permalink

<https://escholarship.org/uc/item/9hw610q5>

Author

Citkowicz, Martyna

Publication Date

2012-04-17

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, MERCED

A Parsimonious Weight Function for Modeling Publication Bias

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Psychological Sciences

by

Martyna Citkowicz

Committee in charge:

Professor Jack L. Vevea, Chair
Professor William R. Shadish
Professor Anna Song

April 2012

Copyright © 2012 by Martyna Citkowicz

All Rights Reserved

The Dissertation of Martyna Citkowicz is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, Merced

April 2012

To my parents
Without whose love and support I would not be where I am today

Table of Contents

List of Abbreviations and Symbols	viii
List of Figures.....	xi
List of Tables	xiv
Acknowledgments	xvi
Abstract	xvii
Chapter 1. Introduction	1
Chapter 2. Some Current Methods Used to Assess Publication Bias	6
Funnel Plots	6
Egger's Linear Regression.....	10
Failsafe N	13
Trim and Fill Method	14
Selection Model Approaches	18
Summary	24
Chapter 3. Modeling Publication Bias Using the Beta Density	28
The Beta Density	28
The Beta Density Weight-Function Model	31
A Solution for the Numerical Problems.....	34
ML Estimation	35
Bayesian Estimation	43

Chapter 4. Maximum Likelihood Mean-Only Simulations	48
The Simulation Design.....	48
Results	56
Conclusions	68
Chapter 5. Maximum Likelihood Moderator Simulations.....	70
The Simulation Design.....	70
Results	76
Conclusions	83
Chapter 6. Analyzing Real Meta-Analytic Data Using Bayesian Estimation	85
Method.....	86
GATB Data.....	91
ECT Temperament Data	97
Behavioral Marital Therapy Data	101
Juvenile Delinquency Data	107
Teacher Expectancy Data	114
Conclusions	123
Chapter 7. Bayesian Simulations	124
Method.....	124
Results	130
Conclusions	133
Chapter 8. Concluding Remarks	135
References	138
References	138
Appendix	146

Data Generation	146
The Beta Density Weight-Function Model in R	149

List of Abbreviations and Symbols

The notation I use in the dissertation generally follows the conventional rules of notation. Random variables are indicated by italic font, while vectors and matrices are indicated by bold font. Although abbreviations and symbols are defined where they first appear in the text, I also outline a list of the most frequently used ones here for convenience.

(A) Abbreviations:

- (1) *AIC*: Akaike's Information Criterion.
- (2) *BIC*: Bayesian Information Criterion.
- (3) CDF: cumulative distribution function.
- (4) CI: confidence interval or credibility interval, depending on the context.
- (5) $f(\cdot)$: function of $(\cdot)$.
- (6) $l(\cdot)$: likelihood of $(\cdot)$.
- (7) $L(\cdot)$: log-likelihood of $(\cdot)$.
- (8) LR: likelihood ratio.
- (9) MCMC: Markov chain Monte Carlo.
- (10) MH: Metropolis-Hastings.
- (11) ML: maximum likelihood.
- (12) MLE: maximum likelihood estimate.
- (13) *PSR*: potential scale reduction.
- (14) RMSE: root mean squared error.
- (15) SMD: standardized mean difference.

(B) Notation for distributions:

- (1) $B(a, b)$: Beta function.
- (2) χ^2 : chi-square distribution.
- (3) $f(p|a, b)$: beta density function.
- (4) $h(\cdot)$: proposal density of $(\cdot)$.
- (5) $N(\mu, \sigma^2)$: normal distribution with mean μ and variance σ^2 .
- (6) $\phi(\cdot)$: standard normal probability density function of $(\cdot)$.
- (7) $\Phi(\cdot)$: standard normal cumulative distribution function of $(\cdot)$.
- (8) $U(0, 1)$: uniform distribution.

(C) Symbols for variables:

- (1) a : first parameter that determines the shape of the beta density.
- (2) b : second parameter that determines the shape of the beta density.
- (3) β_0 : regression coefficient for the intercept.
- (4) β_1 : regression coefficient for the slope.
- (5) β : vector of regression coefficients.
- (6) D : deviance of a model.
- (7) δ_i : unknown regression parameter.
- (8) Δ_i : function of linear predictors or $\mathbf{X}_i\beta$.
- (9) df : degrees of freedom.
- (10) η_i^2 : total variability or $\sigma_i^2 + \tau^2$.
- (11) g : spacing for sub-sampling parameter values (also referred to as lag).
- (12) H_0 : null hypothesis.
- (13) j : number of parameters in the model.
- (14) k : number of effects present in the meta-analysis.
- (15) μ : mean parameter.
- (16) n : sample size.

- (17) ω : individual weight.
- (18) p : p-value.
- (19) q : number of predictors present in the meta-analysis.
- (20) R : ratio that determines whether a value is accepted or rejected.
- (21) σ_i : standard error of an effect-size estimate.
- (22) σ_i^2 : variance, or sampling variability, of an effect-size estimate.
- (23) $\sigma_{\beta_0}^2$: variance of the sampling distribution of the intercept.
- (24) $\sigma_{\beta_1}^2$: variance of the sampling distribution of the slope.
- (25) $\sigma_{\beta_0, \beta_1}$: covariance of the sampling distribution of the intercept and slope.
- (26) τ^2 : between-studies variance component.
- (27) θ^c : candidate parameter value.
- (28) w_i : weight function.
- (29) x_1 : a predictor.
- (30) $\mathbf{X}_i$: vector of predictor variables taken from the i^{th} row of design matrix $\mathbf{X}$.
- (31) ξ_i : random effect.
- (32) Y_i : effect-size estimate.
- (33) z_i : standard normal deviate.
- (34) ζ : parameter vector for parameters a , b , and β .

List of Figures

Figure 2.1 An example of a relatively symmetrical funnel plot with $\mu = 0.5$	8
Figure 2.2 An example of a non-symmetrical funnel plot with $\mu = 0.5$	9
Figure 2.3 An example of a Galbraith plot with a regression line represented by $z = -0.125 + 0.559prec$	12
Figure 3.1 An example of a plot that shows what the beta density looks like when a and b are both 1.0, representing no selection.	29
Figure 3.2 An example of a plot that shows what the beta density looks like when $a = 0.5$ and $b = 2$, representing one-tailed selection.	31
Figure 3.3 An example of a plot that shows what the beta density looks like when $a = 0.5$ and $b = 0.75$, representing asymmetric two-tailed selection.	32
Figure 3.4 A plot of estimated log-likelihood values against a sequence of corresponding mean parameter μ values. The plot indicates that the best estimate of μ is about 0.6.	37
Figure 5.1 An example of an asymmetrical funnel plot comprising two populations with true means of $\mu_1 = 0.1$ and $\mu_2 = 0.8$	72
Figure 6.1 An extended funnel plot of the general aptitude test effects against their standard errors.	91
Figure 6.2 A cumulative sum plot of the four chains of parameter $\hat{b}$ of the GATB data.	93

Figure 6.3 A trace plot of the third chain of the adjusted $\hat{\mu}$ parameter of the GATB data.	94
Figure 6.4 A bivariate trace plot of the first chain of adjusted parameters $\hat{\tau}^2$ (VC) and $\hat{\mu}$ (Mean) of the GATB data.	95
Figure 6.5 An extended funnel plot of the ECT effects against their standard errors. . . .	98
Figure 6.6 An extended funnel plot of the behavioral marital therapy effects against their standard errors.	102
Figure 6.7 A trace plot of the first chain of the adjusted $\hat{\tau}^2$ parameter of the behavioral marital therapy data.	103
Figure 6.8 An extended funnel plot of the juvenile delinquency effects against their standard errors.	107
Figure 6.9 A bivariate trace plot of the first chain of adjusted parameters $\hat{a}$ and $\hat{\mu}$ (Mean) of the juvenile delinquency data.	109
Figure 6.10 A trace plot of the fourth chain of the adjusted $\hat{\tau}^2$ parameter of the juvenile delinquency data.	111
Figure 6.11 An extended funnel plot of the teacher expectancy effects against their standard errors.	115
Figure 6.12 A plot of the beta density for the teacher expectancy data with $\hat{a}$ and $\hat{b}$ ML estimates of 1.61 and 1.53, respectively.	116
Figure 6.13 An extended funnel plot of the teacher expectancy effects dichotomized by the amount of contact teachers had with their students prior to the expectancy induction.	117
Figure 6.14 A plot of the beta density with the results of two different ML mixed-effects analyses of the teacher expectancy data.	118

Figure 7.1 An example of a bad trace plot of the adjusted $\hat{\mu}$ parameter of one data set from the new cell.	128
Figure 7.2 An example of a good trace plot of the adjusted $\hat{\mu}$ parameter of one data set from the new cell.	129

List of Tables

Table 2.1 Characteristics of Publication Bias Methods	27
Table 4.1 Rejection Rates for the Likelihood-Ratio Test for Publication Bias	57
Table 4.2 Coverage Rates of 95% Confidence Intervals for the Parameter Estimate $\hat{\mu}$ for the Unadjusted and Adjusted Models	60
Table 4.3 Coverage Rates of 95% Confidence Intervals for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model When Bias is Not Present	61
Table 4.4 $1000 \times$ Bias for the Parameter Estimate $\hat{\mu}$ for the Unadjusted and Adjusted Models	62
Table 4.5 $1000 \times$ Bias for the Parameter Estimate $\hat{\tau}^2$ for the Unadjusted and Adjusted Models	64
Table 4.6 $1000 \times$ Bias for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model When Bias is Not Present	65
Table 4.7 $1000 \times$ RMSE for the Parameter Estimate $\hat{\mu}$ for the Unadjusted and Adjusted Models	66
Table 4.8 $1000 \times$ RMSE for the Parameter Estimate $\hat{\tau}^2$ for the Unadjusted and Adjusted Models	67
Table 4.9 $1000 \times$ RMSE for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model When Bias is Not Present	68
Table 5.1 Type I Errors for the Likelihood-Ratio Test for Publication Bias	77

Table 5.2 Coverage Rates of 95% Confidence Intervals for the Parameter Estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ for the Unadjusted and Adjusted Models	78
Table 5.3 Coverage Rates of 95% Confidence Intervals for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model	79
Table 5.4 $1000 \times$ Bias for the Parameter Estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ for the Unadjusted and Adjusted Models	80
Table 5.5 $1000 \times$ Bias for the Parameter Estimate $\hat{\tau}^2$ for the Unadjusted and Adjusted Models	81
Table 5.6 $1000 \times$ Bias for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model . .	81
Table 5.7 $1000 \times$ RMSE for the Parameter Estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ for the Unadjusted and Adjusted Models	82
Table 5.8 $1000 \times$ RMSE for the Parameter Estimate $\hat{\tau}^2$ for the Unadjusted and Adjusted Models	82
Table 5.9 $1000 \times$ RMSE for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model	83
Table 7.1 Model Fit Statistics for the Four Cells	131
Table 7.2 Averaged Posterior Means of the Parameter Estimates of the Unadjusted and Adjusted Models for the Four Cells	132
Table 7.3 Credibility Intervals for the Parameter Estimates of the Unadjusted and Adjusted Models for the Four Cells	133

Acknowledgments

I would like to acknowledge the support of a number of people. First, I thank Professor Jack Vevea for being the most supportive and understanding advisor anyone can imagine. I thank him for his guidance and mentorship, and look forward to future collaboration. Second, I thank the members of my committee. Professor Anna Song's constant encouragement and Professor William Shadish's belief in me have been invaluable. Third, I would like to acknowledge the support of my friends and family. I cannot thank them enough for listening and calming me down each time something did not go as planned. I thank my parents, Ewa and Andrzej, for always believing in me; their love, understanding, and trips that forced me to get away and take a break helped me in an immeasurable way. I thank my brother, Olek, for his brotherly support; his teasing comments of "Of course I'm smarter – you are *still* in school!" always made me laugh. Last, I thank my boyfriend, Maksym, for taking care of me and nursing me back to health as I finish this.

Abstract

A Parsimonious Weight Function for Modeling Publication Bias

by

Martyna Citkowicz

Doctor of Philosophy in Psychological Sciences

University of California, Merced, 2012

Professor Jack L. Vevea, Chair

Quantitative research literature is often biased because studies that fail to find a significant effect (or that demonstrate effects that are not in the desired or expected direction) are less likely to be published. This phenomenon, termed publication bias, can cause problems when researchers attempt to synthesize results through the set of techniques known as meta-analysis. Various methods exist that estimate and correct meta-analyses for publication bias. However, no single method exists that 1) can account for continuous

moderators by including them within the model, 2) allow for substantial data heterogeneity, 3) produce an adjusted mean effect size, 4) include a formal test for publication bias, and 5) allow for correction when only a small number of effects is included in the analysis. This dissertation develops a method that tries to encompass those characteristics. The model uses the beta density as a weight function that estimates the selection process in order to produce adjusted parameter estimates. The model is implemented both by maximum-likelihood (ML) and by Bayesian estimation. The utility of the model is assessed by simulations and through use on real data sets. The ML simulations indicate that the likelihood-ratio test has good Type I error performance. However, the test is not very powerful for small data sets. Coverage rates indicate that the model's 95% confidence intervals based on adjusted parameter estimates (those that correct for bias) are more likely to contain the true parameter values than are CIs around the unadjusted parameter estimates (those that do not account for bias). Bias and root mean squared errors of the estimates are better for the adjusted mean effect than the unadjusted mean effect whenever bias is present. The ML simulations also show that the model is good at distinguishing systematic study differences from publication bias. The utility of the Bayesian implementation of the model is demonstrated in two ways: 1) when ML estimation produces nonsensical parameter estimates for real data sets, Bayesian estimation does a good job of adjusting to appropriate estimates; and 2) when bias is present in small data sets, the adjusted Bayesian parameter estimates are generally closer to the true population values than the adjusted ML estimates.

CHAPTER 1

Introduction

Published research is frequently biased because studies that fail to find a significant effect are less likely to be published than studies that obtain significant findings or that exhibit effects in a desired direction (Coursol & Wagner, 1986). One term for this phenomenon is publication bias. Publication bias is what occurs when published and unpublished data are systematically different from one another in direction or strength of effects (Veitch, 2006). The result is a published literature that is unrepresentative of all completed studies (Rothstein, Sutton, & Borenstein, 2005). Moreover, the non-reporting of non-significant findings threatens the validity of reviews of literature that contain only published studies (Gilbody, Song, Eastwood, & Sutton, 2000). Thus, while publication bias impacts studies themselves across many fields (Sterling, 1959), the problem is most often elucidated when studies are synthesized into systematic reviews (e.g., meta-analyses).

Imagine, for example, that 10 studies were performed on the efficacy of an antidepressant medication developed to reduce the symptoms of depression. Five of these studies found significantly positive results and five obtained substantially negative results. If only the five studies that found positive results (i.e., the medication works) were published, researchers conducting a systematic review would conclude that the overall mean effect is significant and that the treatment does appear to reduce depression symptoms. However, unaware of the five unpublished studies that failed to find positive results, the researchers came to the wrong conclusion. The overall mean effect of all 10 effect sizes combined is near zero; hence, the unbiased findings of the review should conclude that, in general, researchers fail to find that the antidepressant reduces depression symptoms

(or, perhaps, that the medicine is effective for some populations and not effective for others). On the other hand, it is also possible that a moderator underlies this null effect. It may be that the medication has positive effects on females and negative effects on males, creating two divergent populations for which the medication should be studied. Proper model specification is important; otherwise significant effects may be hidden, or nullified, by combining two or more populations for whom the effects are very different. While that is a simplification of the complex review process, such systematic reviews are completed regularly in order to summarize the results of studies in a particular field. Thus, in addition to biasing all literature, the non-reporting of non-significant findings can cause reviews of literature that contain only published studies to be misleading.

Meta-analysis is a set of quantitative methods for combining statistical estimates from various studies. It provides ways to synthesize the results of summary statistics from published studies, and allows the analyst to compute tests of significance for those combined results (Hedges, 1992a). The analyses are then used to describe the pattern of results (Lipsey & Wilson, 2001) in order to integrate knowledge that can inform scientific inquiry and public policy (e.g., The Cochrane Collaboration, 2010, informs policy in medicine and The Campbell Collaboration, 2010, informs social policy). In addition, meta-analysis has emerged as a central tool for integrative analysis in such diverse fields as education, criminology, and psychology (National Research Council, 1992). For these reasons, it is important that meta-analyses be as accurate and valid as possible.

However, the combined results that a meta-analysis provides may be biased if publication processes favor studies that have found significant results (see Dickersin, 2005 for a summary of the research that provides evidence for the existence of publication bias). One solution is to include effects from unpublished studies in the meta-analysis, but it is not reasonable to assume that one will be able to find all unpublished studies because unpublished studies are substantially more difficult to identify than published

studies. Another solution is to develop a method that estimates bias and corrects the meta-analytic results. The latter is an approach a number of researchers have taken (e.g., Begg & Mazumdar, 1994; Copas & Li, 1997; Dear & Begg, 1992; Duval & Tweedie, 2000a, 2000b; Egger, Smith, Schneider, & Minder, 1997; Hedges, 1984; Iyengar & Greenhouse, 1988; Rosenthal, 1979; Vevea & Hedges, 1995). Each of the researchers' methods has its own assumptions that allow for only specific kinds of meta-analytic data to be assessed. Unfortunately, no single method can account for moderators by including them within the model, allow for substantial data heterogeneity, produce adjusted parameter estimates, include a formal test for publication bias, and allow for a small number of effects to be included in the analysis without losing power or increasing the chances of Type I errors above the nominal rates. This dissertation develops a method that tries to encompass all those characteristics.

Chapter 2 of the dissertation outlines current methods used to assess publication bias. I describe the theory behind each method, explain how to implement it, and discuss any modifications that have been added over the years. I also point out the type of data that are recommended for analysis with the particular method. I conclude by noting the data characteristics no single method accounts for, and identify a starting point for a new publication bias estimation method.

Chapter 3 presents the new publication bias model. I first describe the rationale behind the model and its basis on an already existing publication bias model. I explain how the new model was developed and eventually made to work. I then define two ways in which I implement the model: 1) maximum-likelihood (ML) estimation using an optimization routine, and 2) Bayesian estimation via the Metropolis-Hastings (MH) algorithm. Ideally, both implementations will estimate the parameters appropriately. However, sometimes ML estimation has problems converging and produces nonsensical estimates.

The hope is that the flexibility of Bayesian estimation will help produce suitable parameter estimates for problematic data sets.

Chapters 4 and 5 discuss the results of two simulation studies used to assess the performance of the model when it is implemented via ML estimation. The simulation study in Chapter 4 focuses on assessing the model when no predictors are included (i.e., a mean-only model). I vary the following data characteristics in order to see how well the model performs on various kinds of data: number of effects included in the meta-analysis, severity of publication bias, and heterogeneity. In Chapter 5, I include a predictor by creating two unbiased populations and combining them into one seemingly biased population. The aim is to assess whether the model can distinguish systematic study differences from publication bias. In this set of simulations, I vary only the number of effects included in the meta-analysis and heterogeneity, as bias is not actually present in Chapter 5's simulation study. In both simulation studies, I examine Type I and Type II errors, coverage rates of 95% confidence intervals, as well as bias and root mean squared errors in order to examine model inference and how well estimation is working.

Chapters 6 and 7 provide the results of two studies in which the model is implemented using Bayesian estimation. Because running a simulation on a simulation is too time consuming, the study in Chapter 6 focuses on assessing the model using real data sets. The data sets are chosen to represent two things: 1) a variety of data characteristics (the ones mentioned above), and 2) problematic data sets (i.e., ones that ran into estimation problems when the model was implemented via ML estimation). Part of the goal is to determine whether the Bayesian implementation of the model can produce estimates that make sense when ML estimation of the model fails. Non-problematic data sets are included as well. In Chapter 7, I present the results of a small simulation study. I chose a few cells (or simulations) from Chapter 4's mean-only ML simulation study to focus on. I sampled data sets from each cell and examined how close the combined results of

the sampled data sets get to the true population parameter values by estimating Bayesian credibility intervals for the parameters. In order to check for the presence of publication bias, in both studies, I calculated the Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC).

The dissertation concludes with an Appendix that includes R code for the simulations. I demonstrate how biased data was generated for the ML simulations. I also provide annotated R code that illustrates how to implement the beta density weight-function model using both R 's optimization algorithm and the Metropolis-Hastings (MH) algorithm.

CHAPTER 2

Some Current Methods Used to Assess Publication Bias

Numerous methods for addressing the problem of publication bias methods exist. Some methods test whether publication bias is present, others adjust the meta-analytic mean for bias, and some are simply visual aids. The goal of this chapter is to describe the most commonly used (i.e., widely cited) and the most sophisticated publication bias estimation methods. By doing so, I will illustrate the kind of publication bias model that is currently missing.

Funnel Plots

Funnel plots are commonly used in order to obtain a visual assessment of possible publication bias. They are scatterplots of the treatment effect-size estimate Y_i plotted against an index of study size or weights employed in the meta-analysis (Sterne, Becker, & Egger, 2005). Conventional funnel plots present the treatment effect on the x-axis and the index of study size (or equivalent) on the y-axis. Effect-size estimates that come from large studies have small sampling error σ_i , so that the top of the plot represents larger studies that cluster around the mean effect size. The bottom of the plot is more spread out, representing smaller studies with larger sampling error variation. That creates the appearance of an upside-down funnel (Rothstein, 2008). (Note that many reverse the plot by presenting small studies at the top of the plot and large studies at the bottom.)

However, funnel plots presented in this dissertation resemble a sideways funnel because the variables presented on the axes are switched (i.e., the index of study size is on the x-axis and the treatment effect is on the y-axis). Now the left-hand side represents

smaller studies and the right-hand side represents larger studies, rather than the bottom and top, respectively. The reason for the change in format is simple: the usual convention in statistical graphics is to put the fixed variable (e.g., index of study size, inverse of the conditional variance, or inverse of the standard error) on the x-axis and the random variable (Y_i) on the y-axis. The plot then contains horizontal aspect ratio where the x-axis is longer in length than the y-axis, allowing for optimal use of space (Cleveland, 1993, 1994; Tufte, 2006). Most funnel plots presented in the literature do not follow that rule, and hence do not make good use of the space available for the graph. I would like to encourage better display of graphics; thus, the sideways funnel is the format presented in this dissertation. Figure 2.1 shows an example of a fairly symmetrical funnel plot with a true mean effect size μ of 0.5.

A symmetrical funnel-shaped plot suggests a lack of bias because larger and therefore more precisely estimated studies cluster around the true mean effect size, while less precise studies show more variability about the true mean. (Note that an exception to that exists as a funnel plot that involves two-tailed selection¹ and has an effect near zero also displays symmetry, although with a hollow base.) That creates the appearance of a funnel (Rothstein, 2008). Conversely, when bias is present, usually the top left-hand side of the plot (when the plot is graphed the way I advocate; i.e., with horizontal aspect ratio) will contain more studies than the bottom left-hand side because smaller studies are more likely to be published when they obtain large, statistically significant effects (Sterne et al., 2005; Rothstein, 2008). Figure 2.2 shows the same effects as Figure 2.1; however, the small non-significant effects are now missing, providing the appearance of a non-symmetrical and biased funnel plot. The two figures are placed in the same location on subsequent pages for ease of comparison. Note that smaller studies are less likely to

¹Two-tailed selection can occur when there is no expectation of an effect in either direction (positive or negative). Thus, published studies represent divergent findings with the middle non-significant studies not published. See Vevea and Woods (2005) or Hedges and Vevea (2005) for a discussion of two-tailed selection.

FIGURE 2.1. An example of a relatively symmetrical funnel plot with $\mu = 0.5$.

obtain significant results in general, while even a small effect may be significant for large studies.

One problem with funnel plots is that it is not easy to determine whether the plot looks like a symmetrical funnel (i.e., the interpretation of funnel plots is a subjective process), especially when the number of effect sizes (k) included in the meta-analysis is small. That is because the smaller the number of effects included in the meta-analysis, the less information there is about what the distribution of effects looks like in the population (i.e., the plotted distribution of effects is less clear and potentially misleading; see Tang & Liu, 2000 for examples of ambiguous funnel plots with small k). In addition, a symmetrical plot does not necessarily indicate a lack of bias. The plot may, for example,

FIGURE 2.2. An example of a non-symmetrical funnel plot with $\mu = 0.5$.

comprise two different populations with different means and different variances, creating the appearance of one symmetrical plot (Wang & Bushman, 1998). That violates an assumption of the funnel plot. As funnel plots assume that the effects come from *one* population, their use is deemed inappropriate with heterogeneous data (Terrin, Schmid, Lau, & Olkin, 2003). Funnel plot asymmetry is also not necessarily caused by publication bias. It may be caused by situations where both sample size and effect size are systematically related to the same study characteristic. Thus, funnel plots can be good visual aids, but they should only be used with more objective methods that include formal tests for bias. Those methods are discussed below.

Egger's Linear Regression

Both correlation and regression approaches have been developed in order to quantify funnel plot asymmetry, thus removing the subjective component in its evaluation. Correlation methods specify the strength of a relationship between two variables. Begg and Mazumdar's (1994) rank correlation method quantifies funnel plot asymmetry by examining the strength of the association between effect-size estimates and their sampling variances. Regression methods, on the other hand, define *what* the relationship is between two variables. Egger's linear regression approach determines whether there is a linear relationship between the effect-size estimates and their standard errors (Egger et al., 1997). The latter is more informative when determining the presence of bias and what the bias is, and is thus preferred. That preference is obvious when one looks at how often each method has been cited. For that reason, only Egger's method will be discussed. See Begg and Mazumdar (1994) or Sterne and Egger (2005) for a discussion of Begg and Mazumdar's rank correlation method.

Egger's linear regression approach consists of regressing the standard normal deviate z_i on its precision $prec_i$:

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

where $z_i = \frac{Y_i}{\sigma_i}$, $prec_i = \frac{1}{\sigma_i^2}$, β_1 represents the slope, β_0 represents the intercept, Y_i represents the effect sizes, and σ_i denotes the standard errors (Sterne & Egger, 2005).² The scatterplot of the standard mean deviates plotted against their precision is known as a Galbraith or radial plot (Galbraith, 1988). The plot, along with its rationale and expected pattern, is discussed below.

²Note that the original paper by Egger et al. (1997) suggested an alternate version of the approach, where the regression of z_i on $prec_i$ is weighted by $\frac{1}{\sigma_i^2}$, or the inverse of the variance. However, the weighted version was not backed by theoretical justification and is no longer supported by Egger et al. (Sterne & Egger, 2005).

Small studies have larger standard errors, thus, their $prec_i$ will be close to zero. Their effect sizes will vary; however, because of their large standard errors, small studies' standard mean deviates z_i will also be close to zero. Large studies, on the other hand, are expected to produce precise effect-size estimates that are close to the true effect, whether it be small or large. That is, large studies will contain smaller standard errors, leading to higher $prec_i$ than small studies, as well as z_i 's that are closer to the true z_i . Thus, a Galbraith plot showing points (i.e., effect sizes) that scatter about a regression line through the origin [0,0] represents a symmetrical funnel plot (i.e., no bias). The slope of the line specifies the size and direction of the effect. Bias, or funnel plot asymmetry, exists when the regression line does not run through the origin, indicating the intercept's deviation from zero (Egger et al., 1997). A simple t -test with $k - 2$ degrees of freedom (df), where k represents the number of effects included in the analysis, can be performed in order to determine whether the intercept significantly differs from zero. If the intercept is significant, there is evidence of bias (Sterne & Egger, 2005). Figure 2.3 shows an example of a Galbraith plot with a regression line of $z = -0.125 + 0.559prec$. The t -test on the intercept fails to support the idea that the intercept differs from zero ($t = -1.030$, $df = 498$, $p = 0.303$), suggesting funnel plot asymmetry has not been found.

Egger's approach is simple and easy implement, but with those advantages come a number of limitations. The approach is not recommended for use with heterogeneous data (or excess between-study variability) due to inflated Type I errors (Peters, Sutton, Jones, Abrams, & Rushton, 2006). The method also does not allow for moderators of the effect size to be assessed without splitting the data into the moderator's categories; thus, continuous moderators cannot be accounted for. Extensions of Egger's linear regression have been developed in order to address both of those issues (see, e.g., Harbord, Egger, & Sterne, 2006; Macaskill, Walter, & Irwig, 2001; Peters et al., 2006; Thompson & Sharp, 1999), and each modification has its advantages and disadvantages. Moreover, according

FIGURE 2.3. An example of a Galbraith plot with a regression line represented by $z = -0.125 + 0.559prec.$

to a series of simulations with $k = 5, 10, 20$, and 30 , at least a moderate number of effect sizes ($k > 10$) should be included in the analysis to avoid a substantial loss in power (Sterne, Gavaghan, & Egger, 2000). Last, as publication bias is not the only possible cause of funnel plot asymmetry, Egger's linear regression approach should not be used to detect publication bias directly. Instead, it should be used as a tool for examining the evidence for small study effects, or the tendency for smaller studies to show larger effects in meta-analyses (Sterne & Egger, 2005). As mentioned in the previous section, this phenomenon may be due to publication bias, but can also arise when there are systematic differences in the population characteristics of large and small studies.

Failsafe N

One of the first methods developed to assess publication bias was the file-drawer analysis (Rosenthal, 1979), also referred to as the failsafe N (Cooper, 1979). The analysis determines the number of studies it would take to reduce to non-significance the overall test of significance for the combined effect size. Stouffer's test of combined significance, often called the 'sum of Z s', is:

$$Z_S = \frac{\sum_{i=1}^k z_i}{\sqrt{k}},$$

where z_i ³ denotes the standard normal deviate associated with a one-tailed p -value and k represents the number of effects included in the analysis (Stouffer, Suchman, DeVinney, Star, & Williams, 1949). Z_S tests the null hypothesis that all k values of Y_i are zero (Becker, 2005).

One can examine the failsafe N with the following formula:

$$N > k \left(\frac{Z_S}{Z_\alpha} \right)^2 - k,$$

where Z_S represents Stouffer's test, Z_α is the z_i at a particular one-tailed α -level, and k represents the number of effects included in the analysis. The idea behind this method is that bias is present when Stouffer's test is large. Stouffer's test is large when studies produce large z_i -values. Large z_i -values can result in two ways: 1) large sample sizes n_i , as large samples are more likely to produce significant effects, and 2) large observed effects. In other words, bias is considered present when mostly large effect-size studies are present and small non-significant effect-size studies are missing. According to Rosenthal, the meta-analytic results are robust to publication bias when the failsafe N is large relative to k (i.e., if the left-hand side of the formula is larger than the right-hand side). That is

³Recall that $z_i = \frac{Y_i}{\sigma_i}$, where Y_i represents the effect sizes and σ_i denotes the standard errors.

because it would be improbable to find a large number of additional unpublished effects. Rosenthal later added a rule of thumb that said if the failsafe N is less than $5k + 10$ study effect sizes then one should be concerned. However, with no real statistical criterion to assess when the failsafe N is large enough (when compared to the number of observed studies k) to indicate that the meta-analytical results are presumed to be free of publication bias (other than the above-mentioned rule of thumb), this measure is subjective and prone to misinterpretation (Becker, 2005).

Since its development, there have been several modifications of the failsafe N in order to reduce subjectivity by including a formal criterion for assessment, as well as to overcome the assumption that unpublished studies have no effects (rather than recognizing that some results are in the opposite direction, which would lead to a smaller failsafe N ; see Orwin, 1983). An advantage is that a large number of effects are not necessary to calculate the failsafe N . However, none of the variations take into account continuous moderators, and they produce varying estimates that can lead to different conclusions. Despite the availability of these more refined approaches, the original failsafe N is still the most commonly used method among these variations even though it asks the wrong question and makes the wrong assumption. It tells us what we already know: that the results of meta-analyses based on a small number of effects can be readily changed when just a few more effects are added to the analysis. What researchers really want to know is the magnitude of the true population effect (Becker, 2005).

Trim and Fill Method

Duval and Tweedie's (2000a, 2000b) trim-and-fill approach is another method that examines funnel plot asymmetry. The trim-and-fill method is a type of sensitivity analysis used to assess the possible impact of hidden study effects on meta-analyses. It does so by determining what additional effects would need to be present in the meta-analysis in

order for the funnel plot to be symmetrical. It then imputes those missing effects using an iterative process and re-estimates the overall mean effect size. The new mean effect size is an adjusted mean that takes into account the outcomes of potential missing effects-size estimates (Duval, 2005).

The idea is relatively simple. First, the key assumption of the trim-and-fill method is that the studies with small effect-size estimates at the bottom of the funnel plot are likely to be missing due to publication bias (recall Figure 2.2). Those effect sizes are least likely to be published due to their probable non-significance. Thus, after assuming that the smallest effects are missing, one estimates the number of missing effects using an L_0 , R_0 , or Q_0 estimator (estimators will be defined below). Then the asymmetric top portion of the plot, which consists of studies that obtained large effect sizes, is removed (or trimmed) from the data set according to the estimated number of missing effects. The trimming, or removal of effects, is only temporary and is completed in order to calculate a new mean effect-size estimate from the remainder of the (presumably more symmetric) effect sizes. The goal is to create a symmetrical plot (via trimming), obtain its mean effect-size estimate, and replace the trimmed effect sizes (and their imputed counterparts) evenly around that new mean estimate and into the already symmetric plot. The counterpart effect sizes represent those potentially missing effects that, when not published, lead to funnel plot asymmetry. If the plot is not fully symmetrical after the first trimming has taken place, the number of missing effects is estimated again, the effects get trimmed accordingly, and a new mean estimate is calculated. The process is repeated until the funnel plot appears symmetric about the latest mean effect-size estimate. Next, one returns the trimmed effects to the data set, along with their counterpart fillers to fill in the asymmetrical side of the plot. That creates a fully symmetrical funnel plot. The filled funnel plot is then used to calculate the estimate of the true overall effect μ , along with its new standard error s (Duval, 2005; Duval & Tweedie, 2000a, 2000b).

The above may sound daunting, but this step-by-step outline of the trim-and-fill algorithm will show how straightforward the trim-and-fill approach really is:

- (1) Order the observed data according to the effect-size magnitude.
- (2) Estimate a meta-analytic mean effect size μ .
- (3) Center the data around μ .
- (4) Estimate the number of k_0 missing effects.
 - (a) Rank the absolute values of the centered effect-size data. Then, give those ranks the sign (+/-) associated with the centered value.
 - (b) Chose an L_0 or R_0 estimator.⁴

$$L_0 = \frac{4S_{rank} - k(k+1)}{2k-1},$$
 where S_{rank} is the sum of the positive ranks and k represents the number of effects included in the analysis.

$$R_0 = \gamma^* - 1,$$
 where γ^* represents the rightmost run of ranks (or the total number of the highest positive ranks that are still in consecutive order).
 - (c) Estimate the number of missing effects and round the estimate up to a whole number. Note that if the L_0 or R_0 estimates are less than or equal to zero, the original mean estimate from Step 2 is the final estimate, as no effects would be considered missing.
- (5) Trim the topmost effects based on the k_0 value calculated in Step 4 and re-estimate μ .
- (6) Center the original data around the new mean estimate.
- (7) Estimate the number of missing effects (refer to Step 4).

⁴Across a k range of 25 to 75 effect sizes, simulations have shown R_0 to be the most accurate over the entire range when $k_0 = 10$ (compared to when $k_0 = 0$ or 5), while L_0 tends to underestimate for smaller values of k . L_0 performs slightly better than R_0 in terms of standard errors when k is smaller. Last, when the number of missing effects is estimated to be more than 25% of those observed, L_0 is the recommended method (Duval & Tweedie, 2000b). Note that a third estimator Q_0 exists. However, it is not recommended because L_0 is a simple linearization of Q_0 , and Q_0 does not possess superior qualities to either L_0 or R_0 (Duval, 2005). If interested in Q_0 's calculation, see Duval and Tweedie (2000a, 2000b).

- (8) If the estimated number of effects missing differs from Step 4 to Step 7, trim and re-estimate following Steps 5 through 7. Repeat this process until the ranks stabilize (i.e., do not change).
- (9) Produce the imputed missing effects by reflecting the latest centered data's topmost effect sizes (fillers) symmetrically about the latest μ . Then take the conditional variances from the original data's topmost effect sizes and apply them as the conditional variances of the new imputed effect sizes.
- (10) Combine the original sorted data with the imputed data, and estimate the final meta-analytic estimates.

As with many methods, a key assumption of the trim-and-fill method is that funnel plot asymmetry is due to publication bias (Duval, 2005), rather than to true differences between small and large studies. Furthermore, the trim-and-fill approach assumes that publication bias follows a deterministic pattern: it is always the most extreme points in the tail of the distribution that are assumed to be missing. The method uses that assumed pattern to impute the missing effects (Rothstein, 2008). However, selection is not deterministic; it does not follow this orderly pattern, as studies within the distribution may be missing as well (not just those in the extreme tail) and the occasional study in the extreme tail may happen to survive. This assumption of a deterministic mechanism is useful, as it allows the for imputation of potentially missing effects, which then allows one to compute an adjusted mean effect. But it is not recommended that the adjusted mean effect be viewed at the true population mean. Instead, it should be compared to the unadjusted mean to examine the meta-analytic data for publication bias.

Until now, the trim-and-fill approach has not been able to include continuous moderators, but a more recent formulation of the method may be able to employ a single moderator (S. Duval, personal communication, July 6, 2010). The existing trim-and-fill method should not be used with heterogeneous data as that allows for the possibility for

the trim-and-fill approach to adjust for publication bias when none actually exists (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin et al., 2003). While this method has great potential, a weakness is that there is no formal test for publication bias. Instead, the adjusted mean effect size is informally compared to the original effect size calculated before imputation in order to determine the robustness of the meta-analytic results to the potential threat of publication bias (Peters et al., 2007; Rothstein, 2008). That is, no statistical criterion exists to determine whether the difference in mean effect sizes is statistically significant.

Selection Model Approaches

The selection model approaches are a set of methods that assess or correct for the presence of publication bias by explicitly modeling the selection process, or the process by which studies are chosen for publication. To correct for publication bias, two models are required: 1) the effect-size model, which notes what the distribution of effect-size estimates would be if there were no selection (i.e., a standard meta-analytic model), and 2) the selection model, which identifies how the distribution is changed by the selection process. The selection model is defined by a weight function that estimates the likelihood that an effect-size estimate with a particular p -value will be observed (Hedges & Vevea, 2005). Different models define the likelihoods (or weights) in different ways, according to their assumptions. Hedges (1984) defined a simple early weight-function model by denoting a weight of zero for a non-significant study ($p > 0.05$) and a weight of one for a significant study ($p < 0.05$). Iyengar and Greenhouse (1988) employed two weight functions, both of which were based on the t -distribution. Under the first function, the weight increased from zero to one as the estimated two-tailed p -value decreased from 1.0 to 0.05. With the second function, the weight was set to a constant, non-zero value when p was non-significant ($p > 0.05$), and one otherwise. Dear and Begg (1992) estimated

the p -value intervals (or ranges of p -values) that define a set of weights by calculating their cutpoints (or points of discontinuity). Vevea and Hedges (1995) defined the weights as step-functions across a number of p -value intervals determined a priori (see Hedges, 1992b; Vevea, Clements, & Hedges, 1993 for previous versions of this weight-function model). The specification of p -value intervals ahead of time, rather than estimating cutpoints, provided Vevea and Hedges' (1995) weight-function approach with more power than other weight-function methods because there are fewer parameters to estimate with pre-specified intervals.

The above-mentioned selection approaches are part of a class of approaches that depend solely on the p -value. The weights are estimated in order to describe the selection process. If selection is not present, the estimated weights will be relatively constant, and the estimated parameters of the effect-size model (i.e., mean effect or regression coefficients, and variance component) will be similar to estimates from a conventional random- or mixed-effects meta-analysis. However, if there is differential selection based on statistical significance, the weights associated with less significant p -values will be smaller, and the estimated mean effect will move toward zero. The goal of this simultaneous estimation of the effect-size and selection models is to produce appropriately adjusted estimates of the parameters of the effect-size model (Hedges & Vevea, 2005).

In addition, other selection models exist that depend both on the effect-size estimate Y_i and its standard error σ_i separately. This second class is presented in a series of papers by Copas (1999; Copas & Li, 1997; Copas & Shi, 2001). The models test whether the standard error still predicts the effect size after selection is taken into account. However, because the selection model cannot be estimated from effect-size data, as there is not enough information, the regression model parameters are fixed, and the Copas model becomes a very complicated sensitivity analysis rather than a direct estimate of selection (Hedges & Vevea, 2005).

An advantage of the selection model approaches is that they are better at examining publication bias than simpler methods when data are heterogeneous (see Terrin et al., 2003) because some can incorporate the between-study variance component into the model. They can also incorporate both discrete and continuous moderators, allowing one to distinguish between systematic study differences and publication bias. The more recent approaches produce adjusted parameter estimates, as well as provide formal tests for publication bias. The cost of all that, however, is that selection model approaches require a large number of effects for precise estimation.⁵ That is because with a small number of effects, there is not much information about the weights. That leads to difficulties in estimating the selection model (Hedges & Vevea, 2005).

Vevea and Hedges' Weight-Function Model

Vevea and Hedges' (1995) weight-function model is considered one of the most sophisticated and powerful selection estimation approaches (see, e.g., Terrin et al., 2003). It is the most generally useful selection model approach because it represents hypothesis testing in fields like psychology by taking into account different p -value intervals. (The literature has shown that social and psychological researchers gain confidence in results that obtain p -values below the conventional levels of significance, specifically, below $p = 0.05$, creating a *cliff effect* at the 0.05 level; see, e.g., Greenwald, 1975; Nelson, Rosenthal, & Rosnow, 1986; Rosenthal & Gaito, 1963, 1964.) That a priori assumption of breaking at different p -values gives the test more power, as there is less to estimate

⁵Vevea and Woods (2005) developed a sensitivity analysis that allows for the assessment of a smaller number of effects. Rather than estimating the weights for the selection model, assumptions are made about how a putative selection process might look in order to set the appropriate weights. Numerous fixed weight functions are created in order to represent varying degrees of biased selection. The patterns of resulting estimates are then examined in order to determine whether publication bias may be an issue. Publication bias may be an issue when the mean estimates change depending on the weight-function pattern, usually decreasing with more severe selection. The advantage of this weight-function approach is that it allows for the examination of heterogeneous data, as well the incorporation of continuous moderators (while allowing for small k). However, if publication bias is plausible, a single adjusted mean effect-size will not be estimated, and the approach does not contain a formal test for publication bias.

in Vevea and Hedges' (1995) model than with a model that estimates both the points of discontinuity (i.e., p -value break points) and the weights (e.g., Dear and Begg's, 1992, approach). Thus, I will now focus only on the Vevea and Hedges' (1995) weight-function approach.

As mentioned above, one must estimate both the effect-size and the selection models in order to correct for bias.

Effect-size model. The effect-size model is the standard meta-analytic model that does not account for bias. There are k studies with effect-size estimates Y_i , such that:

$$Y_i \sim N(\delta_i, \sigma_i^2),$$

where variance σ_i^2 is approximately known because of its dependence on sample size n . Effect parameter δ_i is an unknown regression parameter determined via q known predictors:

$$\delta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + \xi_i,$$

where ξ_i is a random effect assumed to have a normal distribution with a mean of zero and unknown variance τ^2 . To approximate a simple random-effects model with no predictors, one can simply estimate the following model:

$$Y_i \sim N(\beta_0, \sigma_i^2 + \tau^2),$$

where β_0 denotes the random-effects mean and $\sigma_i^2 + \tau^2$ represents sampling variability plus a variance component⁶ that reflects true effect-size variability. In order to include predictors, one must also define a vector of predictor variables as $\mathbf{X}_i = (X_{i1}, \dots, X_{iq})'$ and a vector of unknown regression coefficients as $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_q)$, where q denotes the number of predictors (Hedges & Vevea, 2005).

⁶One can estimate a fixed-effects model by setting the between-studies variance component τ^2 to zero.

The observed effect Y_i is used to test the null hypothesis $H_0 : \delta_i = 0$ using the test statistic:

$$Z_i = \frac{Y_i}{\sigma_i}.$$

The two-tailed p -value associated with the test is:

$$p_i = 2(1 - \Phi(Z_i)),$$

where $\Phi(Z_i)$ denotes the standard normal cumulative distribution function (CDF) evaluated at Z (Vevea & Hedges, 1995; Hedges & Vevea, 2005).

Selection model. The selection model explicitly models the selection process via a weight function that estimates the probability that an effect-size estimate with a particular p -value will be observed. Those probabilities are defined by step-functions across a number of p -value intervals determined a priori. The boundaries of those intervals are set at p -values important to the researcher (Vevea & Hedges, 1995).

For example, cutpoints of $p = 0.001, 0.05$, and 0.10 , representing p -value intervals of 0 to 0.001, 0.001 to 0.05, 0.05 to 0.10, and 0.10 to 1.0, are theoretically important due to their common significance levels. In addition, $p = 0.5$ is theoretically important because it represents the boundary at which effect-size estimates become negative. (Note that effect sizes become negative at $p = 0.5$ when data consist of standardized mean differences, correlations, or log-odds ratios, but this is not true for some other kinds of data, such as risk ratios.) In order to distribute the observed effect sizes more evenly, as they tend to group together across just a few of the intervals, one should add more intervals in between those just mentioned. Next, the weight function $w(p_i)$ is defined as the relative likelihood that an effect-size estimate with a one-tailed p -value of p_i is observed (Vevea & Hedges,

1995). Here is an example of a weight function that contains seven p -value intervals:⁷

$$w(p_i) = \begin{cases} \omega_1 = 1, & \text{if } 0 \leq p < 0.001, \\ \omega_2 & \text{if } 0.001 \leq p < 0.010, \\ \omega_3 & \text{if } 0.010 \leq p < 0.050, \\ \omega_4 & \text{if } 0.050 \leq p < 0.100, \\ \omega_5 & \text{if } 0.100 \leq p < 0.200, \\ \omega_6 & \text{if } 0.200 \leq p < 0.500, \\ \omega_7 & \text{if } 0.500 \leq p < 1.000. \end{cases}$$

Given the weight function $w(p_i)$ and parameters β , τ^2 , and $\omega = (\omega_1, \dots, \omega_k)'$, the weighted probability density function of Y_i is:

$$f(Y_i | \beta, \tau^2, \omega) = \frac{w(p_i) \phi\left(\frac{Y_i - \Delta_i}{\eta_i}\right)}{\eta_i A_i(\beta, \eta_i^2, \omega)},$$

where

$$A_i(\beta, \eta_i^2, \omega) = \frac{1}{\eta_i} \int_{-\infty}^{\infty} w(p_i) \phi\left(\frac{Y_i - \Delta_i}{\eta_i}\right) dY_i,$$

$\Delta_i = \mathbf{X}_i \beta$, $\phi(z)$ is the standard normal density function evaluated at z , and $\eta_i^2 = \sigma_i^2 + \tau^2$.

This function represents an adjusted model that accounts for publication bias by including a weight function $w(p_i)$ that models the selection process (Vevea & Hedges, 1995).

Testing for bias. The parameters of the unadjusted (effect-size) and adjusted (effect-size + selection function) models are estimated via maximum-likelihood (ML) estimation

⁷The estimated weights ω_i are relative because the number of effects present prior to publication biases is unknown. Also, the weight of the first p -value interval is constrained to 1.0 in order to identify the model. That allows for comparisons to be made. For example, if $\omega_3 = 0.5$, the effects in the third interval ($0.010 \leq p < 0.050$) would be only half as likely to be observed as in the first interval ($0 \leq p < 0.001$). Similarly, if $\omega_6 = 2.0$, the effects in the sixth interval ($0.200 \leq p < 0.500$) would be twice as likely to be observed as in the first interval ($0 \leq p < 0.001$; Hedges & Vevea, 2005; Vevea & Hedges, 1995).

on each model's log-likelihood function. Because the models are nested within one another, a formal test can be performed in order to assess whether the differences in the models' estimates are significant or whether they would be likely to arise simply due to chance. The test is a simple likelihood-ratio test that compares the log-likelihoods of the effect-size model with and without the inclusion of the selection model (Vevea & Hedges, 1995). Minus two times the difference between the log-likelihood values of the two models [i.e., $-2(L(\text{adjusted}) - L(\text{unadjusted}))$] approximates the chi-square (χ^2) distribution under the null hypothesis that the weights equal one. The degrees of freedom are $j - 1$, where j represents the number of weight intervals (see, e.g., Rice, 2007). The test examines the significance of the weights, testing the null hypothesis that the weights equal one. If the weights equal one (i.e., the null hypothesis is true), that indicates that bias may not be present as the likelihood of being published would be equal across the p -value intervals. If the null hypothesis is rejected, that indicates that adding a weight function to the effect-size model improves model fit considerably and, thus, bias is present.

Summary

I have examined both visual and statistical techniques that have been developed in order to assess publication bias in meta-analysis. Those methods include: the funnel plot, regression and correlation methods, the failsafe N , the trim-and-fill approach, and the selection model approaches (specifically, Vevea and Hedges' weight-function approach). The methods vary in what they can account for or include in their models. Some methods are not recommended for use with heterogeneous data, or data that are spread about the mean by more than just random sampling of persons (or other primary units) into studies, due to Type I error inflation (or chance of concluding that bias is present when it is actually not). Other methods can account only for categorical moderators, as the data must be split into the moderator's categories in order to separately assess each part of the data for

bias. Only a fraction of the methods provide an adjusted mean effect size after taking publication bias into account. And only a few methods include a formal test for publication bias.

There is no single method that can account for moderators by including them within the model, allow for substantial data heterogeneity, produce an adjusted mean effect size, include a formal test for publication bias, and allow for a small number of effects (e.g., 10 effect sizes) to be included in the analysis without losing power or increasing the chances of Type I errors above the nominal rates. The objective of this dissertation is to create a method that can encompass those characteristics.

The selection model approaches are the most effective of the publication bias methods when large numbers of effects are available for analysis, as their only significant limitation is that they require numerous effects for precise estimation of the weight function (see Table 2.1 for a summary of the publication bias methods' characteristics). (Note that this is a generalization as some selection models are simplistic and may, for example, denote a binary weight of zero or one to represent the selection process, allowing for a small number of effects to be included in the analysis, but they may not estimate selection efficiently.) Given that this is their only real limitation and that the selection method approaches have numerous advantages over some of the other publication bias methods, I would like to follow their technique of creating a weight function that models publication bias in order to develop a new publication bias method. The goal is for the method to be able to incorporate the following (without losing substantial power or increasing Type I errors): include continuous moderators, allow for substantial data heterogeneity, produce adjusted parameter estimates, include a formal test for publication bias, and be effectively estimable even if only a small number of effects are available for analysis. (Note that allowing as few as 10 effects to be included in the analysis would be ideal, but reducing the

minimum required for the analysis to be valid to 20 or 30 effects would still be beneficial and more realistic.)

TABLE 2.1. Characteristics of Publication Bias Methods

Method Characteristics	Funnel Plot	Egger's Linear Regression	Rank Correlation Method	Failsafe N	Trim- and-fill Method	Selection Model Approaches ¹
$k > 10$ necessary	X	X	X			X
Homogeneous data recommended	X	X	X ²	? ³	X	
Moderators can be incorporated into the model		X ⁴			X ⁵	X
Assumes funnel plot asymmetry is due to publication bias	X	X	X		X	
Produces an adjusted mean effect size					X	X
Includes a formal test for publication bias		X	X	X ⁶		X
Subjective	X			X		

Note: k represents the number of effects included in the meta-analysis and n represents the study sample size.

¹ Note that the characteristics presented here are a generalization as some selection models are stronger than others.

² Particularly with binary data, the rank correlation method is not recommended when heterogeneity is present because power has been found to decrease as bias increases (Schwarzer, Antes, & Schumacher, 2007).

³ Because I have been unable to find any simulations that examine heterogeneity when calculating the failsafe N , I cannot conclude whether the approach is recommended for data that are heterogeneous.

⁴ The original Egger's linear regression approach does not incorporate moderators into the model; however, some of the model's extensions can include moderators.

⁵ A more recent formulation of the trim-and-fill approach may be able to employ a single moderator (S. Duval, personal communication, July 6, 2010).

⁶ Whether a test for publication bias is available depends on which variation of the failsafe N one uses.

CHAPTER 3

Modeling Publication Bias Using the Beta Density

Vevea and Hedges' (1995) weight-function model is considered one of the most sophisticated and powerful selection estimation approaches (see, e.g., Terrin et al., 2003). The main disadvantage of the approach is that it requires a large number of effects to precisely estimate the selection model. That is because there are many parameters to estimate, as a set of p -value interval weights must be estimated in addition to the meta-analytic parameter estimates. To resolve this problem, Vevea and Hedges (1996) proposed replacing Vevea and Hedges' (1995) step function (i.e., weight function) with the beta probability density function (also referred to as the beta density or beta pdf). However, they encountered numerical difficulties when estimating the selection model, resulting in a scaled down version that still did not work well. I have overcome those numerical estimation problems. I will focus the discussion of this chapter on the model and its development.

The Beta Density

The beta distribution is a family of continuous probability functions that are used to describe distributions of values that are known to exist within the interval $[0,1]$ (DeGroot & Schervish, 2002). The beta density describes the relative likelihood that a random variable p will occur, given the shape of the distribution. The density is defined by:

$$(3.1) \quad f(p|a, b) = \frac{1}{B(a, b)} p^{a-1} (1 - p)^{b-1},$$

FIGURE 3.1. An example of a plot that shows what the beta density looks like when a and b are both 1.0, representing no selection.

where a and b are parameter values that determine the shape of the beta density and $B(a, b)$ denotes the Beta function (Johnson & Kotz, 1970). Note that, in this beta density weight-function model, p represents the p -values, or probabilities of obtaining the effect-size estimates given their variances. Thus, $f(p|a, b)$ represents the likelihood that a particular p -value will be observed given the shape of the distribution the p -values come from.

There are a number of reasons for using the beta density as the weight function. First, notice that there are only two parameters to estimate. With only a and b to estimate, rather than a set of weights, it should be possible to estimate the selection model without a large number of effects.

A second reason for using the beta density is that when a and b are both equal to one, the beta density becomes the uniform distribution $[0,1]$ (Johnson & Kotz, 1970). The uniform distribution represents no selection (i.e., no publication bias), as the likelihood of being published is the same for all p -values. Figure 3.1 shows a plot of the beta density when it represents no selection. Recall that the effect-size model assumes no selection, estimating what the distribution of effect-size estimates would be if bias were not present. Thus, setting parameters a and b both to 1.0 will estimate the standard meta-analytic model, and if a and b happen to be *estimated* to be 1.0, the standard meta-analytic estimates will also emerge. Fixing a and b to 1.0 will allow for a formal comparison of the effect-size model with and without the inclusion of the selection model in order to assess whether publication bias is present. (The test for bias will be discussed later.)

A third reason for using the beta density to estimate the weight function is that the beta density can represent both one-tailed and two-tailed selection. One-tailed selection occurs when an effect is expected in one direction, resulting in published studies that are skewed towards effects that are either only positive or only negative. Figure 3.2 shows a plot of the beta density when it represents one-tailed selection. Note how the relative likelihood of being published decreases as the p -value increases, indicating the presence of bias. Two-tailed selection can occur when there is no expectation of an effect in either direction (positive or negative).¹ Thus, published studies will represent two divergent findings (e.g., males perform better on a specific task sometimes and other times females are found to perform better on that same task), with the middle non-significant studies remaining unpublished. Figure 3.3 shows a plot of the beta density when it represents asymmetric two-tailed selection. Notice the big dip in the middle of the plot where the non-significant studies are present. The lines curving upward on each end signify the presence of two-tailed selection. Note, however, that while the plot depicts two divergent

¹Note that two-tailed selection can occur when there *is* an expectation of an effect in one direction. The beta density weight-function model can detect two-tailed selection even when it is not assumed to be present.

FIGURE 3.2. An example of a plot that shows what the beta density looks like when $a = 0.5$ and $b = 2$, representing one-tailed selection.

findings, one set of findings is more favored than the other. That is indicated by the asymmetric aspect of the lines in the plot (and the different values for a and b : 0.5 and 0.75, respectively). The beta density weight-function model can represent both symmetric and asymmetric two-tailed selection. Two-tailed selection (symmetric or asymmetric) is not as common as one-tailed selection; however, it can occur and needs to be represented as well.

The Beta Density Weight-Function Model

The beta density weight-function model that adjusts for bias includes both the effect-size and selection models. The effect-size model is the same one as in Chapter 2 (defined

FIGURE 3.3. An example of a plot that shows what the beta density looks like when $a = 0.5$ and $b = 0.75$, representing asymmetric two-tailed selection.

under the section *Vevea and Hedges' Weight Function Model*) and the selection model, or the model that illustrates the relative likelihood that an effect-size estimate with a particular p -value will be observed, is represented by the beta density. The beta density weight-function model is the weighted probability density function of Y_i given parameters a , b , β , τ^2 , and σ_i^2 :

$$(3.2) \quad f(Y_i|a, b, \beta, \tau^2; \sigma_i^2) = \frac{p_i^{a-1}(1-p_i)^{b-1}\phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right)}{\int_{-\infty}^{\infty} p_i^{a-1}(1-p_i)^{b-1}\phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) dY_i},$$

where a and b are the parameter estimates that determine the shape of the distribution, p_i represents the p -value given effect size Y_i and variance σ_i^2 , τ^2 denotes the between-study variance component² that reflects true effect-size variability, Δ_i represents $X_i\beta$, and $\phi(z)$ denotes the standard normal probability density function evaluated at z . Δ_i , $\beta = (\beta_0, \beta_1, \dots, \beta_q)$ is a q -dimensional vector of unknown regression coefficients and $X_i = (X_{i1}, \dots, X_{iq})'$ is a vector of known predictors. Multiplying X_i and β creates a function of linear predictors, which allows one to estimate models with varying numbers of predictors (i.e., different models). When X_i is simply a vector of 1s, only the intercept is estimated, so that Δ_i represents a mean-only model.

Notice that the weight function in the beta density weight-function model is represented by the beta density without including the Beta function $B(a, b)$. $\frac{1}{B(a, b)}$ is a renormalizing constant necessary for the beta density function to integrate to one. Because I have already included a renormalizing constant in the model (the integral in the denominator of the function), I do not need to include a second renormalizing constant.

Recall that a reason for using the beta density as the weight function is to allow for a formal comparison of the effect-size model with and without the inclusion of the selection model (i.e., to test for bias). The comparison can be completed via a simple likelihood-ratio test because the unadjusted model (effect-size model *without* selection) is nested within the adjusted model (effect-size model *with* selection). When both the effect-size and selection model parameters are estimated, the estimates represent adjusted parameter estimates that account for publication bias. Setting parameters a and b to 1.0 sets the selection weight function equal to 1.0. Multiplying the effect-size model by a weight of 1.0

²Note that when the variance component τ^2 is included, the model represents a random-effects analysis. Random-effects analyses are appropriate when one desires to generalize to a hypothetical population of studies. If the goal of the meta-analysis is to focus only on that particular set of studies, a fixed-effects analysis may be more appropriate (see Hedges & Vevea, 1998). If the latter is preferred, one does not include the variance component in the model (i.e., set τ^2 to zero when estimating a fixed-effects model).

is the same as estimating the effect-size model without selection (i.e., the standard meta-analytic model). Thus, the resulting estimates represent unadjusted parameter estimates that do not account for bias. (Note that, in practice, one can simply estimate the standard meta-analytic model. It is formally equivalent to estimating the weighted model with a and b set to 1.0, so the likelihood-ratio test is still valid.) In this way, one can estimate both the unadjusted and adjusted models in order to test for the presence of publication bias. If a significant difference between the two models is found, one can conclude that adding a weight function (with parameters a and b) to the effect-size model improves model fit considerably and, thus, bias is present. If a significant difference is not detected, then bias has not been found in the distribution of effects. Thus, in the latter scenario, the estimates from the effect-size model may be good approximations of the true parameter values.

A Solution for the Numerical Problems

Recall that Vevea and Hedges (1996) ran into numerical problems when estimating the effect-size model with selection (i.e., the adjusted model that accounts for bias). Specifically, they could not estimate parameters a and b simultaneously, forcing them to fix one parameter in order to estimate the other. The adjusted model that includes the selection model resulted in a scaled down version that did not work as well as they had hoped.

My investigation of the numerical difficulties revealed that they were related to the presence of vertical asymptotes in some forms of the beta density. Look back at Figures 3.2 and 3.3, and notice that the curves of the beta density plots do not stop. That is, the lines of the curve continue up towards infinity. This causes problems for the integration in the denominator of the model. Following common practice for working with improper integrals, I implemented a change of variables solution to shift the infinite range of integration to a finite one (Press, Teukolsky, Vetterling, & Flannery, 1992). However, the

solution did not prove to be useful. A solution to the estimation problem that was successful was to simply avoid the asymptotes by fixing the weight function. Thus, for effects that obtain p -values below 0.00001, I fix the weight for p at 0.00001, and for effects that obtain p -values above 0.99999, I fix the weight for p at 0.99999. Because the area in the regions I cut-off is so tiny, I lose virtually no information by implementing this solution.

ML Estimation

In order to estimate the parameters of the effect-size and selection models, I use maximum-likelihood (ML) estimation. The maximum likelihood estimate (MLE) of a parameter is the value of that parameter that is the most likely to have occurred, given the data. (Rice, 2007). Assuming that the effects are independent, the joint likelihood for the effects $\mathbf{Y} = (Y_1, \dots, Y_k)'$ is the product of the individual likelihoods of Equation 3.2:

$$(3.3) \quad l(a, b, \beta, \tau^2 | Y_i, \sigma_i^2) = \prod_{i=1}^k \left\{ \frac{p_i^{a-1} (1 - p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right)}{\int_{-\infty}^{\infty} p_i^{a-1} (1 - p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \right\},$$

where k represents the number of effects present in the meta-analysis. However, because maximizing the logarithm of the likelihood function is easier than maximizing the likelihood function itself, it is standard practice to maximize the log-likelihood function (Rice, 2007):

$$(3.4) \quad L(a, b, \beta, \tau^2 | Y_i, \sigma_i^2) = \sum_{i=1}^k \left\{ (a - 1) \log(p_i) + (b - 1) \log(1 - p_i) - \frac{1}{2} \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right)^2 - \log(A_i) \right\},$$

where A_i is the integral:

$$\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i.$$

The parameter vector $\zeta = (a, b, \beta)'$ may be estimated by solving the log-likelihood equation:

$$(3.5) \quad \frac{\partial L}{\partial \zeta} = 0.$$

The value of any single element may be obtained from:

$$(3.6) \quad \frac{\partial L}{\partial \zeta_l} = \sum_{i=1}^k \left\{ \frac{\partial R_i}{\partial \zeta_l} - \frac{\frac{\partial A_i}{\partial \zeta_l}}{A_i} \right\},$$

where $R_i = (a-1) \log(p_i) + (b-1) \log(1-p_i) - \frac{1}{2} \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right)^2$ and ζ_l denotes the parameter to be estimated. Note that the partial derivatives $\frac{\partial R_i}{\partial \zeta_l}$ and $\frac{\partial A_i}{\partial \zeta_l}$ depend on what type of parameter ζ_l represents. For example, making the appropriate substitutions, the derivative of parameter a is:

$$(3.7) \quad \frac{\partial L}{\partial a} = \sum_{i=1}^k \left\{ \log(p_i) - \frac{\int_{-\infty}^{\infty} p_i^{a-1} \log(p_i) (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \right\}.$$

However, rather than estimating the parameters one at a time, one can estimate all the parameters simultaneously using an optimization routine.

Optimization is the maximization or minimization of an objective function, or a function that measures the outcome of interest (e.g., the log-likelihood function in Equation 3.4), given a vector of parameters ζ . The goal is to obtain values of the parameters that optimize the objective function (Nocedal & Wright, 2006). In other words, the “best”

FIGURE 3.4. A plot of estimated log-likelihood values against a sequence of corresponding mean parameter μ values. The plot indicates that the best estimate of μ is about 0.6.

parameter value is one that produces an outcome that is larger (or smaller, if one is minimizing) at that parameter value than at any other value of that parameter. If the objective function's estimated values are plotted against a set of parameter values, imagine the peak of the curve representing the "best" parameter estimate. Figure 3.4 shows a plot of log-likelihood estimates plotted against a sequence of corresponding values for the mean parameter μ . The peak of the curve is roughly at $\mu = 0.6$. The value 0.6, then, represents the best estimate of the mean of that data set. Note that with multiple parameters, one would need to imagine such a peak occurring on a hypersurface, but the same principle of optimizing the parameter by maximizing the objective function still holds. Thus,

rather than estimating hundreds of log-likelihoods and plotting them against each set of parameter values, one can use a built-in optimization algorithm. The one I use is called *nlminb* in *R* (Version 2.9.2), a public domain statistical computing program; *nlminb* is a numerical optimizer that minimizes an arbitrary objective function. I use *nlminb* to obtain optimal parameter estimates by minimizing the negative of the log-likelihood function (Equation 3.4), which is equivalent to maximizing the log-likelihood function. Note that *nlminb* solves Equation 3.5 for each parameter and decides which parameter estimate is optimal by using numerical approximations of the derivatives. One can feed the analytical derivatives into *nlminb*, but when I did so for this model, it slowed down the estimation process.

The standard errors of the parameters may be calculated by estimating the covariance matrix of the multivariate sampling distribution. The covariance matrix is often estimated via the inverse of Information matrix $\mathbf{I}$, which is the negative of the expected value of the Hessian matrix (or matrix of second partial derivatives). However, obtaining $\mathbf{I}$ may not be a tractable problem for this model. An alternative approximation of the covariance matrix, recommended by Thisted (1988), is the negative inverse of the Hessian matrix. For the beta density weight-function model, the partial derivatives required to construct the Hessian matrix have the general form:

$$(3.8) \quad \frac{\partial^2 L}{\partial \zeta_l \partial \zeta_m} = \sum_{i=1}^k \left\{ \frac{\partial^2 R_i}{\partial \zeta_l \partial \zeta_m} + \frac{\frac{\partial A_i}{\partial \zeta_l} \times \frac{\partial A_i}{\partial \zeta_m}}{A_i^2} - \frac{\frac{\partial^2 A_i}{\partial \zeta_l \partial \zeta_m}}{A_i} \right\},$$

where $\frac{\partial^2 R_i}{\partial \zeta_l \partial \zeta_m}$ and $\frac{\partial^2 A_i}{\partial \zeta_l \partial \zeta_m}$ represent the second partial derivatives of R_i and A_i with respect to parameters ζ_l and ζ_m , and $\frac{\partial A_i}{\partial \zeta_l}$ and $\frac{\partial A_i}{\partial \zeta_m}$ denote the first partial derivatives of parameters ζ_l and ζ_m calculated in Equation 3.6. The Hessian matrix, then, is constructed by making the appropriate substitutions into Equation 3.8 and filling in the matrix for parameters a ,

b , and β :

$$\begin{bmatrix} \frac{\partial^2 L}{\partial a^2} & \frac{\partial^2 L}{\partial a \partial b} & \frac{\partial^2 L}{\partial a \partial \beta} \\ \frac{\partial^2 L}{\partial b \partial a} & \frac{\partial^2 L}{\partial b^2} & \frac{\partial^2 L}{\partial b \partial \beta} \\ \frac{\partial^2 L}{\partial \beta \partial a} & \frac{\partial^2 L}{\partial \beta \partial b} & \frac{\partial^2 L}{\partial \beta^2} \end{bmatrix}.$$

The elements associated with parameters a and b may be calculated by:

$$(3.9) \quad \frac{\partial^2 L}{\partial a \partial b} = \frac{\partial^2 L}{\partial b \partial a} = \sum_{i=1}^k \left\{ \frac{\int_{-\infty}^{\infty} p_i^{a-1} \log(p_i) (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \times \right. \\ \left. - \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} \log(1-p_i) p_i^{a-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} - \right. \\ \left. \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} \log(1-p_i) p_i^{a-1} \log(p_i) \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \right\}.$$

The elements associated with parameters a and β are:

$$(3.10) \quad \frac{\partial^2 L}{\partial a \partial \beta} = \frac{\partial^2 L}{\partial \beta \partial a} = \sum_{i=1}^k \left\{ \begin{aligned} & \frac{\int_{-\infty}^{\infty} p_i^{a-1} \log(p_i) (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \times \\ & \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} p_i^{a-1} \mathbf{X}_i \left(\frac{Y_i - \Delta_i}{\sigma_i^2 + \tau^2} \right) \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} - \\ & \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} \log(p_i) p_i^{a-1} \mathbf{X}_i \left(\frac{Y_i - \Delta_i}{\sigma_i^2 + \tau^2} \right) \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \end{aligned} \right\}.$$

The elements associated with parameters b and β are:

$$(3.11) \quad \frac{\partial^2 L}{\partial b \partial \beta} = \frac{\partial^2 L}{\partial \beta \partial b} = \sum_{i=1}^k \left\{ \frac{\int_{-\infty}^{\infty} p_i^{a-1} \log(1-p_i) (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \times \right. \\ \left. - \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} p_i^{a-1} \mathbf{X}_i \left(\frac{Y_i - \Delta_i}{\sigma_i^2 + \tau^2} \right) \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} - \right. \\ \left. \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} \log(1-p_i) p_i^{a-1} \mathbf{X}_i \left(\frac{Y_i - \Delta_i}{\sigma_i^2 + \tau^2} \right) \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \right\}.$$

When both parameters are either a , b , or β , the elements of the Hessian matrix are the following, respectively:

$$(3.12) \quad \frac{\partial^2 L}{\partial a^2} = \sum_{i=1}^k \left\{ \left(\frac{\int_{-\infty}^{\infty} p_i^{a-1} \log(p_i) (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \right)^2 - \right. \\ \left. \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} p_i^{a-1} \log(p_i)^2 \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi \left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}} \right) dY_i} \right\},$$

$$(3.13) \quad \frac{\partial^2 L}{\partial b^2} = \sum_{i=1}^k \left\{ \frac{\left(\frac{\int_{-\infty}^{\infty} p_i^{a-1} \log(1-p_i)(1-p_i)^{b-1} \phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) dY_i} \right)^2 - \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} p_i^{a-1} \log(1-p_i)^2 \phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) dY_i} \right\},$$

and

$$(3.14) \quad \frac{\partial^2 L}{\partial \beta^2} = \sum_{i=1}^k \left\{ -\frac{\mathbf{X}_i^2}{\sigma_i^2 + \tau^2} + \frac{\left(\frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} p_i^{a-1} \mathbf{X}_i \left(\frac{Y_i - \Delta_i}{\sigma_i^2 + \tau^2}\right) \phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) dY_i} \right)^2 - \frac{\int_{-\infty}^{\infty} (1-p_i)^{b-1} p_i^{a-1} \phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) \left[\left(\mathbf{X}_i \left(\frac{Y_i - \Delta_i}{\sigma_i^2 + \tau^2}\right) \right)^2 - \frac{\mathbf{X}_i^2}{\sigma_i^2 + \tau^2} \right] dY_i}{\int_{-\infty}^{\infty} p_i^{a-1} (1-p_i)^{b-1} \phi\left(\frac{Y_i - \Delta_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) dY_i} \right\}.$$

Once all the elements have been calculated and placed into the matrix, the negative of the Hessian matrix is inverted to produce the covariance matrix. Recall that generally the Information matrix $\mathbf{I}$ is used to obtain the covariance matrix. However, given the complex structure of the model, it is more convenient to use the negative inverse of the Hessian matrix to approximate the covariance matrix. The diagonal elements of the covariance matrix

represent the variances of each parameter. The square root of those elements represent the standard errors of the parameters (Thisted, 1988).

Bayesian Estimation

I also implement the beta density weight-function model using Bayesian estimation, as a second and more flexible way of estimating the model parameters. In ML estimation, one only considers the data in the analysis. In Bayesian estimation, one considers their prior beliefs (or expertise) in addition to what the data have to say. Bayesian estimation summarizes a distribution by deriving the posterior distribution for each parameter. Numerous analyses can then be performed on the posterior distributions. Note that this introduces a different kind of thinking about the data. For example, one can make probabilistic statements about the distribution of the data, which could not be done if working in the classical statistics framework (Lynch, 2007).

In Bayesian estimation, researchers often use sampling algorithms to simulate (or sample) values from the posterior distribution. Sampling is useful in a Bayesian context because sampling from the posterior distribution is considerably simpler than analytically determining the characteristics of the posterior. Sometimes, when a model is complex, an analytical approach to understanding the posterior (and hence the parameters of the model) will be impossible, whereas empirically estimating its characteristics using Monte Carlo methods is comparatively simple. There are a number of advantages afforded by adopting a Bayesian approach. For example, it allows prior knowledge to be incorporated in order to make estimation more efficient (e.g., preventing parameter estimates from wandering into an unreasonable space) or to represent strong subjective beliefs about parameters (Lynch, 2007). Imagine, for example, a data set that you believe (for whatever reason) is not biased. You can represent that belief by setting strong priors on a and b that keep the parameter estimates close to one. That will keep the estimates from diverging

into values that do not make sense as well as make estimation faster. Note that priors can be set on the regression coefficients and variance component as well.

Various sampling algorithms exist for simulating values from the posterior distribution. The one I implement is called the Metropolis-Hastings (MH) algorithm.³ The MH algorithm is a Markov chain Monte Carlo (MCMC) method that obtains samples from a probability distribution in order to use the sequence of samples to approximate the distribution. It generates samples from the posterior distribution using the full joint density function (e.g., Equation 3.4), allowing for use with multivariate distributions (Lynch, 2007). The algorithm consists of the following steps:

- (1) Set starting values for each parameter and set $i = 1$.
- (2) Generate a candidate parameter value θ^c from a proposal density $h(\cdot)$.
- (3) Compute the ratio $R = \frac{f(\theta^c)h(\theta^{i-1}|\theta^c)}{f(\theta^{i-1})h(\theta^c|\theta^{i-1})}$.
- (4) If $U(0, 1) < R$, then set $\theta^i = \theta^c$. Otherwise, set $\theta^i = \theta^{i-1}$.
- (5) Set $i = i + 1$ and return to Step 2.

In Step 1, establish starting values for the parameters. In Step 2, generate a “candidate” value for the parameter θ^c using a chosen proposal density $h(\cdot)$ that is easy to sample from (e.g., normal or uniform density). In Steps 3 and 4, accept or reject this value as a draw from the distribution of interest. The ratio R may be broken down into two parts. The first part $\left(\frac{f(\theta^c)}{f(\theta^{i-1})}\right)$, called the “importance ratio”, is the ratio of the unnormalized posterior density evaluated at the candidate parameter value θ^c to the posterior density evaluated at the previous parameter value θ^{i-1} (i.e., estimate the log-likelihood values at θ^c and θ^{i-1} using Equation 3.4, and compare them). The second part of the ratio $\left(\frac{h(\theta^{i-1}|\theta^c)}{h(\theta^c|\theta^{i-1})}\right)$ is the ratio of the proposal densities evaluated at the candidate and previous values. It adjusts for

³Note that I cannot use Gibbs sampling, a popular and generally more efficient sampling algorithm than Metropolis-Hastings. That is because the derivation of the conditional distributions for the parameters, which the algorithm requires for estimation, is not a tractable problem for this model (see, e.g., Casella & George, 1992; Lynch, 2007).

asymmetrical proposal densities, which could cause some candidate values to be selected more often than others.⁴ (Note that if the proposal density is symmetric, the ratio is equal to one.) R is then compared to a draw from the uniform distribution. If $R > U(0, 1)$, accept θ^c as a draw from the posterior density; otherwise retain the previous value of the parameter. Last, Step 5 allows one to keep count of how many candidate values have been accepted. The ideal acceptance rate is approximately 50%, with a range of 25 to 75% viewed as acceptable (Chib & Greenberg, 1995; Geyer, 1991; Lynch, 2007).

The acceptance rate is important because it plays a key roll in both the autocorrelation of the sampled values and convergence. Notice in Steps 3 and 4 of the algorithm that a candidate value is accepted when the new value is viewed as better than the last value. Thus, if the acceptance rate is high, many “bad” values are sampled and it may take the algorithm a long time to converge. However, if the acceptance rate is low (because many values are rejected), the autocorrelation of the sampled values (discussed next) is high. The acceptance rate is a pull between convergence and the autocorrelation of the sampled values. It is best to find a happy medium between the two in order to maximize efficiency.

One important difference between simple Monte Carlo and Markov chain Monte Carlo methods is independence. In the previous method, values are simulated independently. In the latter method, Markov chains are represented by stationary transition probabilities that allow movement from one value to the next. That creates dependence in the chain (or distribution) of sampled parameter values. The recommended solution is to gain “almost independence” by sub-sampling every g th value in the chain (Geyer, 1991). One way to decide on the g spacing (or chain thinning) is to calculate the autocorrelations of each chain of parameter values and choose a spacing that generates “low” autocorrelations (Lynch, 2007).

⁴Note that without this adjustment ratio, the algorithm would simply be the Metropolis algorithm. Hastings allowed for asymmetrical proposal densities to be used by including this ratio (Geyer, 1991).

Once sampling is completed, obtaining the parameter estimates of the effect-size model with and without selection is simple. First, diagnose each set of sampled parameter distributions for convergence using methods such as trace and cumulative sum plots in order to examine if the parameters have converged to stable values (see, e.g., Lynch, 2007 for convergence diagnostics). Remove the values of the parameters that are in the first non-converged part of the sampled distribution (called the “burn-in” period). Then keep every g th sampled value in order to reduce the autocorrelation within the chain of sampled values, as each estimate is dependent on the previous estimate. Last, simply run summary analyses on the remaining values. For example, the means and standard deviations of the parameters are the means and standard deviations of their sub-sampled posterior distributions. Note that if no (strong) priors are placed on the parameters, the modes of the posterior distributions of the parameters will be identical to the ML parameter estimates when the sample size is infinite. However, taking the mean of the posterior distribution is generally preferred because the mean uses the whole distribution to summarize the posterior, while the mode uses only one piece of information (the point at which the most frequent estimate occurs). The principle is similar in spirit to the common practice of favoring the mean to characterize central tendency in a sample when there are not compelling reasons to use a more resistant statistic. Note that the median may also be a viable choice of statistic to use to summarize the distribution, and has recently gained popularity. However, for reasons stated above, I choose to use the mean to summarize the posterior distribution.

When the beta density weight-function model is implemented via the MH algorithm, theoretically, one can perform the classic likelihood-ratio test in order to determine whether there are significant differences between the effect-size model with and without the inclusion of the selection model because the unadjusted model is nested within the adjusted model in this case as well. However, there are more flexible and appropriate

methods that exist in Bayesian statistics. That is, Bayesian statisticians would never actually compute the likelihood-ratio test as they do not think in terms of formal tests; instead, they compare models in order to determine which one fits the data better (Lynch, 2007), or directly calculate the probability that a parameter is in a particular range. The Bayesian Information Criterion (*BIC*) is a criterion for model selection that approximates the Bayes factor, a ratio of integrated likelihoods used for model comparison in Bayesian statistics that is usually too difficult to estimate when using MCMC estimation (Lynch, 2007). A *BIC* value is calculated for each model:

$$BIC = D + j \log(k) ,$$

where D is the deviance of the model, defined as minus two times the log-likelihood [i.e., $-2L(\text{model})$], j represents the number of parameters in the model, and k denotes the number of effects present in the meta-analysis. The model with the lower *BIC* value is the preferred, or better fitting, model (Congdon, 2003). Akaike's Information Criterion (*AIC*) is another criterion for model selection. It focuses on the goodness of fit of the estimated model and is not dependent on the number of effects present in the meta-analysis. An *AIC* value is calculated for each model:

$$AIC = D + 2j ,$$

where D is the deviance of the model (defined above) and j represents the number of parameters in the model. The better fitting model is the one with the lower *AIC* value (Congdon, 2003). For example, if the adjusted model obtains a lower *BIC* or *AIC* value, then that is the more appropriate model for that meta-analytic data set. That, then, suggests that bias is present, as the selection model is adjusting for bias.

CHAPTER 4

Maximum Likelihood Mean-Only Simulations

Often, complex models such as the beta density weight-function model may not be analytically tractable. In such cases, simulation is the only means by which a model's performance can be assessed. The strategy behind a Monte Carlo simulation is to sample values repeatedly according to an assumed probability model using a set of known parameter values (referred to as generating values). That creates a sample that resembles data drawn from the true population. The generating values represent the true population parameters. The parameters are then estimated using the model of interest (e.g., the beta density weight-function model) and the estimates are compared to the generating values according to a set of criteria in order to assess the model's performance (Mooney, 1997).

In this chapter, I implement a series of Monte Carlo simulations to assess the efficacy of the beta density weight-function model when it is implemented via maximum-likelihood (ML) estimation. I examine model inference and how well the model is working in order to illustrate the model's utility (see Mooney, 1997) on data sets that represent variations in the number of effects included in the meta-analysis (k), severity of publication bias, and heterogeneity. No predictors are included in this set of simulations (i.e., I implemented a mean-only model); however, I add a moderator in Chapter 5 to assess the model's ability to distinguish systematic study differences from publication bias.

The Simulation Design

In order to assess the method's efficacy on the kinds of meta-analytic data that are most likely to appear in the literature, I try to emulate reality as much as possible in both

the way I generate the effect sizes and the characteristics I manipulate in the design of the simulation. I describe both processes below.

Sample-Size and Effect-Size Generation

In order to obtain sample sizes that match reality, I conducted a search in PsycINFO using the terms “meta-analysis” in the title and “standardized mean difference” (SMD) in the abstract. I then looked through the meta-analytic articles in order to find articles that contained both effect-size (D_i) and sample-size (n) information for each study in the meta-analysis. I chose a set of 10 meta-analytic articles that contained the information I needed to create a sample-size distribution from which I could sample. By sampling n from actual studies, instead of creating an artificial positively-skewed distribution, I hope to match the sample sizes present in the real world of psychology.

Next, I generated effects that fit a mean-only model centered at a true population standardized difference between means μ of 0.20. I accomplished this by first estimating the variance of the data. I randomly sampled k sets of sample sizes and effects, and calculated their variances (σ_i^2 's) using the SMD variance formula:

$$\sigma_i^2 = \frac{n_{i1} + n_{i2}}{n_{i1} \times n_{i2}} + \frac{D_i^2}{2(n_{i1} + n_{i2})},$$

where n_{i1} and n_{i2} represent the two group's sample sizes and D_i denotes the effect size from study i . In order to generate observed effects (Y_i), I sampled Y_i for each study from a normal distribution with mean equal to 0.20 and variance equal to the SMD sampling variance calculated plus a variance component. The choice of SMD for the simulated data is due to convenience as it appeared to be the most likely to include sample-size information. Also, SMD data appears very often in the field of psychology; thus, I again hope to emulate the real world of psychology in the simulation. I chose a true SMD of 0.20 for the same reason. I examined the distribution of effects and found that they clustered around an SMD of about 0.20; thus, it made sense to simulate data centered at that value.

Conditions Examined

In order to examine how different kinds of meta-analytic data characteristics affect the model's inference and estimation, I manipulated various design aspects within the simulation. While I would have liked to examine every data characteristic that exists, as that is not feasible, I limited myself to three key data characteristics: number of effects included in the meta-analysis, severity of publication bias, and heterogeneity.

First, I varied the magnitude of the variance component τ^2 . With the average conditional variance in the entire population equal to 0.14, I calculated a third of that to look at a small τ^2 and two-thirds to look at a medium τ^2 (see Hedges & Pigott, 2001). That created τ^2 's of 0.05 and 0.09, respectively (while 0.14 represents a large τ^2). I also included a τ^2 of 0, which represents the case where there is no heterogeneity in the population. The previous three τ^2 's represent random-effects distributions for which three different levels of heterogeneity are present.

I varied the number of effects (k) in order to see how performance in the model changes as the number of effects included in the meta-analysis decreases. Specifically, I considered $k = 30, 90$, and 270. While $k = 30$ may not seem like a small number of effects, it is so for practical reasons. It is sometimes hard to estimate more than one or two parameters with fewer effects than that. It may be the case that for some small data sets, it is easy to estimate the parameters, but for others it is not. Thus, doing a simulation with k less than 30 is not practical. That is, the $k = 30$ limit is imposed by the simulation; it is not necessarily a limit of the model. Also, recall that I hope the model performs better than Vevea and Hedges' (1995) weight-function model. The minimum recommended k for their model is approximately 100 effects. If the beta density weight-function model can perform well with 30 effects included in the analysis, it is already a huge improvement.

Last, I varied the severity of publication bias. I simulated biased data using Vevea and Hedges' (1995) weight-function model. Their step-function imitates hypothesis testing in

fields like psychology by taking theoretically important p -value intervals into account (e.g., $p < 0.05$; see, e.g., Greenwald, 1975; Nelson et al., 1986; Rosenthal & Gaito, 1963, 1964); thus, it is useful in generating biased data with varying degrees of selection. I simulated biased data by defining sets of weights (i.e., weight functions) that represent varying degrees of biased selection. The weights (ω s) represent the relative likelihoods that effects with particular one-tailed p -values will be observed (Vevea & Hedges, 1995). Thus, here, they are used to define the probabilities that effect-size estimates with their specific p -values will be observed within the simulated data set.

For example, in order to simulate non-biased data, I defined the following weight function:

$$w(p_i) = \begin{cases} \omega_1 = 1, & \text{if } 0 \leq p_i < 0.01, \\ \omega_2 = 1, & \text{if } 0.01 \leq p_i < 0.05, \\ \omega_3 = 1, & \text{if } 0.05 \leq p_i < 0.10, \\ \omega_4 = 1, & \text{if } 0.10 \leq p_i < 0.20, \\ \omega_5 = 1, & \text{if } 0.20 \leq p_i < 0.30, \\ \omega_6 = 1, & \text{if } 0.30 \leq p_i < 0.50, \\ \omega_7 = 1, & \text{if } 0.50 \leq p_i < 0.70, \\ \omega_8 = 1, & \text{if } 0.70 \leq p_i < 0.85, \\ \omega_9 = 1, & \text{if } 0.85 \leq p_i < 1.00. \end{cases}$$

Notice that all the weights are one, as that represents no selection. Thus, all effects are equally likely to be observed. I defined three more sets of weights in order to simulate data with small, moderate, and severe degrees of biased selection. After some investigation into what small, moderate, and severe bias may look like, I defined the weights for the above corresponding p -value intervals (from top to bottom) as the following:

- Small bias: ω 's = 1.0, 0.80, 0.60, 0.50, 0.40, 0.40, 0.30, 0.30, 0.20;
- Moderate bias: ω 's = 1.0, 0.70, 0.60, 0.50, 0.40, 0.30, 0.30, 0.20, 0.10;
- Severe bias: ω 's = 1.0, 0.40, 0.30, 0.30, 0.20, 0.20, 0.10, 0.10, 0.10.

The choice for nine p -value intervals was somewhat arbitrary. I wanted to include p -value intervals that made theoretical sense, as well as cover all p -values within the weight function. Nine intervals is simply what provided good p -value coverage and helped simulate the kinds of biased data I wanted. Moreover, note that I focused exclusively on one-tailed selection patterns because I wanted to allocate my computer resources to the more prevalent pattern of selection, as two-tailed selection is believed to be much less common.

Data Simulation

The simulation design may be thought of as a combination of three factors (or data characteristics):

- Heterogeneity: $\tau^2 = 0, 0.05, 0.09, \text{ or } 0.14$;
- Number of effects included in the meta-analysis: $k = 30, 90, \text{ or } 270$;
- Bias severity: none, small, moderate, or severe.

The combination of the factors ($4 \tau^2$'s $\times 3 k$'s $\times 4$ bias levels) resulted in 48 cells. For most cells, I ran 10,000 replications of the simulation. For cells where $k = 30$, I ran 15,000 replications to ensure convergence in the conditions under which the model may have the most trouble estimating the parameters (i.e., small k). The simulations were run in *R* (Version 2.7.1) on a Sun machine using the Solaris 10 operating system.

A single replication of the simulation consisted of estimating the parameters of the unadjusted and adjusted models (i.e., the effect-size model with and without the inclusion of the beta density selection model), calculating the standard errors of all the parameters, and performing the likelihood-ratio test on the two models. I used the built-in optimization algorithm *nlminb* in *R* (Version 2.7.1) to obtain the parameter estimates and log-likelihood values. I minimized the negative of the beta density weight-function model's

log-likelihood function (Equation 3.4) to estimate the adjusted model. To estimate the unadjusted model, I minimized the same log-likelihood function (Equation 3.4) after first fixing parameters a and b to 1.0. However, recall that one could simply estimate the standard meta-analytic model, as it is formally equivalent to estimating the weighted model with a and b set to 1.0.

Criteria For Model Assessment

After estimating the models, I investigated power and Type I errors in order to examine inference about publication bias (see Mooney, 1997 for criteria for model assessment). I conducted likelihood-ratio tests using the log-likelihood values from the unadjusted and adjusted models. Minus two times the difference between the log-likelihood values of the two models [i.e., $-2(L(\text{adjusted}) - L(\text{unadjusted}))$] approximates the chi-square (χ^2) distribution with degrees of freedom (df) equal to the number of parameters constrained in the simpler (or unadjusted) model (see, e.g., Rice, 2007). In this case, df will always equal two because parameters a and b are both constrained to 1.0 in the unadjusted model. I conducted those tests at alpha (α) levels of 0.01, 0.05, and 0.10. I then examined the rejection rates of those tests. In particular, I examined Type I errors (or the probability of improperly rejecting a true null hypothesis) when bias was not present and power (or the probability of correctly rejecting a false null hypothesis) when bias was present. The hope is that when bias is not present, the null hypothesis that bias is not present is rejected at the nominal rates of 0.01, 0.05, and 0.10, respectively. When bias is present, the hope is that the test rejects the null hypothesis as frequently as possible (up to 100% of the time) because the more frequent the rejection, the more powerful the test.

I was also interested in the coverage rates of 95% confidence intervals (CIs) for the parameters.¹ I calculated the standard errors for the unadjusted and adjusted models'

¹Note that I did not calculate the coverage rates of 95% CIs for the variance component. CIs are rarely done for the variance component as it is rarely the parameter of primary interest. Moreover, the actual distribution for the variance component is chi-square χ^2 . Creating CIs based on the normal distribution would produce optimal results only when the number of effects included in the analysis increases because

parameters by estimating the covariance matrix of the multivariate sampling distribution and taking the square root of the diagonal elements of the covariance matrix (see Thisted, 1988). I used those standard errors to calculate 95% CIs for the mean effect ($\hat{\mu}$) and the selection model parameters ($\hat{a}$ and $\hat{b}$) in order to assess how well they conform to their nominal coverage rates. Because I know the true population mean ($\mu = 0.20$), I can examine coverage rates for $\hat{\mu}$ for all 48 cells of the simulation for both the unadjusted and adjusted models. However, for $\hat{a}$ and $\hat{b}$, I only know the true values when bias is not present, as that is when a and b should both be equal to 1.0. The hope is that all coverage rates are equal to about 0.95.

In order to determine how well estimation is working, I examined bias and root mean squared errors (Mooney, 1997). Bias is an average estimate of whether the parameter estimates match the true values across the 10,000 or 15,000 replications, representing the accuracy of the estimates. It is calculated by taking the average of the estimated parameters after subtracting their true values (i.e., the generating values). Root mean squared error (RMSE) is then the square root of the average squared estimation error. Mean squared error (MSE) is the combination of systematic bias and sampling variability. Given that combination, the square root of MSE denotes the average proximity of the estimates to the true values. It represents the precision of the estimates, indicating how closely the estimates cluster together near the true value. Because, to many researchers, precision is considered more important than accuracy, RMSE represents the most useful index of the overall quality of estimation. The hope is that bias and RMSE are lower (or closer to zero) for the adjusted model than for the unadjusted model. However, note that when

the χ^2 distribution approaches the normal distribution as the number of effects move towards infinity. It would be ideal to create CIs for the variance component based on the χ^2 distribution, but that is currently not feasible. Some researchers have figured out how to create CIs for the variance component in the context of meta-analysis (e.g., Biggerstaff & Tweedie, 1997), but those methods are not generalizable to the beta density weight-function model.

publication bias is not present, bias and RMSE cannot be lower for the adjusted model, as the standard meta-analytic model provides unbiased minimum variance estimates.

Convergence and Data Cleaning

Before analyzing the results, I investigated whether the estimation procedures had converged to stable solutions for all 10,000 or 15,000 replications across the 48 simulations (or cells). I looked at parameter convergence diagnostics within *nlminb* to examine whether the optimizer converged. For the unadjusted model, I observed 100% convergence. For the adjusted model, I observed instances of non-convergence ranging from 19 to 239 cases out of the 10,000 or 15,000 replications per cell. That resulted in non-convergence rates of 0.14 to 2.39%. I compared the distributions of the converged estimates with the estimates from non-converged replications for a number of the cells, using summary statistics. Because I did not find any differences between the two distributions, suggesting that non-converged estimation often appears to have converged, despite *nlminb*'s diagnostic, I analyzed the results on both the converged and non-converged data together (i.e., I did cut the non-converged data out of my final data set) in order to include as much information as possible.

I also assessed whether the simulations themselves had reached stable estimates over the 10,000 or 15,000 replications. I checked the stability of the inferential statistics (i.e., the likelihood-ratio test) and the parameters using cumulative sum plots in order to see if the estimates reached stable solutions (i.e., stopped moving up and down to different values). The estimates for all 48 cells stabilized within the 10,000 or 15,000 replications (and in most cases much more quickly), suggesting that the estimates were estimated accurately.

In addition to convergence, I examined the data for replications where the mean estimates ($\hat{\mu}$'s) wandered into a nonsensical sample space. I defined a nonsensical $\hat{\mu}$ as a value that is larger or smaller than its data set's maximum or minimum Y_i value (or

simulated effect-size estimate), respectively. Because I do not want such estimates to taint my findings, I cut the data sets with nonsensical $\hat{\mu}$ estimates out of my final data set. Note that this is what a meta-analyst would typically do if such a problem were encountered. I also found instances where the standard errors ran into estimation issues, resulting in “NAs” that denote “Not Available” (or missing) values. I cut those data sets as well. (Note that there was some overlap in the data sets where $\hat{\mu}$ was a nonsensical value and where the standard errors were not estimable.) As a result, I had to cut data out of each of the 48 cells ranging from 3 to 2,714 replications cut per cell. That resulted in 0.03 to 18.09% replications cut. However, only 3.30% of the replications per cell were cut on average (with a median percent cut of 0.42 replications per cell). The vast majority (and extreme number) of the replications were cut from cells where $k = 30$, which makes sense as that is where the burden of estimation is greatest. Note that most of the data was cut due to nonsensical $\hat{\mu}$ estimates, with only a small portion cut due to the estimation problems of the standard errors.

Results

A large interest in assessing the performance of the beta density weight-function model has to do with inference about publication bias. It is important to examine whether the model is able to adjust for publication bias in such a way as to avoid Type I and Type II errors. Table 4.1 presents the results of the rejection rates for the likelihood-ratio test for publication bias at three different alpha levels ($\alpha = 0.01, 0.05$, and 0.10). When bias is not present, the rejection rates represent Type I errors (or the probabilities of concluding that publication bias is present when it is actually not). When bias is present, the rejection rates represent the model’s ability to avoid Type II errors (or power to detect publication bias when it exists in the population). I first looked at the no bias cells: the Type I errors waver just below and above their nominal rates. In particular, when the number of effects

TABLE 4.1. Rejection Rates for the Likelihood-Ratio Test for Publication Bias

$\alpha = 0.01$					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	0.007	0.007	0.007	0.009
	90	0.010	0.014	0.015	0.013
	270	0.009	0.013	0.013	0.013
Small Bias	30	0.032	0.013	0.014	0.018
	90	0.152	0.033	0.030	0.025
	270	0.646	0.041	0.022	0.038
Moderate Bias	30	0.063	0.019	0.017	0.017
	90	0.329	0.062	0.030	0.021
	270	0.920	0.118	0.025	0.014
Severe Bias	30	0.100	0.045	0.050	0.056
	90	0.538	0.121	0.086	0.091
	270	0.991	0.230	0.195	0.312
$\alpha = 0.05$					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	0.040	0.037	0.042	0.049
	90	0.052	0.066	0.072	0.059
	270	0.044	0.065	0.056	0.060
Small Bias	30	0.119	0.065	0.074	0.081
	90	0.349	0.113	0.089	0.089
	270	0.842	0.116	0.078	0.125
Moderate Bias	30	0.195	0.081	0.078	0.076
	90	0.576	0.170	0.103	0.081
	270	0.977	0.287	0.096	0.066
Severe Bias	30	0.269	0.151	0.149	0.152
	90	0.765	0.256	0.198	0.219
	270	0.998	0.422	0.378	0.526

TABLE 4.1. (Continued)

Model	k	$\alpha = 0.10$			
		τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	0.085	0.081	0.088	0.105
	90	0.104	0.131	0.135	0.121
	270	0.091	0.124	0.109	0.116
Small Bias	30	0.205	0.131	0.140	0.153
	90	0.474	0.188	0.148	0.156
	270	0.904	0.179	0.135	0.210
Moderate Bias	30	0.299	0.153	0.143	0.145
	90	0.703	0.262	0.169	0.142
	270	0.990	0.402	0.165	0.119
Severe Bias	30	0.397	0.239	0.228	0.233
	90	0.857	0.343	0.284	0.316
	270	0.999	0.536	0.494	0.645

in the meta-analysis is small ($k = 30$) and the variance component is less than 0.14, the Type I errors are below their nominal rates across all alpha levels. With a moderate number of effects ($k = 90$), the Type I errors increase to just above their nominal levels. But when the number of effects increases more ($k = 270$), the Type I errors decrease and approach their nominal rates. As the variance component increases, the Type I errors do not have a very consistent pattern, although they generally increase (rather than decrease). Still, the results show that the model is not likely to allow one to conclude that publication bias is present when it is actually not, as all the Type I errors are very close to their nominal values. Their slight deviations most likely represent sampling uncertainty in the simulation.

Now looking at the cells where bias is present, it is not surprising to see that power is highest when there is a large number of effects ($k = 270$). In those cases, when the variance component is 0, power to detect publication bias is above 0.99 across all alpha levels. However, when heterogeneity is introduced (i.e., the variance component changes

from 0 to 0.05), power decreases drastically. Once heterogeneity is present, there is little variation in terms of power across the variance component levels. As the severity of bias increases, power to detect publication bias increases. That is not surprising as bias becomes more noticeable. Power also increases as the alpha level becomes less stringent (i.e., increases) because the null hypothesis that bias is not present becomes easier to reject. As the number of effects decreases, power decreases drastically. When the number of effects is small ($k = 30$), power to detect bias is between 0.01 and 0.40, depending on the alpha level, heterogeneity, and amount of bias present.

Next, I examined the coverage rates of 95% CIs for the parameters of the unadjusted and adjusted models. Table 4.2 presents the results of coverage rates for $\hat{\mu}$ for the unadjusted and adjusted models and Table 4.3 presents the results of coverage rates for $\hat{a}$ and $\hat{b}$ for the adjusted model when bias is not present. (Recall that $\hat{a}$ and $\hat{b}$ are known only when bias is not present – both should equal 1.0. Also note that it does not make sense to look at coverage rates of 95% CIs for $\hat{a}$ and $\hat{b}$ for the unadjusted model because a and b are both set to 1.0; thus, the standard errors are 0.) Looking first at the results for $\hat{\mu}$ (see Table 4.2), when bias is not present and the variance component is 0, the unadjusted and adjusted models obtain coverage rates that are just barely above the nominal rate of 0.95. Once heterogeneity is present and as the variance component increases, the coverage rates increase from 0.93 to 0.95 for the unadjusted model and from 0.91 to 0.96 for the adjusted model. As the number of effects increases, the coverage rates again approach the nominal rate. Overall, the unadjusted model obtains only *slightly* better coverage rates (i.e., closer to the nominal rate of 0.95) than the adjusted model when bias is not present. The lengths of the CIs for the adjusted model are consistently longer than for the unadjusted model, the previous ranging from 0.17 to 1.83 and the latter ranging from 0.06 to 0.35. (The lengths of the CIs increase as the variance component grows in magnitude and decrease as the number of effects increases.) That is not surprising, and will always be

TABLE 4.2. Coverage Rates of 95% Confidence Intervals for the Parameter Estimate $\hat{\mu}$ for the Unadjusted and Adjusted Models

Unadjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	0.960	0.927	0.931	0.935
	90	0.955	0.947	0.944	0.947
	270	0.956	0.946	0.946	0.946
Small Bias	30	0.645	0.327	0.264	0.231
	90	0.201	0.018	0.008	0.004
	270	0.001	0.000	0.000	0.000
Moderate Bias	30	0.549	0.166	0.106	0.070
	90	0.107	0.001	0.000	0.000
	270	0.000	0.000	0.000	0.000
Severe Bias	30	0.299	0.046	0.028	0.019
	90	0.009	0.000	0.000	0.000
	270	0.000	0.000	0.000	0.000
Adjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	0.957	0.919	0.928	0.938
	90	0.952	0.908	0.918	0.941
	270	0.962	0.922	0.945	0.955
Small Bias	30	0.960	0.852	0.829	0.809
	90	0.958	0.856	0.780	0.697
	270	0.939	0.787	0.525	0.331
Moderate Bias	30	0.957	0.848	0.810	0.778
	90	0.956	0.863	0.777	0.676
	270	0.941	0.817	0.576	0.332
Severe Bias	30	0.965	0.834	0.789	0.747
	90	0.939	0.831	0.704	0.556
	270	0.873	0.728	0.344	0.125

true, because variability (and thus uncertainty) about a parameter is increased when there are more parameters to estimate in a model. On average, across the 12 cells, CIs are 65 to 84% longer for the adjusted model than for the unadjusted model. However, once any bias is introduced, the adjusted model obtains *much* better coverage rates than the unadjusted

TABLE 4.3. Coverage Rates of 95% Confidence Intervals for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model When Bias is Not Present

Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{a}$	30	0.958	0.896	0.886	0.886
	90	0.937	0.854	0.881	0.922
	270	0.953	0.911	0.941	0.949
$\hat{b}$	30	0.960	0.889	0.880	0.889
	90	0.950	0.862	0.881	0.905
	270	0.961	0.905	0.923	0.936

model. In fact, for the unadjusted model, the coverage rates range from 0.02 to 0.65 when the number of effects is small ($k = 30$). Once bias and the number of effects increase, the coverage rates quickly drop to 0. For the adjusted model, the coverage rates are closest to the nominal rate when bias and the variance component are small. The coverage rates decrease slightly when bias increases and decrease a lot when the variance component increases. When the number of effects increases, the coverage rates again decrease – only slightly when the variance component is small ($\tau^2 \leq .05$) and more so as the variance component increases.

Now looking at Table 4.3, I observe close to nominal coverage rates for $\hat{a}$ and $\hat{b}$ when the variance component is 0. Once heterogeneity is present and as the variance component increases, the coverage rates generally increase from 0.85 to 0.95. As the number of effects increases, the coverage rates again increase.

Finally, I examined the results of bias and root mean squared errors for all the parameters of the unadjusted and adjusted models in order to investigate how well parameter estimation is working. The estimates indicate deviations of the parameter estimates from their true values, representing accuracy and precision, respectively. The hope is that the estimates are closer to zero for the adjusted model than for the unadjusted model, which would indicate smaller deviations. Tables 4.4 to 4.6 present the results of bias and Tables

TABLE 4.4. $1000 \times$ Bias for the Parameter Estimate $\hat{\mu}$ for the Unadjusted and Adjusted Models

Unadjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	0.449	1.460	0.271	-1.600
	90	0.223	-0.349	0.149	-0.310
	270	-0.197	0.084	0.076	0.181
Small Bias	30	73.559	147.690	185.102	222.601
	90	69.191	147.021	184.228	223.700
	270	67.229	147.724	185.798	224.358
Moderate Bias	30	82.582	171.668	218.592	265.393
	90	77.864	171.584	219.705	267.138
	270	75.923	171.762	219.276	267.867
Severe Bias	30	108.447	222.872	279.034	333.749
	90	103.266	223.716	280.190	335.134
	270	102.284	224.514	280.844	335.350
Adjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	51.622	32.838	19.879	-12.947
	90	23.611	-37.675	-38.433	-35.768
	270	9.536	-23.316	-18.459	-11.083
Small Bias	30	40.479	124.575	163.590	193.735
	90	-1.898	89.794	154.608	211.152
	270	-11.359	96.799	179.412	237.967
Moderate Bias	30	39.987	127.974	169.238	214.800
	90	-2.926	76.816	143.437	210.948
	270	-10.803	86.302	171.873	248.364
Severe Bias	30	19.822	106.033	158.903	210.048
	90	-17.621	85.909	165.923	240.135
	270	-21.396	90.714	184.384	259.570

4.7 to 4.9 present the results of RMSE. Both sets of results are multiplied by 1,000 for ease of presentation. (Note that this multiplication exaggerates the differences between the estimates of the cells; thus, the tables must be read carefully. For example, a bias estimate of 3.52 is actually 0.00352, which is indeed close to the desired value of 0.) Looking

first at the results of bias for $\hat{\mu}$ (see Table 4.4), I observe that bias for the adjusted model is always lower than for the unadjusted model when any publication bias is present and higher when no publication bias is present. For both the unadjusted and adjusted models, in the cases where publication bias is present, bias increases as the variance component increases. When publication bias is not present, the pattern is not as clear (i.e., there is some fluctuation). For the unadjusted model, as the number of effects increases, bias virtually does not change. For the adjusted model, as the number of effects increases, bias generally decreases with little heterogeneity ($\tau^2 = 0.05$) and increases as more heterogeneity is present. When publication bias is not present, the pattern is again not very clear. In the cases where publication bias is present, bias increases as the severity of publication bias increases for the unadjusted model (except when $\tau^2 = 0$). For the adjusted model, the pattern is not as clear, as bias wavers up and down with increases in the severity of publication bias. Nonetheless, note that all the estimates are close to zero.

The pattern of results for $\hat{\tau}^2$ in the cells where publication bias is not present is much clearer (see Table 4.5). For both the unadjusted and adjusted models, bias mostly increases as the variance component increases (with wavering at $k = 270$) and decreases as the number of effects increases. In the cases where publication bias is present, bias increases as the variance component increases. For the adjusted model, bias barely fluctuates as the number of effects increases when heterogeneity is present and decreases when the variance component is 0. For the unadjusted model, bias always decreases as the number of effects increases. When heterogeneity is present, as the severity of publication bias increases, bias increases for the adjusted model and fluctuates for the unadjusted model. When the variance component is 0, bias decreases for both models. These patterns aside, overall, bias is worse for the adjusted model in almost all cases (except when any publication bias is present, $k > 30$, and $\tau^2 = 0$).

TABLE 4.5. $1000 \times$ Bias for the Parameter Estimate $\hat{\tau}^2$ for the Unadjusted and Adjusted Models

Unadjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	3.520	-4.163	-4.341	-4.886
	90	1.955	-1.403	-2.015	-2.624
	270	1.075	-0.674	-0.896	-0.818
Small Bias	30	3.469	-13.740	-23.128	-34.152
	90	2.229	-11.202	-19.666	-31.828
	270	1.612	-10.491	-18.958	-29.581
Moderate Bias	30	2.954	-20.962	-37.018	-56.383
	90	2.093	-18.839	-34.299	-52.866
	270	1.594	-18.150	-33.137	-51.769
Severe Bias	30	5.128	-12.440	-24.570	-40.719
	90	3.990	-11.341	-23.445	-37.615
	270	3.626	-10.704	-21.836	-36.368
Adjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	7.100	20.089	26.951	30.790
	90	4.112	17.555	20.616	17.745
	270	1.406	6.838	6.676	4.038
Small Bias	30	6.115	-7.515	-22.634	-42.201
	90	1.530	-10.208	-23.659	-45.830
	270	0.256	-11.731	-25.826	-48.006
Moderate Bias	30	5.737	-11.065	-28.893	-51.579
	90	1.699	-11.418	-26.664	-46.780
	270	0.279	-12.929	-27.445	-50.607
Severe Bias	30	5.711	-17.754	-38.021	-62.266
	90	0.784	-20.126	-37.386	-61.200
	270	0.064	-19.432	-36.040	-61.672

Looking at Table 4.6, I observe that bias is largest for $\hat{a}$ and $\hat{b}$ when no heterogeneity is present. When heterogeneity is present, bias decreases as the variance component increases for $\hat{b}$ (except when $k = 30$) and is more variable for $\hat{a}$. Also, bias decreases as the number of effects increases in all cases.

TABLE 4.6. $1000 \times$ Bias for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model When Bias is Not Present

Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{a}$	30	163.463	27.276	-0.889	-25.346
	90	79.315	-9.975	-15.945	-15.788
	270	31.040	-9.514	-6.706	-5.601
$\hat{b}$	30	148.645	-25.057	-46.972	-45.591
	90	66.150	49.328	21.637	-2.283
	270	22.867	30.090	11.950	-2.347

Now I move on to examining root mean squared errors, which represent the square roots of the average squared estimation errors and are the most useful indices of the overall quality of estimation. Again, I hope the estimates are lower (i.e., close to zero) for the adjusted model than for the unadjusted model. I first examine RMSE for $\hat{\mu}$ (see Table 4.7). I observe that RMSE is higher for the adjusted model only initially. The difference in RMSE between the unadjusted and adjusted models quickly diminishes as the number of effects and bias increase. In fact, once some bias is present and the number of effects is not small ($k > 30$), there is a crossover effect where the adjusted model now obtains lower RMSE than the unadjusted model. However, this crossover effect is slowed down by the increase of the variance component because as the variance component increases, RMSE increases. This crossover effect happens because, for the adjusted model, RMSE decreases as the number of effects and bias increase. But, for the unadjusted model, RMSE decreases as the number of effects increases and *increases* as bias increases.

The pattern of results for $\hat{\tau}^2$ is very similar to that of $\hat{\mu}$ (see Table 4.8). RMSE decreases as the number of effects increases and increases as the variance component increases for both the unadjusted and adjusted models. For the adjusted model, RMSE decreases as bias increases (except when $k = 270$ and $\tau^2 \geq 0.09$). For the unadjusted model, RMSE fluctuates up and down when bias increases. The crossover effect almost

TABLE 4.7. $1000 \times \text{RMSE}$ for the Parameter Estimate $\hat{\mu}$ for the Unadjusted and Adjusted Models

Unadjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	45.924	68.872	79.202	90.675
	90	26.116	38.965	46.015	51.795
	270	14.686	22.718	26.348	30.135
Small Bias	30	86.001	160.685	198.798	237.553
	90	73.748	151.389	188.889	228.594
	270	68.792	149.185	187.343	226.005
Moderate Bias	30	93.697	182.084	229.226	276.517
	90	81.911	175.152	223.209	270.749
	270	77.345	172.917	220.442	269.097
Severe Bias	30	118.152	231.043	287.532	342.714
	90	106.865	226.423	282.960	338.137
	270	103.609	225.389	281.765	336.353
Adjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	192.816	356.124	406.642	441.194
	90	136.013	292.334	293.309	281.062
	270	48.302	156.729	143.435	132.415
Small Bias	30	168.872	297.468	340.427	376.390
	90	77.827	207.403	250.451	289.392
	270	31.005	142.611	205.470	256.761
Moderate Bias	30	162.147	279.594	327.273	370.025
	90	77.599	201.024	248.180	302.709
	270	31.048	131.645	200.289	269.156
Severe Bias	30	135.903	235.010	281.418	330.947
	90	53.925	156.764	218.793	283.514
	270	31.548	118.805	199.295	269.890

does not occur in the results for $\hat{\tau}^2$; the unadjusted model almost always obtains lower RMSE than the adjusted model. The only time the adjusted model obtains lower RMSE is when some bias is present, the variance component is 0 and the number of effects is large ($k = 270$).

TABLE 4.8. $1000 \times \text{RMSE}$ for the Parameter Estimate $\hat{\tau}^2$ for the Unadjusted and Adjusted Models

Unadjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	9.124	33.050	46.306	60.220
	90	4.395	18.628	26.270	34.684
	270	2.212	10.623	14.889	20.135
Small Bias	30	8.325	31.854	47.293	65.258
	90	4.292	20.167	30.711	44.864
	270	2.583	14.258	23.432	34.903
Moderate Bias	30	7.103	32.272	51.318	74.485
	90	3.903	23.771	39.996	59.726
	270	2.450	19.920	35.188	54.171
Severe Bias	30	10.392	31.747	48.409	68.459
	90	6.164	20.268	33.466	49.838
	270	4.555	14.406	25.869	40.881
Adjusted Model					
Model	k	τ^2			
		0.000	0.050	0.090	0.140
No Bias	30	43.227	128.931	159.844	198.896
	90	14.819	81.163	100.726	111.196
	270	3.605	38.793	45.359	51.806
Small Bias	30	28.970	71.103	89.523	111.423
	90	6.994	37.060	50.062	70.074
	270	1.132	20.993	32.523	53.730
Moderate Bias	30	30.287	64.992	83.282	105.944
	90	9.055	36.705	50.196	72.172
	270	1.134	20.999	33.803	55.692
Severe Bias	30	30.612	52.117	69.425	95.617
	90	6.539	30.861	47.722	71.457
	270	0.432	23.225	38.934	64.169

Looking at Table 4.9, I observe a decrease in RMSE for $\hat{a}$ and $\hat{b}$ as the number of effects increases. When heterogeneity is introduced (i.e., the variance component changes

TABLE 4.9. $1000 \times \text{RMSE}$ for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model When Bias is Not Present

Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{a}$	30	387.463	347.758	332.709	311.389
	90	232.381	240.751	210.847	172.629
	270	112.264	132.605	101.773	79.507
$\hat{b}$	30	511.088	476.170	454.638	410.512
	90	301.025	403.763	332.806	266.439
	270	129.485	240.587	184.226	139.056

from 0 to 0.05) and the number of effects is not small ($k > 30$), RMSE increases. However, once heterogeneity is present, the RMSE decreases as the variance component increases.

Conclusions

Overall, the results of these simulations suggest that the beta density weight-function model is useful in adjusting the parameter estimates of realistic meta-analytic data sets for publication bias. Not surprisingly, this model detects bias most often (via the likelihood-ratio test) when the number of effects included in the meta-analysis is large and the variance component is small. However, when the amount of publication bias present is severe, the power to detect publication bias is still decent when a small number of effects are included in the analysis. The model also obtains Type I errors that are close to the nominal rates in all cells of the simulation, suggesting that the model will very rarely cause one to erroneously conclude that bias is present when it is actually not. Moreover, even if the model's test does not detect publication bias, the coverage rates show that the true parameter values are *much* more likely to be caught within the 95% confidence intervals of the adjusted model's parameter estimates (i.e., those that use the beta density as the weight function) than of the unadjusted model's parameter estimates (i.e., the effect-size

model estimates). That is true even when only a small number of effects is present in the analysis, with coverage rates ranging from 0.75 to 0.96 for the adjusted mean estimate and from 0.02 to 0.65 for the unadjusted mean estimate when any bias is present. Bias and root mean squared errors also suggest that the adjusted estimates are usually better than the unadjusted estimates, as bias and RMSE for the adjusted estimates indicate smaller deviations from the true values. (Note that this is not the case with variance component; however, the variance component is rarely the parameter of primary interest.) Bias is always lower for the adjusted mean estimate than for the unadjusted mean estimate when any bias is present. RMSE is lower for the adjusted mean estimate than for the unadjusted mean estimate mostly when the number of effects and bias are moderate to large/severe. When the number of effects is small, the RMSE is again lower for the adjusted mean estimate when the amount of publication bias present is severe and heterogeneity is moderate to large. All this suggests that including the beta density as the weight function that models publication bias helps produce adjusted parameter estimates that resemble the true population values in many kinds of data sets (even when the power to detect publication bias is low for the cases when a small number of effects are present in the meta-analysis).

CHAPTER 5

Maximum Likelihood Moderator Simulations

In this chapter, I investigate the performance of the beta density weight-function model when a moderator is included in the meta-analysis. In particular, I want to know how well the model can distinguish systematic study differences from publication bias. I create a biased population by simulating two unbiased populations with varying data characteristics and combining them into one population. I then examine model inference and how well the model is working in order to illustrate the model's utility in this particular scenario when the model is implemented via maximum-likelihood (ML) estimation (see Mooney, 1997). The hope is that the likelihood-ratio test will not detect bias, as none exists, and that the model will produce appropriate adjusted parameter estimates for both populations.

The Simulation Design

The design of this set of simulations is very similar to the design of Chapter 4's ML mean-only simulations. However, this set of simulations is much smaller as it includes only cases where bias is not present.

Sample-Size and Effect-Size Generation

The sampling distribution from which I sampled the effects (D_i 's) and sample sizes (n 's) for this simulation is the same distribution as the one created for the ML mean-only simulations in Chapter 4. I used that distribution to sample n 's that represent sample sizes present in the real world of psychology, as well as to generate observed effects (Y_i 's) for two populations centered at true SMDs of 0.1 and 0.8. I accomplished this by simulating

data for each population separately in the same way I simulated data in Chapter 4: 1) randomly sample k sets of sample sizes and effects, 2) calculate their SMD sampling variances σ_i^2 , and 3) generate observed effects Y_i by sampling Y_i for each study from a normal distribution with mean equal to the true population mean ($\mu_1 = 0.1$ or $\mu_2 = 0.8$) and variance equal to σ_i^2 plus a variance component.

In order to create one biased population from the two unbiased populations, I varied the characteristics of each population by defining the sample sizes that may be present in each population. First, I calculated the median of the all the sets of sample sizes combined (i.e., median of $ntot_i = n_{i1} + n_{i2}$), which resulted in a median ($ntot_{med}$) of 47. Next, I defined weight functions for each population as the probabilities that effect-size estimates, given their generating combined sample sizes ($ntot_i$'s), will be observed within the simulated data set. For the first population, I defined the following weight function:

$$w(ntot_i) = \begin{cases} 0.9, & \text{if } ntot_i > ntot_{med}, \\ 0.1, & \text{if } ntot_i < ntot_{med}, \end{cases}$$

so that effects with combined sample sizes above the median are much more likely to be observed. That resulted in the following data characteristics for Population 1: small Y_i 's ($\mu_1 = 0.1$), large n_i 's, and small σ_i^2 's. For the second population, I reversed the situation by defining the following weight function:

$$w(ntot_i) = \begin{cases} 0.9, & \text{if } ntot_i < ntot_{med}, \\ 0.1, & \text{if } ntot_i > ntot_{med}, \end{cases}$$

so that Population 2 may include large Y_i 's ($\mu_2 = 0.8$), small n_i 's, and large σ_i^2 's. I then combined the two populations to create one biased population. Figure 5.1 shows an example of this process working.

FIGURE 5.1. An example of an asymmetrical funnel plot comprising two populations with true means of $\mu_1 = 0.1$ and $\mu_2 = 0.8$.

Conditions Examined

Except for the presence of bias in the population, I manipulated the same design aspects within this set of simulations as in the ML mean-only simulations. I varied the magnitude of the variance component to look at three different levels of heterogeneity, as well as the case where heterogeneity is not present in the population. I again considered $\tau^2 = 0, 0.05, 0.09$, and 0.14 . I varied the number of effects in order to see how performance changes in the model as fewer effects are included in the meta-analysis. I again considered $k = 30, 90$, and 270 .

Data Simulation

The simulation design is based on the combination of two data characteristics: the variance component τ^2 and the number of effects included in the meta-analysis k . With four levels of τ^2 and 3 levels of k , the result is 12 cells. For each cell, I ran 15,000 replications of the simulation. The simulations were run in *R* (Version 2.7.1) on a Sun machine using the Solaris 10 operating system.

Once again, a single replication of the simulation consisted of estimating the parameters of the unadjusted and adjusted models (i.e., the effect-size model with and without the inclusion of the beta density selection model), calculating the standard errors of all the parameters, and performing the likelihood-ratio test on the two models. However, recall that this time there are more parameters and standard errors to estimate as a moderator is included within the model. Specifically, a slope must now be estimated in order to include the difference in Populations 1 and 2 as a systematic study difference in the model. To estimate the slope β_1 , I first dummy coded Population 1 as 0 and Population 2 as 1. That resulted in a design vector $\mathbf{X}_i$ consisting of a column of 1's for the intercept followed by a column containing the predictor x_1 . The first half of the column containing the predictor is a series of 0's, representing Population 1, and the second half is a series of 1's, representing Population 2. The vector of unknown regression coefficients β contains the intercept and slope, β_0 and β_1 . Multiplying $\mathbf{X}_i$ and β creates the conditional mean effect Δ_i necessary to estimate a model with an intercept and one predictor. That is the effect-size model I estimate in this set of simulations (with and without the inclusion of the beta density weight function).

Accordingly, I used the following formula in order to obtain the conditional mean of each population:

$$\mu = \beta_0 + \beta_1 x_1 ,$$

where β_0 represents the intercept, β_1 represents the slope, and x_1 denotes the dummy coded predictor for Populations 1 and 2. Because the dummy code for Population 1 is 0, the conditional mean of Population 1 is simply the intercept (i.e., $\mu_1 = \beta_0 + \beta_1(0) = \beta_0$). The conditional mean of Population 2 is: $\mu_2 = \beta_0 + \beta_1(1)$. In order to calculate the standard errors of the conditional means, I used the following formula along with the information provided from the covariance matrix of the multivariate sampling distribution:

$$\sigma = \sqrt{\sigma_{\beta_0}^2 + x_1^2 \sigma_{\beta_1}^2 + 2x_1 \sigma_{\beta_0, \beta_1}},$$

where $\sigma_{\beta_0}^2$ denotes the variance of the intercept, $\sigma_{\beta_1}^2$ denotes the variance of the slope, x_1 represents the predictor (dummy coded 0 and 1), and $\sigma_{\beta_0, \beta_1}$ denotes the covariance of the intercept and slope (see Rice, 2007).

I used the built-in optimization algorithm *nlminb* in *R* (Version 2.7.1) to obtain the parameter estimates and log-likelihood values. I minimized the negative of the beta density weight-function model's log-likelihood function (Equation 3.4) to estimate the adjusted model. After fixing parameters a and b to 1.0, I minimized the same log-likelihood function (Equation 3.4) to estimate the unadjusted model.

Criteria For Model Assessment

In order to assess the efficacy of the beta density weight-function model when funnel plot asymmetry is actually due to a linear predictor and not publication bias, I examined model inference and how well estimation is working in the same way I did for the ML mean-only simulations. That is, I investigated Type I errors in order to examine inference about publication bias; I examined coverage rates of 95% confidence intervals (CIs) for the parameters¹; and I determined how well estimation is working according to bias and root mean squared errors (RMSE). Notice, however, that this time I am not interested

¹Note that I did not calculate the coverage rates of 95% CIs for the variance component. Refer to the section *Criteria For Model Assessment* in Chapter 4 for an explanation.

in power, as publication bias does not exist in this set of simulations. Instead, I look only at Type I errors, hoping that the likelihood-ratio test rejects the null hypothesis that publication bias is not present at the nominal rates of 0.01, 0.05, and 0.10, respective of the alpha levels the test is conducted at ($\alpha = 0.01, 0.05$, and 0.10).

Convergence and Data Cleaning

Before analyzing the results, I investigated the rates at which the estimation procedures had converged to stable solutions for all 15,000 replications across the 12 simulations (or cells). According to the parameter convergence diagnostic in *nlminb*, which assesses convergence of the optimizer, the unadjusted model obtained 100% convergence. For the adjusted model, I observed instances of non-convergence ranging from 30 to 986 cases out of the 15,000 replications per cell. That equated to non-convergence rates of 0.20 to 6.57%. I compared the distributions of the converged and non-converged estimates for a number of the cells using summary statistics. Because I did not find any differences between the two distributions, suggesting that non-converged estimation often appears to have converged, despite *nlminb*'s diagnostic, I did not cut the non-converged data out of my final data set in order to include as much information as possible.

I assessed whether the simulations had reached a stable result by looking at cumulative sum plots of the inferential statistics (i.e., the likelihood-ratio test) and the parameters. I wanted to see if the estimates stabilized (i.e., stopped moving up and down) over the 15,000 replications. The estimates appear to have stabilized within the 15,000 replications for all 12 cells, suggesting that the estimates were estimated accurately.

I then examined the data for replications where the conditional mean estimates ($\hat{\mu}_1$ and $\hat{\mu}_2$) wandered into a nonsensical sample space. I defined a nonsensical $\hat{\mu}_1$ or $\hat{\mu}_2$ as a value that is larger or smaller than its data set's maximum or minimum Y_i value (or simulated effect), respectively. I cut the data sets with nonsensical $\hat{\mu}_1$ or $\hat{\mu}_2$ estimates out of my final data set in order to keep such estimates from tainting my findings. I

also cut data sets which contained “NAs” (or “Not Available” values) for their standard error estimates due to estimations problems. As a result, I had to cut data out of each of the 12 cells ranging from 0 to 2,603 replications cut per cells. That resulted in 0 to 17.35% replications cut. However, only 5.69% of the replications were cut on average (with a median percent cut of 1.53 replications per cell). Not surprisingly, the vast majority (and extreme number) of the replications were cut from cells where $k = 30$ because that is where the burden of estimation is greatest. (That is true even more so in this set of simulations because there are more parameters to estimate in these simulations than in the ML mean-only simulations.) Note that most of the data were cut due to nonsensical $\hat{\mu}_1$ and $\hat{\mu}_2$ estimates, with only a small portion cut due to the estimation problems of the standard errors.

Results

I was first interested in assessing the beta density weight-function model’s performance in terms of inference about publication bias. Specifically, I want to know whether the model is able to distinguish systematic study differences (i.e., the predictor) from publication bias by avoiding Type I errors (or the probabilities of concluding the publication bias is present when it is actual not). Table 5.1 presents the results of the Type I errors for the likelihood-ratio test for publication bias at three different alpha levels ($\alpha = 0.01, 0.05$, and 0.10). The Type I errors vacillate just below and above their nominal rates. When the number of effects in the meta-analysis and the variance component are small ($k = 30$ and $\tau^2 \leq 0.05$), the Type errors are below their nominal rates across all alpha levels. As the variance component increases, the Type I errors vary but generally increase to rates just above the nominal rates. But when the number of effects increases, the Type I errors approach their nominal rates. Overall, the results show that the model does not find bias

TABLE 5.1. Type I Errors for the Likelihood-Ratio Test for Publication Bias

$\alpha = 0.01$				
τ^2				
k	0.000	0.050	0.090	0.140
30	0.003	0.008	0.011	0.013
90	0.009	0.011	0.014	0.014
270	0.009	0.014	0.014	0.011

$\alpha = 0.05$				
τ^2				
k	0.000	0.050	0.090	0.140
30	0.025	0.045	0.058	0.065
90	0.048	0.062	0.064	0.060
270	0.048	0.061	0.059	0.052

$\alpha = 0.10$				
τ^2				
k	0.000	0.050	0.090	0.140
30	0.063	0.096	0.118	0.128
90	0.097	0.123	0.124	0.114
270	0.095	0.113	0.110	0.104

often, as all the Type I errors are very close to their nominal values. Thus, the model is not likely to allow one to conclude that publication bias is present when it is actually not.

Next, I examined coverage rates of 95% CIs for the parameters of the unadjusted and adjusted models. Table 5.2 presents the results of coverage rates for $\hat{\mu}_1$ and $\hat{\mu}_2$ for the unadjusted and adjusted models and Table 5.3 presents the results of coverage rates for $\hat{a}$ and $\hat{b}$ for the adjusted model. Note that I focus the results on the conditional means $\hat{\mu}_1$ and $\hat{\mu}_2$, rather than the regression estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ from which the means are calculated. I do that in order to keep the focus on the parameters of interest – the mean effects of Populations 1 and 2. Also, recall that because a and b are both fixed to 1.0 in the unadjusted model, it does not make sense to look at coverage rates of 95% CIs for $\hat{a}$ and $\hat{b}$ in this case because the standard errors are 0. I first look at the results for $\hat{\mu}_1$ and $\hat{\mu}_2$ (see Table 5.2): when the variance component is 0, the unadjusted and adjusted models obtain

TABLE 5.2. Coverage Rates of 95% Confidence Intervals for the Parameter Estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ for the Unadjusted and Adjusted Models

Unadjusted Model					
Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{\mu}_1$	30	0.957	0.912	0.919	0.924
	90	0.957	0.944	0.941	0.941
	270	0.955	0.946	0.948	0.950
$\hat{\mu}_2$	30	0.953	0.933	0.937	0.936
	90	0.952	0.944	0.944	0.944
	270	0.953	0.949	0.951	0.951
Adjusted Model					
Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{\mu}_1$	30	0.978	0.912	0.914	0.922
	90	0.963	0.924	0.936	0.951
	270	0.967	0.939	0.953	0.961
$\hat{\mu}_2$	30	0.956	0.881	0.870	0.881
	90	0.951	0.895	0.912	0.931
	270	0.957	0.937	0.948	0.953

coverage rates that are just barely above the nominal rate of 0.95. Once heterogeneity is present (i.e., $\tau^2 > 0$) and as the variance component increases, the coverage rates increase from 0.91 to 0.95 for the unadjusted model and from 0.87 to 0.96 for the adjusted model. As the number of effects increases, the coverage rates again approach the nominal rate. Overall, the unadjusted model obtains only *slightly* better coverage rates (i.e., closer to the nominal rate of 0.95) than the adjusted model. However, the lengths of the CIs are consistently longer for the adjusted model than for the unadjusted model, the previous ranging from 0.13 to 1.59 and the latter ranging from 0.07 to 0.54. (The lengths of the CIs increase as the variance component grows in magnitude and decrease as the number of effects increases.) Recall that this will always be the case, as more uncertainty about the parameters (i.e., variability) is present when there are more parameter to estimate in a model. On average, across the cells of both $\hat{\mu}_1$ and $\hat{\mu}_2$, CIs are 49 to 74% longer for the

TABLE 5.3. Coverage Rates of 95% Confidence Intervals for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model

Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{a}$	30	0.966	0.893	0.881	0.891
	90	0.951	0.895	0.918	0.941
	270	0.956	0.936	0.951	0.954
$\hat{b}$	30	0.967	0.873	0.866	0.883
	90	0.956	0.879	0.914	0.942
	270	0.960	0.920	0.946	0.957

adjusted model than for the unadjusted model. Note that the pattern of results is the same for $\hat{\mu}_1$ and $\hat{\mu}_2$; however, the coverage rates are generally *slightly* lower for $\hat{\mu}_2$ than for $\hat{\mu}_1$ for the adjusted model.

Now looking at Table 5.3, I observe coverage rates that are just barely above the nominal rate of 0.95 for $\hat{a}$ and $\hat{b}$ when the variance component is 0. Once heterogeneity is present and as the variance component increases, the coverage rates generally increase from 0.87 to 0.96. As the number of effects increases, the coverage rates again increase and approach their nominal rate. (Note that the pattern of results for $\hat{a}$ and $\hat{b}$ is the same as for $\hat{\mu}_1$ and $\hat{\mu}_2$.)

Last, I examined the results of bias and root mean squared errors for the parameters of the unadjusted and adjusted models in order to assess the quality of estimation. Bias indicates whether the parameter estimates match the true values (i.e., accuracy) and RMSE indicates how closely the estimates cluster together near the true value (i.e., precision) after taking systematic bias and sampling variability into account. The hope is that the estimates are lower (i.e., closer to zero) for the adjusted model than for the unadjusted model, which would indicate smaller deviations. Tables 5.4 to 5.6 present the results of bias and Tables 5.7 to 5.9 present the results of RMSE. Both sets of results are multiplied by 1,000 for ease of presentation; thus, the results must be examined carefully, as many of

TABLE 5.4. $1000 \times$ Bias for the Parameter Estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ for the Unadjusted and Adjusted Models

Unadjusted Model					
Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{\mu}_1$	30	1.291	-0.006	-3.030	-4.108
	90	0.203	-0.385	-0.085	-0.009
	270	-0.055	0.013	0.230	0.369
$\hat{\mu}_2$	30	2.881	4.348	1.999	0.202
	90	0.124	0.020	0.340	0.216
	270	0.011	0.336	-0.931	-0.292
Adjusted Model					
Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{\mu}_1$	30	0.623	24.969	19.426	12.603
	90	-8.005	-18.704	-23.613	-27.448
	270	-3.576	-8.528	-10.773	-9.827
$\hat{\mu}_2$	30	63.670	13.664	1.185	-6.267
	90	22.090	-16.240	-20.884	-20.796
	270	6.404	-6.510	-10.053	-8.180

the values are indeed close to the desired value of 0. Looking first at the results of bias for $\hat{\mu}_1$ and $\hat{\mu}_2$ (see Table 5.4), I observe that bias fluctuates as the number of effects increases and as the variance component increases (i.e., the pattern is not consistent). However, keeping in mind how close these values are to 0, the estimates really do not change, or waver, very much. Generally, bias is lower for the unadjusted model than for the adjusted model for both $\hat{\mu}_1$ and $\hat{\mu}_2$.

The pattern of results for $\hat{\tau}^2$ is clearer and more consistent (see Table 5.5). For both the unadjusted and adjusted models, bias increases as the variance component increases when heterogeneity is present. However, when heterogeneity is introduced (i.e., the variance component changes from 0 to 0.05) and at least a moderate number of effects is present ($k \geq 90$), bias decreases for the adjusted model. For both models, as the number

TABLE 5.5. $1000 \times$ Bias for the Parameter Estimate $\hat{\tau}^2$ for the Unadjusted and Adjusted Models

Model	k	τ^2			
		0.000	0.050	0.090	0.140
Unadjusted	30	2.921	-8.744	-10.773	-13.334
	90	1.728	-2.816	-3.428	-4.533
	270	0.988	-0.814	-1.201	-1.558
Adjusted	30	3.788	-11.779	-16.537	-21.620
	90	3.061	-1.505	-2.956	-3.458
	270	1.397	-0.357	-0.854	-1.189

TABLE 5.6. $1000 \times$ Bias for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model

Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{a}$	30	84.059	-11.448	-28.340	-33.147
	90	34.491	-20.393	-23.796	-19.066
	270	11.578	-7.697	-8.132	-5.992
$\hat{b}$	30	225.338	-105.791	-132.682	-120.849
	90	120.193	-36.267	-31.118	-16.484
	270	49.156	-10.861	-4.310	-1.672

of effects increases, bias decreases. Moreover, bias is generally lower for the unadjusted model than for the adjusted model (except when $\tau^2 \geq 0.05$ and $k = 270$).

The pattern of results for $\hat{a}$ and $\hat{b}$ is very similar to that of $\hat{\tau}^2$ (see Table 5.6). Bias decreases as the number of effects increases for both the unadjusted and adjusted models. The pattern is not as clear when the variance component increases; however, when heterogeneity is introduced (i.e., the variance component changes from 0 to 0.05), there is a large drop in bias for both $\hat{a}$ and $\hat{b}$. Note that bias is largest when no heterogeneity is present.

I now move on to examine the results of root mean squared errors, which denote the square roots of the average squared estimation errors and represent useful indices of

TABLE 5.7. $1000 \times \text{RMSE}$ for the Parameter Estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ for the Unadjusted and Adjusted Models

		Unadjusted Model			
		τ^2			
Parameter	k	0.000	0.050	0.090	0.140
$\hat{\mu}_1$	30	52.983	85.120	101.269	117.532
	90	29.710	48.422	58.192	67.582
	270	16.818	27.982	33.292	38.395
$\hat{\mu}_2$	30	94.219	116.735	129.279	143.308
	90	53.368	67.456	75.450	83.383
	270	30.282	38.647	42.900	47.392

		Adjusted Model			
		τ^2			
Parameter	k	0.000	0.050	0.090	0.140
$\hat{\mu}_1$	30	130.816	254.358	312.124	366.193
	90	86.343	201.281	227.300	250.695
	270	38.322	108.407	121.384	129.885
$\hat{\mu}_2$	30	249.454	338.458	371.187	392.363
	90	138.074	219.554	228.043	229.388
	270	63.344	115.467	118.004	119.353

TABLE 5.8. $1000 \times \text{RMSE}$ for the Parameter Estimate $\hat{\tau}^2$ for the Unadjusted and Adjusted Models

		τ^2			
Model	k	0.000	0.050	0.090	0.140
Unadjusted	30	8.081	32.259	45.539	60.032
	90	4.034	18.693	26.223	34.847
	270	2.106	10.658	14.856	19.888
Adjusted	30	16.634	55.896	80.367	106.523
	90	10.099	39.996	53.133	65.842
	270	4.402	22.149	28.749	35.171

the overall quality of estimation. Here, the patterns are very clear. Because the pattern of results is identical for parameter estimates $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\tau}^2$, I describe all of them in one. Tables 5.7 and 5.8 present the results. For all three parameters, the unadjusted model always obtains lower RMSE than the adjusted model. RMSE decreases as the

TABLE 5.9. $1000 \times \text{RMSE}$ for the Parameter Estimates $\hat{a}$ and $\hat{b}$ for the Adjusted Model

Parameter	k	τ^2			
		0.000	0.050	0.090	0.140
$\hat{a}$	30	264.396	297.035	299.363	285.251
	90	150.711	184.220	169.834	144.517
	270	72.133	94.787	80.853	67.602
$\hat{b}$	30	631.916	488.650	452.193	413.300
	90	368.215	353.535	296.943	233.760
	270	186.910	204.879	150.227	110.199

number of effects increases and increases as the variance component increases for both the unadjusted and adjusted models.

The pattern of results for $\hat{a}$ and $\hat{b}$ is a little different than that of $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\tau}^2$ (see Table 5.9). RMSE decreases as the number of effects increases. RMSE also tends to decrease as the variance component increases when heterogeneity is present. When heterogeneity is introduced (i.e., the variance component changes from 0 to 0.05), however, RMSE increases.

Conclusions

Overall, the results of these simulations suggest that the beta density weight-function model does a great job of distinguishing systematic study differences from publication bias when the model is correctly specified (i.e., the correct predictor is identified). The model obtains Type I errors for the likelihood-ratio test for publication bias that are close to the nominal rates in all cells of the simulation, indicating that the model will very rarely cause one to erroneously conclude that publication bias is present in the meta-analysis when it is actually not. Moreover, the coverage rates show that the true parameter values are practically just as likely to be caught within the 95% confidence intervals of the adjusted estimates (i.e., those that use the beta density as the weight function that models

publication bias) as of the unadjusted estimates (i.e., the effect-size model estimates that assume bias is not present). That is comforting, as the adjusted estimates are then close to the true values and do not attempt to adjust for bias when it does not exist. The indices of how well parameter estimation is working based on estimate deviation did not provide equally nice results. Both bias and RMSE indicate that the unadjusted model almost always provides better parameter estimates by indicating that they are closer to their true values (i.e., they deviate less). That is not surprising because in this set of simulations publication bias is not present; thus, there is no reason for the unadjusted model to do poorly as it assumes publication bias is not present. One can worry about this, or simply think of it this way: when the model fails to detect publication bias (which it does well in this scenario), simply use the unadjusted parameter estimates. However, note that using the adjusted parameter estimates is not a bad thing. Recall that the coverage rates suggest that the true parameter values are very likely to be observed within the 95% CIs of both the unadjusted and adjusted parameter estimates (although the CIs are longer for the adjusted model than for the unadjusted model). Moreover, even though bias and RMSE show that the unadjusted model estimates deviate from their true values less than the adjusted model estimates, neither set of estimates deviate very much.

CHAPTER 6

Analyzing Real Meta-Analytic Data Using Bayesian Estimation

Maximum-likelihood (ML) estimation of the beta density weight-function model sometimes fails either by producing nonsensical estimates or by simply not converging. For that reason, among others, I also implemented the model using Bayesian estimation. Bayesian estimation summarizes the characteristics of a data set by finding the posterior distribution for each parameter. Many researchers use Markov chain Monte Carlo (MCMC) sampling algorithms to do so, as sampling from the posterior distribution is simpler than analytically determining the characteristics of the posterior; indeed, in many complex models, including the one presented here, MCMC represents the only way to obtain the posterior distributions. Simple summary statistics can then be obtained on the sampled posterior distributions (Lynch, 2007). Moreover, Bayesian estimation allows for subjective beliefs about the parameters to be included by incorporating prior knowledge (or priors). These priors also may be used to prevent the estimates from wandering into an unreasonable space. For those reasons, Bayesian estimation is seen as more flexible.

In this chapter, I analyze five real meta-analytic data sets using the Bayesian implementation of the beta density weight-function model. I sample from the posterior distribution using the Metropolis-Hastings (MH) algorithm (see, e.g., Chib & Greenberg, 1995; Geyer, 1991; Lynch, 2007).¹ I then compare the Bayesian estimation results of those data sets to the ML estimation results that I obtain using *R*'s (Version 2.9.2) optimizer *nlm*. I start with real data sets, rather than a simulation, because a simulation on a simulation

¹Recall that I cannot use Gibbs sampling, the more commonly used and generally more efficient sampling algorithm, because the derivation of the conditional distributions for the parameters is not a tractable problem for the beta density weight-function model (see, e.g., Casella & George, 1992; Lynch, 2007).

is demanding of computer resources and because these estimation procedures typically need human intervention to work well. I first want to determine how well the model can handle data with various characteristics using the Bayesian implementation. In Chapter 7, I will focus on a small simulation study to investigate the performance of the beta density weight-function model on data sets with specific data characteristics when the model is implemented using Bayesian estimation. The data sets I include in the current study represent variations in the number of effects included in the meta-analysis (k), severity of publication bias, and heterogeneity. In addition, some of the data sets obtain seemingly nonsensical parameter estimates when the model is implemented using ML estimation. The goal is to see whether Bayesian estimation can succeed in estimating the parameters of these problematic data sets when ML estimation fails.

Method

In order to assess how well a Bayesian implementation of the beta density weight-function model can handle meta-analytic data with various characteristics, I first run the model on real data sets. I chose five meta-analytic data sets to work with (i.e., run the model on) based on variations in the following characteristics: number of effects included in the meta-analysis (k), severity of publication bias, heterogeneity, and whether the data set was problematic in that it obtained nonsensical parameter estimates when implemented using ML estimation. The data sets included are: a large collection of validity coefficients for the General Aptitude Test Battery (GATB)², Else-Quest, Hyde, Goldsmith, and Van Hulle's (2006) gender differences in effortful control temperament data (ECT), Shadish and Baldwin's (2005) behavioral marital therapy data (BMT), Petrosino, Turpin-Petrosino, and Guckenburg's (2010) system processing effects on juvenile delinquency data (JD), and Raudenbush's (1984) effects of teacher expectancy on pupil IQ data (TE)³.

²The GATB data set was retrieved from Vevea et al. (1993).

³The TE data set was retrieved from Raudenbush and Bryk (1985).

The estimation process is unique for each data set, thus, I focus on one data set at a time. In each section, I describe the data set, the estimation process, and the results. This will allow the reader to follow the story behind each data set more easily. But first I briefly explain the procedure that is common to all the data sets.

General Procedure

In Bayesian estimation, the posterior density is the product (or summation, if one is working in the log metric) of the likelihood and prior beliefs. I used the MH algorithm (described in Chapter 3's section *Bayesian Estimation*) to sample values for each parameter from the posterior distribution using the log-likelihood function in Equation 3.4 as the likelihood. In order to obtain sampled values for the adjusted model parameters, I used the full log-likelihood function (i.e., including the selection model). To obtain values for the unadjusted model parameters, I first set a and b to 1.0, as that allows for the estimation of the effect-size model without selection. (But recall that one could simply estimate the standard meta-analytic model, as it is formally equivalent to estimating the weighted model with a and b set to 1.0.) I ran the unadjusted and adjusted models on the data sets as mean-only random-effects models. However, if it was clear that a moderator was necessary, I included it in subsequent mixed-effects analyses to ensure proper model specification. (Note that including τ^2 in all analyses allows for the generalization of the results to a hypothetical population of studies.)

As my prior beliefs, I used weak generic priors for each of the data sets because I did not have strong opinions about how much bias each data set may contain. I used priors for a and b that made theoretical sense in order to keep the parameter estimates from diverging into values that are unrealistic. The prior for a was based on a χ^2 distribution with the posterior density equal to $\chi^2(10 - 10(\text{value}), 2)$ with 2 df . That keeps a below one, as values of a above one do not make sense under most circumstances because they suggest the opposite of publication bias (while $a = 1$ simply indicates no bias when b also

equals 1). The prior for b was set to a normal distribution with the mean centered at one and the standard deviation set to 0.4 [i.e., $N(1, 0.4^2)$], which keeps b within a reasonable range.⁴ Note that these are the priors I simply started with; if a data set did not converge and necessitated modifications, adjustments were made and the MH algorithm was run again on that data set.

I ran the MH algorithm for each data set four times with various starting values for each of the parameters in order to ensure the estimates converge and to make sure the distribution does not have multiple modes at which the estimates could get stuck. The idea is that if all the chains (or sets of sampled values) for each of the parameters (with their varying starting values) converge to approximately the same values, then that is a good indication that the parameters have converged to the appropriate values (Lynch, 2007). I ran the four chains for each data set with the following starting values for the adjusted model's parameters:

- (1) $a = 1, b = 1, \mu = \text{unadjusted ML } \hat{\mu}, \tau^2 = \text{unadjusted ML } \hat{\tau}^2$;
- (2) $a = 0.5, b = 1.5, \mu = \text{unadjusted ML } \hat{\mu}, \tau^2 = 0$;
- (3) $a = 0.8, b = 1.2, \mu = 1\sigma \text{ above the unadjusted ML } \hat{\mu}, \tau^2 = \text{unadjusted ML } \hat{\tau}^2$;
- (4) $a = 0.8, b = 1.2, \mu = 1\sigma \text{ below the unadjusted ML } \hat{\mu}, \tau^2 = \text{unadjusted ML } \hat{\tau}^2$.

The unadjusted model's parameter starting values are the same as those of the adjusted model's with the exception that a and b are always set to 1.0 in the unadjusted model.

Last, I varied the number of iterations (or replications) run in each chain for each data set to ensure convergence. I ran the study in *R* (Version 2.9.2) on a computer using the Microsoft Windows operating system.

⁴I examined beta density plots with *many* combinations of a and b in order to determine what a reasonable range of values would look like in order to model anything from no publication bias to severe publication bias.

Convergence

Before analyzing the results, I first assessed whether all the chains have converged and whether the acceptance rates (or rates at which the candidate values have been accepted) for all the chains are within the recommended 0.25 to 0.75 range (see, e.g., Chib & Greenberg, 1995; Geyer, 1991; Lynch, 2007). I determined convergence by examining plots, calculating the potential scale reduction factor (or *PSR*), and estimating the MCMC error for the parameter μ (see, e.g., Lynch, 2007). I examined each chain separately by first looking at trace plots and cumulative sum plots to determine whether the chain of parameter values has stabilized (i.e., has stopped moving up and down). I then examined bivariate trace plots of each set of parameter chains (a and b , a and μ , b and μ , etc.) in order to examine whether the values of those parameters have mixed well in the bivariate posterior distributions. That is, I wanted to see if the sets of parameter values moved around in the posterior parameter space in such a way as to indicate convergence.

After determining the burn-in period (the section of a chain where the values have not yet stabilized) by looking at trace and cumulative sum plots, I cut the non-converged values from the set of sampled values. I then calculated the *PSR* factor in order to further assess the remaining values for convergence. The *PSR* factor assesses the reduction of between-chain variance for the sets of chains run for each parameter to examine if each chain of a parameter has converged to approximately the same value. It is estimated by dividing the total variance (or variance that includes both variation between and within the chains) by the within variance of the sets of chains for each parameter. Convergence is reached when the square-root of the *PSR* factor is "close to 1" (Lynch, 2007). Or, according the *Mplus* User's Guide, the *PSR* should be within the boundaries of 0.95 and 1.05 (Muthén & Muthén, 2010).

After determining convergence via the *PSR* factor, I calculated the autocorrelation of each chain at a series of g lag values. I determined the "best" g by looking at a plot

of potential g values and their corresponding autocorrelations for all the chains of all the parameters. I chose g to be the lag value at which the autocorrelations of all the chains stop decreasing at such a drastic rate (i.e., when the autocorrelations start leveling out). Consequently, I reduced (or thinned) the chains of sampled values further by sampling every g th value in order to reduce the autocorrelation within each chain.

After I obtained my final set of sampled values, I assessed convergence once more by examining how close the means of the chains of parameter μ are to one another. According to the Central Limit Theorem, the means of the sampled values, or chains, should be normally distributed with a mean equal to the population mean and a standard error equal to the population standard deviation divided by the square-root of the sample size. I estimated the MCMC error of each $\hat{\mu}$ chain by dividing the posterior standard deviation of $\hat{\mu}$ by the square-root of the length of the final reduced chain. Ideally, 95% of the MCMC runs should produce posterior means within ± 1.96 standard errors (Lynch, 2007). However, because the runs of chains are far from infinite, I simply want to make sure the means of the chains for $\hat{\mu}$ are approximately within the other chains' intervals ($\hat{\mu} \pm$ MCMC error). Note that I am only assessing the closeness of the means of the posterior distributions for μ . The reason is two-fold: 1) τ^2 is not normally distributed, but rather follows a χ^2 distribution⁵, and a may not be normally distributed if the χ^2 prior is strongly in effect; and 2) if the means of the chains of μ are vastly different from one another, the means of the chains of a and b should also differ, as a and b determine the bias adjustment for the adjusted μ (i.e., all the parameters affect one another, so if μ is vastly different across the chains, the other parameters will differ as well).

FIGURE 6.1. An extended funnel plot of the general aptitude test effects against their standard errors.

GATB Data

The General Aptitude Test Battery, or GATB, was intended to predict job performance by measuring cognitive, perceptual, and psychomotor skills. From 1947 to 1993, a total of 755 studies had been completed in order to assess the validity of the GATB and its nine scales. The GATB has been found to be a moderately valid predictor of job performance (Vevea et al., 1993). The data set I employ here consists of validity coefficients for the GATB's General Ability scale; the correlations have been transformed by Fisher's Z (a commonly used variance stabilizing transformation). When all 755 effect-size estimates

⁵The χ^2 distribution follows a normal distribution as the sample size approaches infinity. However, the sample size here is far from infinite.

are combined, the unadjusted ML mean effect size $\hat{\mu}$ is 0.26. The variance component $\hat{\tau}^2$ equals about 0.01, indicating that there is heterogeneity, or excess between-study variability. Recall that a population is homogeneous if sampling error is the sole source of variability in the observed effect sizes (Lipsey & Wilson, 2001). In the ML analysis, the null hypothesis that the GATB distribution of effects is homogeneous, or that $\tau^2 = 0$, is rejected ($Q = 1493.74$, $df = 754$, $p < 0.001$). That suggests heterogeneity is present in the data, indicating that the distribution of GATB effect-size estimates is spread about the mean by more than just random sampling of persons into the studies. Moreover, despite the small value for τ^2 , according to the I^2 statistic, the amount of heterogeneity present in the data is not trivial ($I^2 = 49.52\%$). When the parameter estimates are adjusted for publication bias using the beta density weight-function model, the ML estimation results indicate that $\hat{a}$ and $\hat{b}$ are 0.91 and 1.93, respectively. Consequently, $\hat{\mu}$ reduces to 0.17, $\hat{\tau}^2$ is 0.02, and the likelihood-ratio test suggests that bias is present ($\chi^2 = 13.42$, $df = 2$, $p = 0.001$).

Procedure

I ran four separate sets of chains of the MH algorithm to sample unadjusted and adjusted parameter values from the posterior distribution of the GATB data. To ensure the values do not converge to separate modes of the distribution, I used different sets of starting values for each of the sets of chains:

- (1) $a = 1$, $b = 1$, $\mu = 0.260$, $\tau^2 = 0.011$;
- (2) $a = 0.5$, $b = 1.5$, $\mu = 0.260$, $\tau^2 = 0$;
- (3) $a = 0.8$, $b = 1.2$, $\mu = 0.266$, $\tau^2 = 0.011$;
- (4) $a = 0.8$, $b = 1.2$, $\mu = 0.254$, $\tau^2 = 0.011$.

(Note that when estimating the unadjusted model, a and b are always set to 1.0; μ and τ^2 follow the pattern above.) Because I did not have strong prior beliefs about the data set, I used the generic weak priors mentioned in the *General Procedure* section ($a =$

FIGURE 6.2. A cumulative sum plot of the four chains of parameter $\hat{b}$ of the GATB data.

$\chi^2(10 - 10(\text{value}), 2)$ and $b = N(1, 0.4^2)$) to keep the estimates from wandering into an unrealistic space.

I first ran each chain with 20,000 iterations. After examining the trace and cumulative sum plots of each chain for each parameter, I was not convinced that all the parameters converged to stable values. Although the results did not differ vastly between the chains, I decided to re-run the chains with 30,000 iterations. Cumulative sum plots containing generally overlapping chains for each parameter convinced me that this time the estimates have stabilized. Figure 6.2 shows an example of a good cumulative sum plot where the chains of parameter $\hat{b}$ are mostly overlapping and stabilizing to the same value. Some of

FIGURE 6.3. A trace plot of the third chain of the adjusted $\hat{\mu}$ parameter of the GATB data.

the trace plots still show a bit of movement among the values in the chains, but overall look fine. The adjusted $\hat{\mu}$ is the parameter that had the most trouble stabilizing. Figure 6.3 displays the third chain of the adjusted $\hat{\mu}$, which shows a bit of wavering, but still decent convergence. Moreover, the bivariate trace plots indicated that each set of parameter chains mixed sufficiently (i.e., converged) in the bivariate posterior distributions. Figure 6.4 shows an example of the first chain of adjusted parameters $\hat{\tau}^2$ and $\hat{\mu}$ mixing very well, indicated by the nice oval shape of the plot (the lines outside of the oval represent the burn-in period). Also, the acceptance rates were within the recommended 0.25 to 0.75

FIGURE 6.4. A bivariate trace plot of the first chain of adjusted parameters $\hat{\tau}^2$ (VC) and $\hat{\mu}$ (Mean) of the GATB data.

range (see, e.g., Chib & Greenberg, 1995; Geyer, 1991; Lynch, 2007), ranging from 0.29 to 0.61.

After examining the plots, I identified the first 9,999 iterations as my burn-in period (or section of a chain where the values have not yet stabilized). I calculated the *PSR* factor on the remaining 20,001 iterations, which resulted in a *PSR* that ranges from 1.00 to 1.01 across the six parameters (with the unadjusted parameters obtaining lower *PSR* values than the adjusted parameters). Note that this is well within Muthén and Muthén's (2010) recommended range and thus suggests convergence.

Next, recall that the sampled values are not obtained independently (see, e.g., Geyer, 1991; Lynch, 2007); thus, I sub-sampled from each chain by taking every 50th sampled

value. That reduced the autocorrelations of each chain to approximately 0 for the unadjusted parameters and between 0.10 and 0.60 for the adjusted parameters. It also reduced the number of sampled values to 401 estimates per chain. I then calculated the MCMC error of $\hat{\mu}$ on those final sampled values in order to assess whether the $\hat{\mu}$ chains converged to approximately the same values. For the unadjusted $\hat{\mu}$, the intervals for one standard error below and above the mean across the four chains are: (0.2593, 0.2599), (0.2601, 0.2607), (0.2596, 0.2602) and (0.2598, 0.2604).⁶ For the adjusted $\hat{\mu}$, the intervals are: (0.1949, 0.1980), (0.1988, 0.2015), (0.1955, 0.1984), and (0.1911, 0.1940). For both the unadjusted and adjusted $\hat{\mu}$, the means appear very close to one another, as well as mostly within one another's intervals, suggesting the estimates converged to approximately the same values.

Results

I was first interested in assessing whether bias is present in the GATB data. I determined which model (the unadjusted model that assumes publication bias is not present or the adjusted model that includes a weight function that models publication bias) fits the data better by estimating the BIC ⁷ and AIC ⁸. The model that obtains a lower BIC or AIC is considered the better fitting model (Congdon, 2003; Lynch, 2007). Across all four chains, the BIC is $-1,973.80$ for the unadjusted model and $-1,973.14$ for the adjusted model. With such a trivial difference, the BIC suggests both models fit equal well. The AIC , on the other hand, is lower for the adjusted model ($-1,991.65$) than for the unadjusted model ($-1,983.05$), suggesting some bias may be present. With posterior

⁶Note that I increase the number of decimal points presented in the intervals to four (rather than the two presented throughout the dissertation) in order to show the small variation present between the intervals. Otherwise, many intervals may look like they have a length of zero (e.g., 0.26 to 0.26).

⁷ $BIC = D + j \log(k)$, where D is the deviance of the model, defined as minus two times the log-likelihood, j represents the number of parameters in the model, and k denotes the number of effects present in the meta-analysis.

⁸ $AIC = D + 2j$, where D is the deviance of the model and j represents the number of parameters in the model.

mean values of 0.93 and 1.67 for $\hat{a}$ and $\hat{b}$, respectively, I agree that some bias must be present (otherwise $\hat{b}$ should be closer to 1.0). The posterior mean of $\hat{\mu}$ is then adjusted from 0.26 to 0.20. The adjustment makes sense. If one looks at the extended funnel plot of the GATB data in Figure 6.1, it does appear that a few effects may be missing from the bottom left-hand side of the plot. It is the heavily-weighted effects on the top left-hand side of the plot that may be inflating the mean effect-size estimate. Thus, a $\hat{\mu}$ adjustment from 0.26 to 0.20 does not seem unreasonable. This suggests the GATB may not be as valid as once assumed. This is a 23% reduction in the magnitude of the transformed validity coefficient, which may change one's opinion in whether to use the GATB to predict job performance. In fact, the credibility interval for the adjusted $\hat{\mu}$ indicates that 95% of the GATB Fisher's Z transformed correlation validity coefficients are between 0.13 and 0.25, which does not include the unadjusted mean validity coefficient of 0.26.

Discussion

Overall, Bayesian estimation does a decent job of summarizing the characteristics of the GATB data. While I am not pleased with the remaining movement in some of the trace plots, I believe the estimates generally converged to reasonable values. Note that the z-transformed correlation coefficient is the same in the ML and Bayesian estimations. That suggests that with such a large non-problematic data set, both Bayesian and ML estimation seem to do a good job of estimating the parameters of the GATB distribution.

ECT Temperament Data

The effortful control temperament data, or ECT, is taken from Else-Quest et al.'s (2006) meta-analysis, which assessed gender differences in child temperament. In order to create a medium-sized data set ($k = 88$), only a portion of the data (distractibility and persistence categories of attention from the effortful control dimension of temperament in Table 1) are included in this study. Distractibility refers to a child being sidetracked by

FIGURE 6.5. An extended funnel plot of the ECT effects against their standard errors.

things in his or her environment, while persistence refers to a child staying on task. This data set is interesting in that it may represent two-tailed selection. Two-tailed selection is likely to occur when there is no expectation of an effect in either direction (positive or negative; see, e.g., Vevea & Woods, 2005; Hedges & Vevea, 2005). With the ECT data, the authors noted that, based on prior research, either males or females might score higher on temperament. Furthermore, in practice, the effects present in the meta-analysis show that both males and females performed better on temperament on different occasions. Thus, if any selection exists, two-tailed selection is more plausible than one-tailed selection. Recall that the beta density weight-function model is able to capture both one- and two-tailed selection. The unadjusted ML mean effect size $\hat{\mu}$ for the ECT data is -0.04 and

$\hat{\tau}^2$ is 0.06, the latter indicating a large amount of heterogeneity ($Q = 357.55$, $df = 87$, $p < 0.001$; $I^2 = 75.67\%$). The adjusted ML parameter estimates are 1.08 and 1.01 for $\hat{a}$ and $\hat{b}$, respectively, 0.01 for $\hat{\mu}$ and 0.07 for $\hat{\tau}^2$. With $\hat{a}$ and $\hat{b}$ approximately equal to 1.0, the likelihood-ratio test is not significant ($\chi^2 = 0.35$, $df = 2$, $p = 0.84$), indicating that there is no evidence that publication bias is present.

Procedure

In order to sample from the posterior distribution of the ECT effects, I ran four separate sets of chains of the MH algorithm with the following starting values for the unadjusted and adjusted parameters (with a and b set to 1.0 in the unadjusted model):

- (1) $a = 1$, $b = 1$, $\mu = -0.036$, $\tau^2 = 0.059$;
- (2) $a = 0.5$, $b = 1.5$, $\mu = -0.036$, $\tau^2 = 0$;
- (3) $a = 0.8$, $b = 1.2$, $\mu = -0.003$, $\tau^2 = 0.059$;
- (4) $a = 0.8$, $b = 1.2$, $\mu = -0.069$, $\tau^2 = 0.059$.

Using the generic priors of $a = \chi^2(10 - 10(value), 2)$ and $b = N(1, 0.4^2)$, I ran each chain with 35,000 iterations. Once the algorithm finished running, I first examined the acceptance rates, which ranged from 0.30 to 0.42 and were well within the recommended range. I then examined convergence by looking at plots and by calculating the *PSR* factor and MCMC error for μ . The trace plots, cumulative sum plots, and bivariate trace plots for all the parameters (and sets of parameters) indicated that the chains of the parameters converged to stable values. The chains converged by around 20,000 iterations, thus, I cut the first 19,999 iterations from all the chains and did not include them in my analyses (i.e., the burn-in period is 19,999 iterations). I then estimated *PSR* values of about 1.00 for all six parameters, indicating convergence.

Next, I sub-sampled every 25th sampled value from each chain in order to reduce the autocorrelations of each chain. The autocorrelations were reduced to 0 for the unadjusted parameters and between 0.10 and 0.48 for the adjusted parameters. That also resulted

in cutting down the number of sampled values to 601 per chain. Based on those final values, I assessed convergence one more time by estimating the MCMC error of $\hat{\mu}$ and calculating the intervals for the unadjusted and adjusted means of $\hat{\mu}$. For the unadjusted $\hat{\mu}$, the intervals for one standard error below and above the mean are the following for the four chains: $(-0.0385, -0.0355)$, $(-0.0374, -0.0345)$, $(-0.0377, -0.0351)$, and $(-0.0384, -0.0356)$. For the adjusted $\hat{\mu}$, the intervals are: $(-0.0846, -0.0782)$, $(-0.0820, -0.0753)$, $(-0.0777, -0.0713)$, and $(-0.0782, -0.0713)$. With very large overlap in estimates within both the unadjusted and adjusted sets of intervals, I am again confident that the chains have converged to the same values.

Results

I first assessed the effortful control temperament data for publication bias. I estimated the *BIC* and *AIC* on the final set of values. Across the four chains, the *BIC* and *AIC* are -95.82 and -100.78 for the unadjusted model, respectively, and -85.99 and -95.90 for the adjusted model, respectively. With consistently lower values for the unadjusted model, the results indicate that the unadjusted model is the better fitting model. This suggests that publication bias has not been found in the ECT data, as adding the weight function to the effect-size model does not improve model fit. This is not surprising given that the posterior means of $\hat{a}$ and $\hat{b}$ are estimated to be close to 1.0 ($\hat{a} = 0.94$ and $\hat{b} = 1.00$). Moreover, the posterior means of $\hat{\tau}^2$ (0.06 for both the unadjusted and adjusted models) and $\hat{\mu}$ (-0.04 and -0.08 for the unadjusted and adjusted model, respectively) almost do not change. The results are in line with those of ML estimation, which also failed to find evidence of bias in the ECT data. This suggests that the standardized mean difference (SMD) for children's scores on the effortful control dimension of temperament is most likely about -0.04 . In addition, there is a 95% chance that the mean SMD across the random-effects distribution is between -0.10 and 0.03 . Note that a Bayesian approach supports such probabilistic statements because the mean SMD is viewed as a random variable.

Discussion

Bayesian estimation appears to have no trouble obtaining the posterior distributions of the parameters of the temperament (or ECT) data. Although the estimates may take a substantial amount of time to converge and sometimes show some wavering (similar to that of the GATB data, but less extreme), the parameters do stabilize to appropriate values. However, it is possible that slightly longer runs of the MH algorithm may provide even more consistent estimates.

Behavioral Marital Therapy Data

The behavioral marital therapy data, or BMT, comes from Shadish and Baldwin's (2005) meta-analysis. The meta-analysis examined 30 SMD effect sizes in order to assess whether behavioral marital therapy is better than no therapy in distressed couples. The unadjusted ML mean effect size $\hat{\mu}$ is 0.58 and $\hat{\tau}^2 = 0$. Despite the fact that τ^2 is zero, a significant amount of heterogeneity is present in the data ($Q = 42.25$, $df = 29$, $p = 0.05$; $I^2 = 31.36\%$). The adjusted ML estimates obtain values of 0, 7.59, -0.99, and 0.02 for $\hat{a}$, $\hat{b}$, $\hat{\mu}$, and $\hat{\tau}^2$, respectively. Notice how extreme those adjusted estimates are: $\hat{a}$ hits its lower-bound of 0,⁹ while $\hat{b}$ is almost 8. Those estimates suggest that *a lot* of bias is present, which is validated via the significant likelihood-ratio test ($\chi^2 = 25.60$, $df = 2$, $p < 0.001$) and the huge mean adjustment (from 0.58 to -0.99). While publication bias may truly be present in this meta-analysis, it is doubtful that the mean should be adjusted down by 271%. That is, although the ML parameter estimates are converging, I am not sure they are converging to reasonable values; the beta density weight-function model may be adjusting too much in this case.

The same phenomenon occurs when a mixed-effects analysis is run with a moderator (outcome specificity) included in the model. Outcome specificity refers to the idea that

⁹Note that a lower-bound of 0 must be placed on both a and b when estimating the model parameters as values below 0 are undefined for the beta density.

FIGURE 6.6. An extended funnel plot of the behavioral marital therapy effects against their standard errors.

some studies used outcome measures that were more closely tied to the treatment than others. For example, using the number of arguments as an outcome for a therapy that trains couples to reduce arguments would lead to a high score on outcome specificity. With extreme estimates for $\hat{a}$ and $\hat{b}$ (0 and 7.99, respectively), the ML estimation results again indicate that a large amount of publication bias must be present in the meta-analysis ($\chi^2 = 22.08$, $df = 2$, $p < 0.001$). Consequently, $\hat{\beta}_0$ is over-adjusted down from -0.07 to -1.58 and $\hat{\beta}_1$ decreases from 1.04 to 0.70 ($\hat{\tau}^2$ increases from 0 to 0.03). With a mean of 0.66 for the outcome specificity moderator, the mean effect for a typical study (i.e., one where outcome specificity = 0.66) is adjusted down from 0.62 to -1.12 .

FIGURE 6.7. A trace plot of the first chain of the adjusted $\hat{\tau}^2$ parameter of the behavioral marital therapy data.

Procedure

I ran four separate chains of the MH algorithm using the following starting values for the unadjusted and adjusted parameters (with a and b set to 1.0 in the unadjusted model) in order to sample from the posterior distribution of the BMT effects:

- (1) $a = 1, b = 1, \mu = 0.580, \tau^2 = 0$;
- (2) $a = 0.5, b = 1.5, \mu = 0.580, \tau^2 = 0$;
- (3) $a = 0.8, b = 1.2, \mu = 0.668, \tau^2 = 0$;
- (4) $a = 0.8, b = 1.2, \mu = 0.491, \tau^2 = 0$.

I used the generic priors for a and b and ran each chain with 45,000 iterations. With values ranging from 0.40 to 0.62, the acceptance rates are within the recommended range. The trace and cumulative sum plots generally display convergence, although a few of the plots show that the parameter estimates occasionally obtain extreme values. Look, for example, at Figure 6.7 for a trace plot of the first chain of the adjusted $\hat{\tau}^2$. Note, however, that such values are to be expected, especially when estimating τ^2 . They are not a concern so long as the chain of estimates does not diverge systematically for an extended period of time. The bivariate trace plots display adequate convergence in the bivariate posterior distributions. I then cut the burn-in period of 19,999 iterations and estimated the *PSR* factor on the remaining 25,001 iterations. With *PSR* values of about 1.00 for all six parameters, the parameters have reached convergence.

I reduced the autocorrelations of the chains by sub-sampling every 25th sampled value. That reduced the number of sampled values to 1,001 estimates per chain. It also reduced the autocorrelations of each chain to between 0 and 0.22 for the unadjusted parameters and between 0.02 and 0.75 for the adjusted parameters. On those final sampled values, I estimated the MCMC error of $\hat{\mu}$ and calculated the intervals for one standard error below and above the means of the unadjusted and adjusted $\hat{\mu}$ estimates. For the unadjusted $\hat{\mu}$, the intervals are: (0.6081, 0.6140), (0.6156, 0.6216), (0.6076, 0.6134), and (0.6114, 0.6175). For that adjusted $\hat{\mu}$, the intervals are: (0.2050, 0.2180), (0.2177, 0.2306), (0.2013, 0.2145), and (0.2323, 0.2446). The intervals of the chains do not always overlap, indicating that the values are not quite the same across the four chains. An even longer run of the MH algorithm may be required to obtain the same values for the four chains. However, because the intervals are very close to one another, I am not too concerned that one chain is slightly different (e.g., by 0.02 units) than the other.

Next, I incorporated the outcome specificity moderator into the analysis and ran the MH algorithm again with the following starting values for the parameters:

- (1) $a = 1, b = 1, \beta_0 = -0.072, \beta_1 = 1.041, \tau^2 = 0$;
- (2) $a = 0.5, b = 1.5, \beta_0 = -0.072, \beta_1 = 1.041, \tau^2 = 0$;
- (3) $a = 0.8, b = 1.2, \beta_0 = 0.302, \beta_1 = 1.567, \tau^2 = 0$;
- (4) $a = 0.8, b = 1.2, \beta_0 = -0.446, \beta_1 = 0.515, \tau^2 = 0$.

I used the generic priors for a and b and put a moderately weak normal prior of $N(0.03, 0.2^2)$ on the adjusted τ^2 in order to keep it from diverging to nonsensical values (a prior run of the algorithm indicated large adjusted τ^2 divergence). I first tried putting a χ^2 distribution prior on τ^2 , which is equivalent to the common practice of placing an inverse gamma distribution prior on $\frac{1}{\tau^2}$. However, a χ^2 distribution prior did not keep τ^2 from obtaining nonsensical values. I then tried a normal distribution prior, which did help. I ran each chain with 250,000 iterations. While both the trace and cumulative sum plots indicated that the values of each of the chains stabilized, just as in the random-effects analysis, longer runs of the MH algorithm may be desired to obtain estimates that are even more similar. Nonetheless, bivariate trace plots indicated appropriate convergence in the bivariate posterior distributions. Because the estimates took a long time to converge, I cut the first 149,999 iterations for the burn-in period. The remaining 100,001 iterations showed convergence with *PSR* values of about 1.00 for all eight parameters.

The autocorrelations remained high for many lag values, thus, I sub-sampled every 60th value in order to reduced to autocorrelations to between 0.04 and 0.55 for the unadjusted parameters and between 0.01 and 0.58 for the adjusted parameters. That reduced the number of sampled values to 1,667 estimates per chain. I then estimated the MCMC error of $\hat{\beta}_0$ and calculated the intervals for one standard error below and above the means of the unadjusted and adjusted $\hat{\beta}_0$ estimates. For the unadjusted $\hat{\beta}_0$, the intervals are the following for the four chains: $(-0.0381, -0.0190)$, $(-0.0482, -0.0300)$, $(-0.0396, -0.0202)$, and $(-0.0748, -0.0561)$. For the adjusted $\hat{\beta}_0$, the intervals are: $(-0.2901, -0.2703)$, $(-0.2643, -0.2445)$, $(-0.2723, -0.2519)$, and $(-0.2595, -0.2399)$. Despite what

the cumulative sum plots indicated about the necessity of potentially longer runs of the MH algorithm for the four chains to provide the same estimates, these intervals indicate sufficient overlap across the four chains for both the unadjusted and adjusted $\hat{\beta}_0$. That suggests the estimates converged to approximately the same values.

Results

I assessed the random-effects meta-analysis of the behavioral marital therapy effects for publication bias by calculating the *BIC* and *AIC* for the effect-size model with and without the inclusion of the beta density selection model. The results indicate that the better fitting model is the adjusted model that models the selection process as the *BIC* and *AIC* are lower for the adjusted model (3.34 and -2.26 , respectively) than for the unadjusted model (6.07 and 3.27, respectively) across the four chains. That suggests publication bias may be present in the BMT effects. With posterior means of 0.66 and 1.61 for $\hat{a}$ and $\hat{b}$, respectively, the posterior mean of $\hat{\mu}$ is adjusted from 0.61 to 0.22. Thus, the beta density weight-function model implies that the SMD for behavioral marital therapy is most likely around 0.22, with 95% of the SMDs lying between -0.19 and 0.59.

The results of the mixed-effects analysis, which includes outcome specificity as a moderator, also indicate that bias may be present in the BMT effects. The *AIC* is clearly lower for adjusted model (-3.52) than for the unadjusted model (-0.07) and the *BIC* is only slightly so (2.08 and 2.73, respectively). Moreover, posterior means of $\hat{a}$ and $\hat{b}$ are definitely not close to 1.0 (0.75 and 1.60, respectively). Thus, the SMD for behavioral marital therapy is more likely to be about 0.32 than 0.64 for studies with typical outcome specificity when both publication bias and outcome specificity are taken into account.

Discussion

Obtaining the posterior distributions of the behavioral marital therapy effects took a generous amount of computing resources because the estimates took a long time to

FIGURE 6.8. An extended funnel plot of the juvenile delinquency effects against their standard errors.

converge. In fact, it might be beneficial to run the chains even longer. However, the resource allocation was well worth it because the Bayesian adjusted parameter estimates make more sense than the overly adjusted ML estimates. In other words, Bayesian estimation really helped to obtain reasonable parameter estimates (and adjusted SMDs) for the behavioral marital therapy meta-analysis.

Juvenile Delinquency Data

The juvenile delinquency data, or JD, comes from a meta-analysis of 27 SMD effects by Petrosino et al. (2010) who aimed to assess whether juvenile system processing reduces subsequent delinquency. I focus only on the results from the first follow-up session, which

occurred an average of 10 to 11 months after the intervention. The unadjusted ML mean effect size $\hat{\mu}$ is -0.11 and $\hat{\tau}^2 = 0.08$, the latter indicating a large amount of heterogeneity ($Q = 96.99$, $df = 26$, $p < 0.001$; $I^2 = 73.19\%$). The adjusted ML estimates are 1.55, 0.94, 0.81, and 0.29 for $\hat{a}$, $\hat{b}$, $\hat{\mu}$, and $\hat{\tau}^2$, respectively. What is strange about this data set is that bias is adjusted upwards (from -0.11 to 0.81), suggesting that the opposite of publication bias is present (although the likelihood-ratio test is not significant; $\chi^2 = 2.19$, $df = 2$, $p = 0.33$). Note, however, that the adjusted estimate simply does not make sense, as there is only one observed effect that is larger than 0.81 (see Figure 6.8). This is a case where ML estimation appears to produce a nonsensical estimate, and where I hope Bayesian estimation will be able to help.

The authors noted that the type of control group included in a study may moderate the results of the meta-analysis, thus, I included it as a predictor in a mixed-effects analysis. There were two types of control groups included: one with diversion services (e.g., counseling) and one with “simple diversion” (i.e., the group did nothing). The unadjusted ML conditional mean estimates for the two groups are $\hat{\mu}_1 = -0.05$ and $\hat{\mu}_2 = -0.18$, respectively ($\hat{\tau}^2 = 0.07$). The adjusted ML estimates are 1.57, 0.92, 0.94, 0.66, and 0.27 for $\hat{a}$, $\hat{b}$, $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\tau}^2$, respectively. Again, the mean effects are adjusted upwards – in the wrong direction. Perhaps putting weak priors on a and b will keep the estimates from producing values that do not make sense (for this data set).

Procedure

I ran four separate chains of the MH algorithm to sample from the posterior distribution of the JD effects. The chains contained the following starting values for the unadjusted and adjusted parameters (with a and b set to 1.0 in the unadjusted model):

- (1) $a = 1$, $b = 1$, $\mu = -0.108$, $\tau^2 = 0.077$;
- (2) $a = 0.5$, $b = 1.5$, $\mu = -0.108$, $\tau^2 = 0$;
- (3) $a = 0.8$, $b = 1.2$, $\mu = -0.040$, $\tau^2 = 0.077$;

FIGURE 6.9. A bivariate trace plot of the first chain of adjusted parameters $\hat{a}$ and $\hat{\mu}$ (Mean) of the juvenile delinquency data.

$$(4) \ a = 0.8, b = 1.2, \mu = -0.177, \tau^2 = 0.077.$$

I used the generic priors of $a = \chi^2(10 - 10(\text{value}), 2)$ and $b = N(1, 0.4^2)$ and ran each chain with 40,000 iterations. The acceptance rates ranged from 0.30 to 0.49, representing values within the recommended range. The trace and cumulative sum plots of the parameter values indicated that the estimates converged to stable values. However, bivariate trace plots that include parameter a showed worrisome mixing (at first glance) in some of the bivariate posterior distributions. For example, look at figure 6.9 for a bivariate trace plot of the first chain of adjusted parameters $\hat{a}$ and $\hat{\mu}$. Rather than a nice oval, the plot displays a sideways cone. Recall, however, that I am forcing a to be below 1.0; thus,

many values of $\hat{a}$ are hitting the upper-bound value. That is likely to happen any time a minimum or maximum value is set for a parameter (e.g., τ^2 cannot be below zero) so I am not concerned about this bivariate trace plot (or others that contain a). Next, I cut the first 14,999 values as the burn-in period and calculated the *PSR* factor for the parameters on the remaining 25,001 iterations. The *PSR* values are approximately 1.00 for all six parameters, again indicating convergence.

In order to reduce the autocorrelations of the chains, I sub-sampled every 25th sampled value. That reduced the autocorrelations of each chain to approximately 0 for the unadjusted parameters and between 0.18 and 0.43 for the adjusted parameters. It also reduced the number of sampled values to 1,001 estimates per chain. I then assessed the convergence of those final values by estimating the MCMC error of $\hat{\mu}$ and calculating the intervals for the unadjusted and adjusted means of $\hat{\mu}$. For the unadjusted $\hat{\mu}$, the intervals for one standard error below and above the mean are the following for the four chains: $(-0.1038, -0.0989)$, $(-0.1055, -0.1009)$, $(-0.1068, -0.1019)$, and $(-0.1070, -0.1023)$. For the adjusted $\hat{\mu}$, the intervals are: $(-0.1782, -0.1678)$, $(-0.1823, -0.1716)$, $(-0.1742, -0.1635)$, and $(-0.1872, -0.1762)$. With such a large overlap in estimates within both the unadjusted and adjusted sets of intervals, I am assured the chains have converged to the same values.

To ensure that I did not impose bias on a data set that may not be biased (or that may contain the opposite of publication bias) by putting a prior on a that forces it to be below 1.0, I re-ran the MH algorithm with normal priors on both a and b (priors for a and $b = N(1, 0.4^2)$). I ran the chains with the same starting values over 40,000 iterations and found that the estimates did not converge. The values of the parameters are very unstable, diverging to nonsensical values throughout the chains. For example, look at Figure 6.10 for a trace plot of the fourth chain of the adjusted $\hat{\tau}^2$. The $\hat{\tau}^2$ posterior estimates are trying to stay at values close to zero, however, they keep systematically diverging to values as

FIGURE 6.10. A trace plot of the fourth chain of the adjusted $\hat{\tau}^2$ parameter of the juvenile delinquency data.

high as 15. The other parameter estimates show the same kind of pattern because as one parameter diverges, the others are affected as well. Thus, it seems that with this meta-analytic data set, the generic priors really help stabilize the parameter estimates. However, because they are weak, they do not have much influence over the values of parameter estimates of the posterior distribution.

Next, because the results of the first analysis indicated that the better fitting model is the unadjusted model that assumes publication bias is not present, to ensure proper model specification, I added a potentially important moderator (type of control group) into the

model and ran the MH algorithm one more time. I used the following starting values for the parameters:

- (1) $a = 1, b = 1, \beta_0 = -0.052, \beta_1 = -0.124, \tau^2 = 0.070$;
- (2) $a = 0.5, b = 1.5, \beta_0 = -0.052, \beta_1 = -0.124, \tau^2 = 0$;
- (3) $a = 0.8, b = 1.2, \beta_0 = 0.038, \beta_1 = 0.009, \tau^2 = 0.070$;
- (4) $a = 0.8, b = 1.2, \beta_0 = -0.142, \beta_1 = -0.257, \tau^2 = 0.070$.

I used the original generic priors for a and b and ran each chain with 100,000 iterations. With reasonable acceptance rates, ranging from 0.45 to 0.54, I moved on to assess plots of the parameter estimates. The trace and cumulative sum plots of the parameter estimates indicated convergence within 29,999 iterations, while bivariate trace plots also showed decent mixing among the parameter estimates of the bivariate posterior distributions.¹⁰ I cut the burn-in period and estimated PSR values of about 1.00 for all eight parameters, indicating convergence.

I reduced the autocorrelations of the chains to about 0.02 for the unadjusted parameters and ranging from 0.02 to 0.42 for the adjusted parameters by sub-sampling every 25th sampled value. That reduced the number of sampled values to 2,801 estimates per chain. I estimated the MCMC error of $\hat{\beta}_0$ and calculated the intervals for one standard error below and above the means of the unadjusted and adjusted $\hat{\beta}_0$ estimates. For the unadjusted $\hat{\beta}_0$, the intervals are the following for the four chains: $(-0.0507, -0.0468)$, $(-0.0530, -0.0490)$, $(-0.0516, -0.0477)$, and $(-0.0507, -0.0469)$. For the adjusted $\hat{\beta}_0$, the intervals are: $(-0.1519, -0.1452)$, $(-0.1435, -0.1365)$, $(-0.1497, -0.1432)$, and $(-0.1373,$

¹⁰Note that bivariate trace plots that included parameter a again displayed $\hat{a}$ values that congregated around 1.0. However, because of the restriction placed on a , I am not concerned. Moreover, the second analysis (with normal priors placed on a and b) showed that the restriction helps in estimating appropriate values by allowing the estimates to converge to stable values.

–0.1308). The intervals of the chains for the unadjusted $\hat{\beta}_0$ overlap a lot, while the intervals for that adjusted $\hat{\beta}_0$ are very close to one another (if not overlapping). That indicates the estimates converged to approximately the same values.

Results

I first assessed the random-effects meta-analysis of the juvenile delinquency effects for publication bias. Based on the final set of values, the *BIC* and *AIC* for the unadjusted model are –17.66 and –20.25, respectively, across the four chains. For the adjusted model, the *BIC* and *AIC* are –10.49 and –15.68, respectively. The values are lower for the unadjusted model, thus, the unadjusted model is the better fitting model. Moreover, the results indicate that publication bias has not been found in the juvenile delinquency effects. With posterior means of 0.90 and 1.00 for $\hat{a}$ and $\hat{b}$, respectively, these results are not surprising.

Next, because the authors noted that the type of control group included in a study may moderate the results of the meta-analysis, I included it as a predictor in the analysis in order to ensure proper model specification. The results for the mixed-effects analysis indicate that the *BIC* and *AIC* are again lower for the unadjusted model (–18.14 and –20.73, respectively) than for the adjusted model (–11.04 and –16.22, respectively). And posterior means of $\hat{a}$ and $\hat{b}$ are again close to 1.0 (0.90 and 1.03, respectively). Thus, publication bias has not been found in the juvenile delinquency effects once again. Moreover, note that the *BIC* and *AIC* are very similar for the unadjusted and adjusted models between the random-effects and mixed-effects analyses. This suggests that the type of control group present in the study may not matter. In that case, one can think parsimoniously and choose the unadjusted mean-only model as the “best” fitting model. That would indicate that the most likely SMD is about –0.10 for juvenile delinquency, which indicates an increase in crime after juveniles are processed through the court system. Moreover, 95% of the SMDs will lie between –0.25 and 0.05.

Discussion

Obtaining the posterior distributions of the parameters of the juvenile delinquency effects was relatively simple. However, the key to that success may be attributed to the generic priors placed on a and b . Those priors kept all the parameter estimates from diverging to nonsensical values, allowing the values to converge. Moreover, while both the likelihood-ratio test from ML estimation and the BIC and AIC from Bayesian estimation allow one to conclude that bias has not been found in this data set, only the priors in the Bayesian estimation keep the adjusted parameter estimates from diverging to nonsensical values. Note that this is an example of a data set for which the ML implementation of the beta density weight-function model may not perform well, according to the simulations in Chapters 4 and 5. That is because 1) power for the likelihood-ratio test for publication bias is low for small data sets, and 2) the adjusted estimates are poorer for data sets with a large amount of heterogeneity (especially for small data sets). Thus, it is best to take heed when making conclusions about the data. Look at a funnel plot of the data and see whether the unadjusted or adjusted mean estimate makes sense for that particular data set.

Teacher Expectancy Data

The teacher expectancy data, or TE, is from Raudenbush's (1984) meta-analysis on the effects of teacher expectancy on pupil IQ. The meta-analysis examined 19 SMD effects in order to assess whether teachers' expectations of how well their students are supposed to do (i.e., their alleged potential for growth) influences students' intelligence. The mechanism through which Raudenbush believed this would happen is how much contact the teachers had with their students prior to the expectancy induction. He believed that more contact would lead the teachers to be less likely to change their expectations due to the experimental treatment (being told which students have greater potential for growth) because more contact means more stable prior expectations. He examined this by splitting

FIGURE 6.11. An extended funnel plot of the teacher expectancy effects against their standard errors.

the data into two groups: the low contact group whose teachers had less than or equal to two weeks of contact with their students and the high contact group whose teachers had more than two weeks of contact with their students prior to the expectancy induction. I will first run a mean-only model when assessing the teacher expectancy effects meta-analysis for publication bias. Because the moderator is crucial to the model, I will follow that up with a mixed-effects analysis.

For the random-effects ML analysis, the unadjusted ML mean effect size $\hat{\mu}$ is 0.08 and $\hat{\tau}^2 = 0.01$, the latter indicating a substantial amount of heterogeneity ($Q = 35.83$, $df = 18$, $p = 0.01$; $I^2 = 49.76\%$). The adjusted ML estimates are 1.61, 1.53, 1.33, and 0.53 for $\hat{a}$, $\hat{b}$, $\hat{\mu}$, and $\hat{\tau}^2$, respectively. With an upward mean adjustment from 0.08 to 1.33

FIGURE 6.12. A plot of the beta density for the teacher expectancy data with $\hat{a}$ and $\hat{b}$ ML estimates of 1.61 and 1.53, respectively.

and a barely significant likelihood-ratio test ($\chi^2 = 5.82$, $df = 2$, $p = 0.05$), the results (at first glance) suggest the opposite of publication bias is present. Note, however, that the studies that are most likely to be missing are those that show a teacher expectancy effect. That may be because the literature wants to show that the teacher expectancy effect does not occur, thus, those studies that do show an effect (positive effects) could be missing. In that case, the mean effect size *should* be adjusted upwards, indicating reverse bias. See Figure 6.12 for a plot of the beta density for the teacher expectancy data, which shows what reverse bias looks like in this data set. But what is irrational about the adjusted mean estimate is the fact that the mean is larger than every observed effect size. That is, the adjustment is too strong.

FIGURE 6.13. An extended funnel plot of the teacher expectancy effects dichotomized by the amount of contact teachers had with their students prior to the expectancy induction.

For the mixed-effects ML analysis, a different problem arises: depending on what starting values one uses when estimating the adjusted parameter estimates, the estimates converge to different values. The unadjusted ML conditional mean effect sizes for the low and high contact groups are $\hat{\mu}_1 = 0.20$ and $\hat{\mu}_2 = -0.06$, respectively (see Figure 6.13 for a funnel plot of the effects dichotomized by the amount of student-teacher contact). $\hat{\tau}^2$ is 0, which indicates homogeneity of the effects ($Q = 22.82$, $df = 17$, $p = 0.16$) now that teacher-student contact has been taken into account. The two different sets of adjusted ML estimates are: 1) 1.61, 2.97, 0.89, -2.28, and 0.16 for $\hat{a}$, $\hat{b}$, $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\tau}^2$, respectively; and 2) 0.32, 2.98, -0.14, -0.53, and 0.004, for $\hat{a}$, $\hat{b}$, $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\tau}^2$, respectively. Look at Figure

FIGURE 6.14. A plot of the beta density with the results of two different ML mixed-effects analyses of the teacher expectancy data.

6.14 and note that the two sets of results appear to paint very different pictures of what the relative likelihood of being published in the teacher expectancy meta-analysis given prior teacher-student contact may be. However, because of the lack of effects with smaller p -values, there is not much information about the region on the left-hand side of the plot; thus, it may be too difficult to model selection in that region. Moreover, the right-hand side of the plot is not that different for the two sets of results. But the log-likelihood of the first set of results (-24.92) is smaller than that of the second set (-24.01), indicating that the second set of results has not reached its optimal estimates. Nonetheless, note that neither set of results appears to make sense. The first set of ML mixed-effects results adjusts the

mean of the first group in the expected direction (although quite high – from 0.20 to 0.89), while the mean of the second group is adjusted down by an irrational amount (from -0.06 to -2.28 – much lower than the smallest observed effect size of -0.32). The second set of results adjusts both group means down, which is not consistent with the presence of reverse publication bias. I hope that Bayesian estimation will display more consistent and appropriate results.

Procedure

In order to sample from the posterior distribution of the TE effects, I ran four separate sets of chains of the MH algorithm with the following starting values for the unadjusted and adjusted parameters (with a and b set to 1.0 in the unadjusted model):

- (1) $a = 1, b = 1, \mu = 0.078, \tau^2 = 0.013$;
- (2) $a = 0.5, b = 1.5, \mu = 0.078, \tau^2 = 0$;
- (3) $a = 0.8, b = 1.2, \mu = 0.130, \tau^2 = 0.013$;
- (4) $a = 0.8, b = 1.2, \mu = 0.026, \tau^2 = 0.013$.

I first ran the chains using the generic priors for a and b and found that the estimates of the parameters did not converge. Given that I forced a to be below 1.0, that makes sense. Remember that reverse bias may be present in this data set, which would require a to be above 1.0 in order to model it. Consequently, I re-ran the MH algorithm with normal priors on both a and b (priors for a and $b = N(1, 0.4^2)$), as well as a weak normal prior of $N(0.5, 0.4^2)$ on the adjusted τ^2 in order to keep it stable (trial runs of the algorithm indicated large adjusted τ^2 divergence when a prior was not placed on it). Using the same starting values, I ran the chain over 100,000 iterations. The acceptance rates were well within the recommended range, ranging from 0.40 to 0.58. The trace, cumulative sum, and bivariate trace plots all indicated that the parameter estimates converged to stable values. The values stabilized by about 30,000 iterations so I cut the first 29,999 iterations

out as the burn-in period. I then estimated *PSR* values of approximately 1.00 for all six parameters, again indicating convergence.

I sub-sampled every 25th sampled value from each chain in order to reduce the autocorrelations of each chain. That reduced the autocorrelations to between 0 and 0.08 for the unadjusted parameters and between 0.08 and 0.37 for the adjusted parameters. It also reduced that number of sampled values to 2,801 per chain. Based on those final sampled values, I assessed convergence once again by estimating the MCMC error of $\hat{\mu}$ and calculating the intervals for one standard error below and above the means of the unadjusted and adjusted $\hat{\mu}$. The intervals for the unadjusted $\hat{\mu}$ are: (0.0922, 0.0947), (0.0922, 0.0947), (0.0928, 0.0953), and (0.0909, 0.0934). The intervals for the adjusted $\hat{\mu}$ are: (0.8522, 0.8846), (0.8070, 0.8388), (0.7486, 0.7815), and (0.7881, 0.8203). The intervals for the unadjusted $\hat{\mu}$ overlap almost completely, while the intervals for the adjusted $\hat{\mu}$ are very close to one another (if not overlapping). Running the MH algorithm longer may help the adjusted $\hat{\mu}$ estimates converge to even more consistent values, but overall the estimates are close to the same value.

Next, I incorporated the teacher-student contact moderator into the model and ran a mixed-effects analysis. I ran the MH algorithm with the following starting values for the parameters:

- (1) $a = 1, b = 1, \beta_0 = 0.201, \beta_1 = -0.264, \tau^2 = 0$;
- (2) $a = 0.5, b = 1.5, \beta_0 = 0.201, \beta_1 = -0.264, \tau^2 = 0$;
- (3) $a = 0.8, b = 1.2, \beta_0 = 0.277, \beta_1 = -0.167, \tau^2 = 0$;
- (4) $a = 0.8, b = 1.2, \beta_0 = 0.125, \beta_1 = -0.361, \tau^2 = 0$.

I again used the normal priors for a and b and put a moderately weak normal prior of $N(0.05, 0.2^2)$ on the adjusted τ^2 in order to keep it from diverging to nonsensical values (prior runs of the algorithm indicated large adjusted τ^2 divergence when a prior was not included). I ran each chain with 150,000 iterations and observed adequate acceptance

rates ranging from 0.46 to 0.67. The trace, cumulative sum, and bivariate trace plots displayed very nice convergence to stable values. I cut the first 39,999 iterations as the burn-in period and calculated the *PSR* factor on the remaining 110,001 iterations. With *PSR* values of about 1.00 for all eight parameters, convergence has been reached.

I reduced the autocorrelations of each chain to between 0.07 to 0.28 for the unadjusted parameters and between 0.16 and 0.53 for the adjusted parameters by subsampling every 25th sampled value. That reduced the number of sampled values to 4,401 per chain. I then estimated the MCMC error of $\hat{\beta}_0$ and calculated the intervals for one standard error below and above the means of the unadjusted and adjusted $\hat{\beta}_0$. The intervals for the unadjusted $\hat{\beta}_0$ are: (0.2417, 0.2440), (0.2386, 0.2410), (0.2417, 0.2440), and (0.2401, 0.2424). The intervals for the adjusted $\hat{\beta}_0$ are: (0.8472, 0.8616), (0.8467, 0.8615), (0.8481, 0.8624), and (0.8358, 0.8500). With generous overlap in the intervals of both the unadjusted and adjusted $\hat{\beta}_0$, the intervals suggest convergence.

Results

I assessed the random-effects teacher expectancy meta-analysis for publication bias by calculating the *BIC* and *AIC* for both the unadjusted and adjusted models. Across the four chains, the *BIC* and *AIC* are -21.66 and -23.55 for the unadjusted model, respectively, and -22.06 and -25.84 for the adjusted model, respectively. With lower model fit values for the adjusted model, the results suggest that the adjusted model is the better fitting model. That also indicates that publication bias may be present in TE effects. With posterior means of 1.48 for both $\hat{a}$ and $\hat{b}$, the results actually suggest the presence of reverse bias (just as in the ML random-effects analysis). However, unlike the ML analysis, the posterior mean of $\hat{\mu}$ was adjusted to a more reasonable (and less extreme) estimate of 0.82 from an unadjusted posterior mean estimate of 0.09.

I then examined the teacher expectancy effects for publication bias after taking teacher-student contact into account. The results again suggest that bias may be present in the

effects, with lower *BIC* and *AIC* values for the adjusted model (-33.57 and -37.35 , respectively) than for the unadjusted model (-32.65 and -34.54 , respectively). The posterior means for $\hat{a}$ and $\hat{b}$ are 1.47 and 1.70 , respectively, indicating the presence of reverse bias. The posterior mean of the low contact group $\hat{\mu}_1$ was adjusted from 0.24 to 0.85 . The posterior mean of the high contact group $\hat{\mu}_2$ was adjusted from -0.07 to -0.73 . That suggests that for the low contact group the teacher expectancy effect should be bigger than what one observes from this meta-analysis, while for the high contact group there should be a reverse teacher expectancy effect (rather than a null effect). Moreover, note that the *BIC* and *AIC* are lower for the mixed-effects analyses than for the random-effects analyses. That indicates the teacher-student contact moderator is important in this model. The SMD of teacher expectancy for the low contact group will most likely be around 0.85 , with 95% of the SMDs lying between -0.02 and 1.84 . The SMD of teacher expectancy for the high contact group will be about -0.73 , with 95% of the SMDs lying between -3.27 and 1.45 .

Discussion

Sampling from the posterior distribution of the teacher expectancy effects was harder than from any of the other data sets included in this study. There are a number of reasons for this. First, the data set is small so it does not contain much information. This is even more problematic in the adjusted mixed-effects analysis where five parameters are to be estimated from only 19 effects. Also, this data set is different in that it may contain reverse bias as studies that obtain the teacher expectancy effect appear to be missing. That is why the original priors I put on a and b ($a = \chi^2(10 - 10(\text{value}), 2)$ and $b = N(1, 0.4^2)$) did not allow the parameter estimates to converge. Nonetheless, Bayesian estimation did a decent job of obtaining the parameter estimates. While some of the adjusted mean effects may still be adjusted too much, they appear much more reasonable than the ML adjusted mean effects. Stronger priors may be added in the future to keep a and b from adjusting too much.

Conclusions

In this chapter, my goal was to investigate how useful it would be to implement the beta density weight-function model via Bayesian estimation. From the above analyses, I believe this implementation is very useful. When ML estimation has no trouble estimating the adjusted parameter estimates, Bayesian estimation does an equally good job of estimating the parameters. In cases where ML estimation produces nonsensical parameter estimates, Bayesian estimation seems to do a better job than ML estimation of adjusting to more appropriate estimates (e.g., behavioral marital therapy data, juvenile delinquency data, and teacher expectancy data). This is especially helpful for publication bias estimation with small meta-analytic data sets where there is not much information about the data. Priors can provide that extra information and usually help stabilize the parameter estimates.

Moreover, often, one worries about whether an estimate has stabilized to an appropriate value. With real data, we do not know the truth about the population, so we cannot know whether the adjusted mean estimate is the appropriate one. However, we sometimes do know when an estimate is inappropriate. For example, with the behavioral marital therapy data, it is obvious that a mean adjustment from 0.58 to -0.99 does not make sense. The adjustment is simply too extreme because 1) the smallest observed effect is 0.07 and 2) an adjustment of 271% does not seem reasonable. Thus, one can tell whether an estimate does not make sense by simply looking at the data. This is comforting as one does not want to make conclusions (or decisions) based on unreasonable estimates.

CHAPTER 7

Bayesian Simulations

The preferred method of model assessment is via simulation, rather than through use of the model on real data. That is because in a simulation one knows the truth about the population, allowing for comparison of the model estimates with the true population values. In Chapters 4 and 5, I assessed the beta density weight-function model's utility using simulations when the model is implemented via maximum-likelihood (ML) estimation. In this chapter, I evaluate the model's performance when it is implemented using Bayesian estimation via the Metropolis-Hastings (MH) sampling algorithm, a Markov chain Monte Carlo (MCMC) method (see, e.g., Chib & Greenberg, 1995; Geyer, 1991; Lynch, 2007).

Note that this simulation study is very small because it consists of doing simulations on simulations. That is, I use simulated data in order to sample (or simulate values) from the posterior distribution of each of the simulated data sets. This is very demanding of both computer and human resources. Thus, I implemented only four simulations, focusing mostly on small data sets. I focus on small data sets because I have already seen the model work well with large data sets (via ML estimation). Part of the goal is to see whether a Bayesian implementation of the model will be particularly useful for small meta-analytic data sets.

Method

In order to assess the utility of a Bayesian implementation of the beta density weight-function model, I conducted a small simulation study that consists of repeatedly sampling

from the posterior distributions of a set of simulated data sets. I describe that process below.

The Data

Instead of generating new data, I used the already simulated data from the ML mean-only simulations in Chapter 4 for this simulation study. I chose three cells of that study to work with (or sample from):

- Cell 2a: no bias, $k = 30$, and $\tau^2 = 0.05$;
- Cell 7b: small bias, $k = 90$, and $\tau^2 = 0.09$;
- Cell 4d: severe bias, $k = 30$, and $\tau^2 = 0.14$.

I chose those three cells in order to represent variations in both heterogeneity and severity of publication bias. Notice that I focus on smaller data sets, including only one cell that contains moderately-sized data sets (Cell 7b). The reason for that is simple: I want to determine whether the flexibility of Bayesian estimation can help improve the performance of the beta density weight-function model with small data sets, as the model already appears to perform well with large data sets (according to the ML simulations). In addition, because some might argue that a data set of 30 effects is not that small, I created a new cell for the study that consists of 600 simulated data sets with the following characteristics: moderate bias, $k = 15$, and $\tau^2 = 0$. I generated this biased data using the same method I used to generate the data for Chapter 4's ML mean-only simulations.

Procedure

In order to assess how well, on average, the Bayesian implementation of the beta density weight-function model performs on small to medium-sized data sets, I ran the model on numerous replications of each of the cells of simulated data. Remember that each replication within a cell represents one simulated data set. Thus, I am essentially running the model on numerous data sets sampled from the same population (or cell). I can then see how well the model adjusts for publication bias (if any exists), on average,

for that population (or meta-analytic data with specific characteristics). For Cells 2a and 4d (where $k = 30$), I randomly sampled 400 data sets (or replications) on which to run the model. For Cell 7b (where $k = 90$), I randomly sampled 200 data sets. For the new cell (where $k = 15$), I simulated 600 data sets and ran the model on all of them.

For each data set within each cell, I used the MH algorithm (described in Chapter 3's section *Bayesian Estimation*) to sample values for each parameter from the posterior distribution using the log-likelihood function in Equation 3.4 as the likelihood. I used the full log-likelihood function (i.e., including the selection model) to obtain the sampled values for the adjusted model parameters. In order to obtain values for the unadjusted model parameters, I first set a and b to 1.0, as that allows for the estimation of the effect-size model without selection. I ran both models as mean-only random-effects models since including τ^2 allows for the generalization of the results to a hypothetical population of studies.

Recall that in Bayesian estimation, the posterior density is the product (or summation, if one is working in the log metric) of the likelihood and prior beliefs. Because I did not want to impose strong opinions about the presence of publication bias on the data sets (even though I know how much bias each data sets contains because I simulated the data), I used the weak generic priors from Chapter 6 as my prior beliefs. In order to keep a below one, as values of a above one do not make sense under most circumstances, the prior for a was based on a χ^2 distribution with the posterior density equal to $\chi^2(10 - 10(value), 2)$ with 2 df . The prior for b was set to a normal distribution with the mean centered at one and the standard deviation set to 0.4 [i.e., $N(1, 0.4^2)$], as that keeps b within a reasonable range.

I ran the MH algorithm once for each data set within each cell with the starting values for both the unadjusted and adjusted μ and τ^2 set to the true population values (i.e., the cell's generating values) in order to save computer time and reduce the burn-in period.

Because the true values are unknown for a and b , the starting values for a and b were set to 1.0 in the adjusted model. In the unadjusted model, a and b were set to 1.0 permanently and not estimated. I ran each chain of parameter values for each of the data sets with $k = 30$ and 90 (Cells 2a, 7b, and 4d) with 10,000 iterations and the data sets with $k = 15$ (new cell) with 30,000 iterations. These simulations on simulations were run in *R* (Version 2.7.1) on a Sun machine using the Solaris 10 operating system.

Convergence

Before analyzing the results, I first assessed whether the chains converged to stable values. Because it is not feasible to look at thousands of trace and cumulative sum plots to assess convergence (each data set contains six chains of sampled parameter values and there are 200, 400, or 600 data sets per cell), I relied on the *PSR* factor. The *PSR* factor determines whether each chain of a parameter has converged to approximately the same value by assessing the reduction of between-chain variance for the set of chains run for that parameter (Lynch, 2007). Normally, the *PSR* factor is calculated on the multiple chains of each parameter for one data set. In this simulation study, each cell contained 200, 400, or 600 data sets. Within a cell, those data sets were all sampled from the same population. Thus, I treated each cell as one data set with multiple runs of chains. I then calculated the *PSR* factor on the 200, 400, or 600 chains of each parameter for each cell.

For Cell 2a, I obtained *PSR* values ranging from 1.03 to 1.37, with the adjusted model parameters obtaining lower *PSR* values than the unadjusted model parameters. For Cell 7b, the *PSR* values ranged from 1.16 to 1.44. For Cell 4d, the *PSR* values ranged from 1.20 to 1.52. For the new cell, the *PSR* values ranged from 1.01 to 1.23. Almost all the values are outside of *Mplus* User's Guide boundaries of 0.95 and 1.05 (Muthén & Muthén, 2010), indicating the parameters have not converged. Subsequently, I looked at a number of trace plots of the parameters for each of the cells. Given the huge divergence of values in many of the plots, the results are not surprising.

FIGURE 7.1. An example of a bad trace plot of the adjusted $\hat{\mu}$ parameter of one data set from the new cell.

If such a result were to occur in an MCMC run with real data, the solution would be to throw out the non-converged data and re-run the sampling algorithm after adjusting the starting values or priors (or both). However, from what the results indicate, it may not make sense to run the beta density weight-function model via the MH algorithm in this (simulation) context because it may not be possible to find generic starting values or priors that are appropriate for all data sets to converge. (Remember that small data sets are the conditions under which the model may have the most trouble estimating the parameters.) And it is not feasible to find specific priors and starting values for each of the data sets. Thus, instead of re-running the simulations, I examined trace plots of the adjusted $\hat{\mu}$ parameter for each of the cells in order to find the first 20 data sets that indicate convergence.

FIGURE 7.2. An example of a good trace plot of the adjusted $\hat{\mu}$ parameter of one data set from the new cell.

I decided to seek the first 20 converged data sets because 1) it is not feasible to look at all the adjusted $\hat{\mu}$ trace plots and 2) 20 seems like a practical number of data sets to find in order to summarize the results of each cell. Moreover, it is acceptable to look at just one of the parameters for convergence because when one parameter diverges, the others tend to do so as well due to the correlations in the multivariate sampling distribution. When examining convergence, I avoided keeping data sets whose adjusted $\hat{\mu}$ trace plots displayed systematic up or down movement of the estimates. Figure 7.1 shows an example of a bad adjusted $\hat{\mu}$ trace plot displaying this phenomenon. Figure 7.2 shows an example of a good trace plot that indicates the adjusted $\hat{\mu}$ has converged.

As I examined the trace plots of the adjusted $\hat{\mu}$, I noted the burn-in period for each data set in order to cut that first non-converged part out of the final set of values. For each cell, I also looked at a number of g lag versus autocorrelation plots in order to determine the best g value with which I could sub-sample (or thin) the chains of parameter values. Recall that in MCMC sampling, the values within a Markov chain are dependent on one another because movement from one value to the next is represented by a stationary transition probability (Geyer, 1991). Sub-sampling (or chain thinning) is done to reduce the autocorrelations within each chain (Lynch, 2007). For Cells 2a and 4d, I sub-sampled every 15th value, and for the new cell and Cell 7b, I sub-sampled every 20th value. Across the four cells, that reduced the autocorrelations to close to zero for the unadjusted parameters and no larger than 0.50 for the adjusted parameters. Cutting the burn-in period and thinning the chains also reduced the final sets of parameter values sampled for each data set. For Cell 2a, the data sets consisted of 267 to 667 estimates per chain. For Cell 7b, the number of final estimates ranged from 300 to 500 values per chain. For Cell 4d, the chains ranged from 467 to 667 sampled values in a data set. For the new cell, the data sets contained 1,000 to 1,500 estimates per chain.

Results

In order to assess how well the beta density weight-function model adjusts the simulated meta-analytic data with various characteristics for publication bias, I first estimated the *BIC* and *AIC* for each data set to examine which model (the unadjusted model that assumes publication bias is not present or the adjusted model that includes a weight function that models publication bias) fits the data better. The model that obtains a lower *BIC* or *AIC* is considered the better fitting model (Congdon, 2003; Lynch, 2007). I also calculated the posterior means and credibility intervals (CIs) for each of the parameters in order to see what the distribution of each parameter looks like. I then averaged those

TABLE 7.1. Model Fit Statistics for the Four Cells

Model	Cell 2a		Cell 7b		Cell 4d		New Cell	
	<i>BIC</i>	<i>AIC</i>	<i>BIC</i>	<i>AIC</i>	<i>BIC</i>	<i>AIC</i>	<i>BIC</i>	<i>AIC</i>
Unadjusted	-13.898	-16.700	-56.925	-61.924	-14.178	-16.981	-19.329	-20.745
Adjusted	-6.680	-12.284	-47.748	-57.747	-7.401	-13.006	-14.065	-16.897

results across the 20 data sets in order to see how well the beta density weight-function model performs on average.

For all four cells, the model fit indices in Table 7.1 indicate that the better fitting model is the unadjusted model, as the unadjusted model obtains consistently lower values for the *BIC* and *AIC*. This indicates that evidence of publication bias has not been found in any of the cells. However, bias is present in all but one of the cells (Cell 2a). This suggests that the beta-density weight-function model is good at avoiding the conclusion that publication bias is present when it is actually not, but it is not good at leading one to the correct model when publication bias is present. This is akin to finding in the ML mean-only simulation study where the likelihood-ratio test lacked power to detect publication bias in small data sets. The posterior means of the parameters tell a different story.

Table 7.2 shows the results for the posterior means of the parameters for both the unadjusted and adjusted models. For Cell 2a (no bias, $k = 30$, and $\tau^2 = 0.05$), the posterior means of the unadjusted parameters are indeed closer to the true population values of 0.20 for μ and 0.05 for τ^2 (0.23 versus 0.08 for $\hat{\mu}$ and 0.057 versus 0.063 for $\hat{\tau}^2$). With values of 0.90 and 1.25 for $\hat{a}$ and $\hat{b}$, respectively, the adjusted model adjusts $\hat{\mu}$ down too much. Given that the *BIC* and *AIC* correctly indicate that the unadjusted model is the better fitting model, this is not a concern, as one should use the unadjusted parameter estimates to summarize Cell 2a's distribution. For the three cells that contain bias, the results for $\hat{\mu}$ are better (i.e., closer to the true mean) for the adjusted model than for the unadjusted model (even though the fit indices suggest otherwise). For Cell 7b (small bias,

TABLE 7.2. Averaged Posterior Means of the Parameter Estimates of the Unadjusted and Adjusted Models for the Four Cells

Parameter	Cell 2a	Cell 7b	Cell 4d	New Cell
True μ	0.20	0.20	0.20	0.20
True τ^2	0.05	0.09	0.14	0.00
Unadjusted $\hat{\mu}$	0.234	0.384	0.554	0.338
Unadjusted $\hat{\tau}^2$	0.057	0.069	0.071	0.031
Adjusted $\hat{a}$	0.896	0.920	0.884	0.893
Adjusted $\hat{b}$	1.252	1.057	1.077	1.382
Adjusted $\hat{\mu}$	0.079	0.297	0.429	0.204
Adjusted $\hat{\tau}^2$	0.063	0.064	0.069	0.039

$k = 90$, and $\tau^2 = 0.09$), $\hat{\mu}$ is adjusted down from 0.38 to 0.30. For Cell 4d (severe bias, $k = 30$, and $\tau^2 = 0.14$), $\hat{\mu}$ decreases from 0.55 to 0.43. For the new cell (moderate bias, $k = 15$, and $\tau^2 = 0$), $\hat{\mu}$ is adjusted down from 0.34 to 0.20.

Although it may not be appropriate to make generalizations across the cells as there are only four very different cells to compare, it seems that meta-analytic data with less heterogeneity (new cell) result in better adjusted parameter estimates than data with a larger amount of heterogeneity (Cell 4d). This finding is also similar to that of the ML mean-only simulation study. Moreover, heterogeneity appears to be more indicative of appropriate $\hat{\mu}$ adjustment than the size of the meta-analytic data, as it is the cell with the smallest data sets (new cell: $k = 15$) that adjusts $\hat{\mu}$ closest to the true mean. Last, note that the results for the cells that contain larger data sets are more consistent than those that contain smaller data sets, as the credibility intervals for Cell 7b are smaller than those for Cells 2a and 4d, and the CIs for Cells 2a and 4d are generally smaller than for the new cell (see Table 7.3).

Recall that a goal of this study is to assess whether a Bayesian implementation of the beta density weight-function model will be particularly useful for small meta-analytic data. I have assessed that by comparing the fit statistics and parameter estimates of the unadjusted and adjusted models. Another way to address that question is to see whether

TABLE 7.3. Credibility Intervals for the Parameter Estimates of the Unadjusted and Adjusted Models for the Four Cells

Parameter	Cell 2a		Cell 7b		Cell 4d		New Cell	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Unadjusted $\hat{\mu}$	0.103	0.370	0.306	0.462	0.420	0.696	0.184	0.507
Unadjusted $\hat{\tau}^2$	0.014	0.145	0.034	0.121	0.021	0.171	0.001	0.132
Adjusted $\hat{a}$	0.627	0.997	0.752	0.997	0.668	0.994	0.626	0.997
Adjusted $\hat{b}$	0.703	1.839	0.611	1.551	0.472	1.740	0.705	2.095
Adjusted $\hat{\mu}$	-0.317	0.316	0.086	0.448	0.125	0.636	-0.192	0.425
Adjusted $\hat{\tau}^2$	0.008	0.222	0.023	0.148	0.016	0.202	0.001	0.221

the Bayesian implementation of the model adjusts the parameters more appropriately than the ML implementation of the model. The unadjusted ML $\hat{\mu}$ estimates are 0.20, 0.38, 0.53, and 0.28 for Cells 2a, 7b, 4d, and New, respectively. The adjusted ML $\hat{\mu}$ estimates are 0.23, 0.36, 0.41, and 0.27 for Cells 2a, 7b, 4d, and New, respectively. For Cell 2a, where publication bias is not present, the ML adjusted $\hat{\mu}$ is a much more appropriate estimate than the Bayesian adjusted $\hat{\mu}$ (although in both cases one should use the unadjusted parameter estimates). However, for the cells where bias is present and heterogeneity is not large (i.e., $\tau^2 \leq 0.14$), the Bayesian adjusted $\hat{\mu}$ estimates are quite a bit closer to the true mean than the ML adjusted $\hat{\mu}$ estimates.

Conclusions

In this chapter, my goal was to determine whether a Bayesian implementation of the beta density weight-function model is useful for small meta-analytic data. The analyses indicate that this implementation is indeed useful. However, note that this very small simulation study may be misleading, as 20 runs of a cell is not sufficient to lead to stable conclusions. Thus, the results should be discerned with caution. When bias is not present, the fit statistics correctly identify the unadjusted model (whose parameter estimates are close to the true population values) as the better fitting model. When bias is present, the

adjusted parameter estimates are more appropriate (i.e., closer to the true values) than the unadjusted parameter estimates. This is comforting; however, the fit indices do not lead one to the correct model, suggesting that the unadjusted model is the better fitting model. It is not clear why this happens. Perhaps it is analogous to the lack of power that the likelihood-ratio test contains when assessing bias in small data sets.

I also examined whether the Bayesian estimates are more appropriate than the ML estimates for the four cells. When bias is not present, it does not matter which approach one uses as both lead one to the correct (unadjusted) model for which the parameter estimates are close to the true population values (although the Bayesian estimates are slightly larger). When bias is present, neither Bayesian nor ML estimation leads one to the correct model often. However, the adjusted Bayesian parameter estimates are usually closer to the true population values than the adjusted ML parameter estimates. This suggests that the Bayesian implementation of the beta density weight-function model may be more useful for small data sets than the ML implementation of the model.

CHAPTER 8

Concluding Remarks

Studies that are not published are often those that fail to find a significant effect or that demonstrate effects that are not in the desired or expected direction. The result of that phenomenon, referred to as publication bias, is a biased published literature (Coursol & Wagner, 1986; Veitch, 2006). Meta-analyses, or quantitative summaries of statistical estimates used to summarize a particular field, are affected by publication bias, as the studies available for inclusion in the analysis are mostly those that consist of significant effects (Gilbody et al., 2000; Rothstein et al., 2005). As these syntheses have become vital to informing policy makers (e.g., The Cochrane Collaboration, 2010, informs policy in medicine and The Campbell Collaboration, 2010, informs social policy), it is important that meta-analyses be as accurate as possible; otherwise biased meta-analyses may bring about misinformed policies. Researchers have developed numerous methods that share a common goal of assessing and correcting meta-analyses for publication bias (e.g., Begg & Mazumdar, 1994; Copas & Li, 1997; Dear & Begg, 1992; Duval & Tweedie, 2000a, 2000b; Egger et al., 1997; Hedges, 1984; Iyengar & Greenhouse, 1988; Rosenthal, 1979; Vevea & Hedges, 1995). Each method incorporates different sets of assumptions and allows for only specific kinds of meta-analytic data to be assessed. Those assumptions have left a gap in the publication bias methodology literature: no single method exists that 1) can account for continuous moderators by including them within the model, 2) allow for substantial data heterogeneity, 3) produce an adjusted mean effect size, 4) include a formal test for publication bias, and 5) allow for correction when only a small number of

effects is included in the analysis. The goal of this dissertation was to develop a method that could fill that gap.

The new model, referred to in this dissertation as the beta density weight-function model, uses the beta density as the weight function that estimates the selection process in order to adjust the meta-analytic parameter estimates for publication bias. The model was implemented in two ways: 1) maximum-likelihood estimation using an optimization routine, and 2) Bayesian estimation via the Metropolis-Hasting (MH) algorithm. The efficacy of both implementations of the model was assessed by simulations and through use on real data sets.

The ML simulations indicate that the model is good at distinguishing systematic study differences from publication bias. The likelihood-ratio (LR) test for publication will rarely cause one to conclude that publication bias is present when it is actually not, as the Type I errors are very close to their nominal values. However, the test is not very powerful for small data sets; thus, it does not lead one to the correct model often when the number of effects included in the analysis is small and bias is present. Nevertheless, coverage rates show that the true parameter values are *much* more likely to be caught within the 95% confidence intervals of the adjusted model's parameter estimates (those that correct for bias) than of the unadjusted model's parameter estimates (those that do not account for bias). Bias and root mean squared errors of the estimates are better for the adjusted mean effect than the unadjusted mean effect when any bias is present, suggesting less deviation from the true mean effect. For both the unadjusted and adjusted models, the statistics also indicate that the mean estimates deviate from the true mean effect more as the amount of heterogeneity present increases.

The Bayesian simulation study presents a very similar pattern of results to the ML mean-only simulation study, although the Bayesian results are based on a very limited number of simulated data sets. The lack of power to detect publication bias indicated by

the LR test in ML estimation is also displayed in Bayesian estimation by the smaller *BIC* and *AIC* values for the unadjusted model than the adjusted model when bias is present. On the other hand (and also mirroring ML findings), the results indicate that the adjusted mean estimates are closer to the true mean effect when less heterogeneity is present in the meta-analytic data.

Overall, the results suggest that the beta density weight-function model is useful in adjusting the parameter estimates of realistic meta-analytic data sets for publication bias. The model can accommodate continuous moderators, it allows for heterogeneity to be present in the meta-analysis, it produces adjusted parameter estimates that account for bias, it includes a formal test for publication bias, and it sometimes works well when only a small number of effects is included in the analysis. Specifically, the model can adjust well for publication bias even when there are few effects if too much heterogeneity is not present. Moreover, the Bayesian implementation of the model appears to be even more useful than the ML implementation when working with small meta-analytic data sets, indicated by less deviance from the true mean effect when bias is present and by more appropriate estimates with problematic data sets. That said, one should be cautious when interpreting the results from small meta-analytic data sets with a large amount of heterogeneity, as that is when the beta density weight-function model is weakest. It is important to plot your data and look at whether the unadjusted or adjusted mean estimate makes more sense, rather than blindly accept one value or the other according to the test for publication bias. It is also useful to assess and adjust the data set for publication bias using some of the other publication bias methods (as long as the data set meets the assumptions of that method) in order to compare the results.

References

- Becker, B. J. (2005). Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessments, and adjustments* (pp. 111–125). West Sussex, England: John Wiley and Sons, Ltd.
- Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4), 1088–1101.
- Biggerstaff, B. J., & Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, 16, 753–768.
- Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3), 167–174.
- Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4), 327–335.
- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.
- Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.
- Congdon, P. (2003). *Applied Bayesian modelling*. Wiley series in probability statistics. West Sussex, England: John Wiley and Sons, Ltd.
- Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 37(1), 131–146.
- Copas, J. B. (1999). What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 162(1), 95–109.

- Copas, J. B., & Li, H. G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society, Series B (Methodological)*, 59(1), 55–95.
- Copas, J. B., & Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research*, 10, 251–265. doi: 10.1177/096228020101000402.
- Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, 17(2), 136–137.
- Dear, K. B. G., & Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, 7(2), 237–245.
- DeGroot, M. H., & Schervish, M. J. (2002). *Probability and statistics (3rd ed.)*. Boston, MA: Addison-Wesley.
- Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessments, and adjustments* (pp. 11–33). West Sussex, England: John Wiley and Sons, Ltd.
- Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessments, and adjustments* (pp. 127–144). West Sussex, England: John Wiley and Sons, Ltd.
- Duval, S., & Tweedie, R. (2000a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449), 89–98.
- Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463.

- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629–634.
- Else-Quest, N. M., Hyde, J. B., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender difference in temperament: A meta-analysis. *Psychological Bulletin*, 132(1), 33–72. doi: 10.1037/0033-2909.132.1.33.
- Galbraith, R. F. (1988). Graphical display of estimates having differing standard errors. *Technometrics*, 30(3), 271–281.
- Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In E. M. Keramidas (Ed.), *Computing science and statistics: Proceedings of the 23rd symposium on the interface* (pp. 156–163). Fairfax Station, VA: Interface Foundation of North America.
- Gilbody, S. M., Song, F., Eastwood, A. J., & Sutton, A. (2000). The causes, consequences and detection of publication bias in psychiatry. *Acta Psychiatrica Scandinavica*, 102(4), 241–249.
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82(1), 1–20.
- Harbord, R. M., Egger, M., & Sterne, J. A. C. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, 25, 3443–3457. doi: 10.1002/sim.2380.
- Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Education Statistics*, 9(1), 61–85.
- Hedges, L. V. (1992a). Meta-analysis. *Journal of Education Statistics*, 17(4), 279–296.
- Hedges, L. V. (1992b). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2), 246–255.

- Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6(3), 203–217.
- Hedges, L. V., & Vevea, J. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessments, and adjustments* (pp. 145–174). West Sussex, England: John Wiley and Sons, Ltd.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486–504.
- Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3(1), 109–135.
- Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions – 2: Distributions in statistics*. Wiley series in probability and mathematical statistics. New York, NY: John Wiley and Son, Inc.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications, Inc.
- Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer series in statistics for social and behavioral sciences. New York, NY: Springer Science + Business Media, LLC.
- Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20, 641–654.
- Mooney, C. Z. (1997). *Monte Carlo simulation*. Sage series in quantitative applications in the social sciences (vol. 116). Thousand Oaks, CA: Sage Publications, Inc.
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide (6th ed.)*. Los Angeles, CA: Muthén & Muthén.
- National Research Council. (1992). *Combining information: Statistical issues and research opportunities*. Washington, DC: National Academy Press.

- Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, 41(11), 1299–1301.
- Nocedal, J., & Wright, S. J. (2006). *Numerical optimization (2nd ed.)*. Springer series in operations research and financial engineering. New York, NY: Springer Science + Business Media, LLC.
- Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8(2), 157–159.
- Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, 295(6), 676–680.
- Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, 26, 4544–4562. doi: 10.1002/sim.2889.
- Petrosino, A., Turpin-Petrosino, C., & Guckenburg, S. (2010). Formal system processing on juveniles: Effects of delinquency. *Campbell Systematic Reviews*, 1, doi: 10.4073/csr.2010.1.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in Fortran 77: The art of scientific computing (2nd ed.)*. Cambridge, MA: Cambridge University Press.
- R. (Version 2.7.1). [Computer software]. Vienna, Austria: The R Foundation for Statistical Computing. Available from <http://www.r-project.org/>
- R. (Version 2.9.2). [Computer software]. Vienna, Austria: The R Foundation for Statistical Computing. Available from <http://www.r-project.org/>
- Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of credibility of expectancy induction: A synthesis of findings from 18

- experiments. *Journal of Educational Psychology*, 76(1), 85–97.
- Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, 10(2), 75–98.
- Rice, J. A. (2007). *Mathematical statistics and data analysis (3rd ed.)*. Belmont, CA: Brooks/Cole, Cengage Learning.
- Rosenthal, R. (1979). The ‘file drawer problem’ and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.
- Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 55(1), 33–38.
- Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, 15(2), 570.
- Rothstein, H. R. (2008). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology*, 4(1), 61–81. doi: 10.1007/s11292-007-9046-9.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessments, and adjustments* (pp. 1–7). West Sussex, England: John Wiley and Sons, Ltd.
- Schwarzer, G., Antes, G., & Schumacher, M. (2007). A test for publication bias in meta-analysis with sparse binary data. *Statistical Medicine*, 26, 721–733. doi: 10.1002/sim.2588.
- Shadish, W. R., & Baldwin, S. A. (2005). Effects of behavioral marital therapy: A meta-analysis of randomized control trials. *Journal of Consulting and Clinical Psychology*, 73(1), 6–14.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical*

Association, 54(285), 30–34.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessments, and adjustments* (pp. 75–98). West Sussex, England: John Wiley and Sons, Ltd.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessments, and adjustments* (pp. 99–110). West Sussex, England: John Wiley and Sons, Ltd.

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, 53(11), 1119–1129.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M. (1949). *The American Soldier: Adjustment during Army life*. Studies in social psychology in World War II (vol. 1). Princeton, NJ: Princeton University Press.

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

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22(13), 2113–2126. doi: 10.1002/sim.1461.

The Campbell Collaboration. (2010). *About us*. Available from <http://www.campbellcollaboration.org/aboutus/index.php>

The Cochrane Collaboration. (2010). *About us*. Available from <http://www.cochrane.org/about-us>

Thisted, R. A. (1988). *Elements of statistical computing*. New York, NY: Chapman and Hall.

- Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, 18, 2693–2708.
- Tufte, E. R. (2006). *Beautiful evidence*. Cheshire, CT: Graphics Press LLC.
- Veitch, E. (2006). Introducing PLoS Clinical Trials: A different approach to publishing trial results. *PLoS Clinical Trials*, 1(1), 1. doi: 10.1371/journal.pctr.0010008.
- Vevea, J. L., Clements, N. C., & Hedges, L. V. (1993). Assessing the effects of selection bias on validity data for the general aptitude test battery. *Journal of Applied Psychology*, 78(6), 981–987.
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419–435.
- Vevea, J. L., & Hedges, L. V. (1996). *Meta-analysis under publication bias: Compensating for selection in small data sets*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, 10(4), 428–443. doi: 10.1037/1082-989X.10.4.428.
- Wang, M. C., & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*, 3(1), 46–54.

Appendix

Many numerical estimation issues arise when one chooses to take on a project such as this one. In the Appendix, I discuss how I handled some of those issues. First, I will show how the biased data were generated for the maximum-likelihood (ML) simulations. Then I will illustrate how the beta density weight-function model is defined in *R*. Last, I will provide annotated *R* code that illustrates how the model may be run using both *R*'s optimizer *nlminb* and the Metropolis-Hastings (MH) algorithm. All *R* code will be represented by use of a `verbatim` font.

Data Generation

In order to generate data that represents sample sizes (n 's) present in the real world of psychology, I sampled n 's from a sample size distribution (defined as “Ns” in the *R* code below) that I created from a set of 10 meta-analytic articles. Based on those samples, I calculated SMD variances and generated observed effects. I then saved those effects with probabilities defined by a weight function that represents a specific amount of publication bias (no bias, small bias, moderate bias, or severe bias). The details of each of those steps are explained for the mean-only simulation in the annotated *R* code below.

```
#Create a matrix to save the data:  v, y, and p.  
save <- matrix(0,20*k,3)
```

```
#Simulate 20 x k effects over loop j.  Note that more than k  
effects need to be generated because not all the effects will be
```

```

accepted given the weight function.
for(j in 1:(20*k)) {

#Randomly sample one row of the sample size (n) data from the
sample size distribution (Ns).
n <- Ns[sample(nrow(Ns),1),]

#Calculate the SMD variance (v) on that sample.
v <- ((n$n1 + n$n2)/(n$n1*n$n2)) + ((n$es^2)/(2*(n$n1 + n$n2)))

#Generate an observed effect (y) by sampling y from a normal
distribution with mean equal to the true population mean (truemu)
and standard deviation equal to the square root of v plus a
variance component (vc).
y <- rnorm(1,truemu,sqrt(v+vc))

#Calculate the p-value (p).
p <- 1-pnorm(y/sqrt(v))

#Define p-value intervals that will represent the weight function.
pint = 1 #p < .01
if(p > .01 & p <= .05) pint = 2
if(p > .05 & p <= .10) pint = 3
if(p > .10 & p <= .20) pint = 4
if(p > .20 & p <= .30) pint = 5
if(p > .30 & p <= .50) pint = 6
if(p > .50 & p <= .70) pint = 7
if(p > .70 & p <= .85) pint = 8

```

```

if(p > .85) pint = 9

#Define a set weights that represent the probabilities with
which an observed effect size with a particular p-value will be
observed. (The set of weights defined here contains moderate
bias.)
w <- c(1.0,.70,.60,.50,.40,.30,.30,.20,.10)

#Save the data (y,v,p) if the probability of obtaining that data,
given its p-value, is greater than a random draw from the uniform
distribution.
if(runif(1) < w[pint]) save[j,1:3] = c(y,v,p) }

#After this process has run over 20 x k replications, save the
first k effects, variances, and p-values.
save <- save[rowSums(save) !=0, , drop=TRUE]
y <- save[1:k,1]
v <- save[1:k,2]
p <- save[1:k,3]

```

For the ML simulations that included a moderator, the code is very similar. The only difference is in the weight functions. For that set of simulations, I generated data for two unbiased populations with varying data characteristics and combined them to create one population that appeared biased when assessed by the funnel plot criterion. I wanted the first population to contain small effects with large sample sizes and small variances. Thus, I defined a weight function that saved data with large total sample sizes (defined as larger than the median sample size) with a higher probability:

```

if (ntot > nmed & runif(1) < .9) savepop1[j,1:3] = c(y,v,p)
else if (ntot < nmed & runif(1) < .1) savepop1[j,1:3] = c(y,v,p).

```

For the second population, I reversed the situation by defining a weight function that saved data with small total sample sizes (defined as smaller than the median sample size) with a higher probability, resulting in a larger proportion of effects that are large and contain large variance:

```

if (ntot < nmed & runif(1) < .9) savepop2[j,1:3] = c(y,v,p)
else if (ntot > nmed & runif(1) < .1) savepop2[j,1:3] = c(y,v,p).

```

For both populations, I saved $\frac{1}{2} \times k$ data so that when I combined the two populations, I had one biased data set with length of k . That means the populations were equally represented in combined data set.

The Beta Density Weight-Function Model in R

I now present the *R* code for the log-likelihood of the beta density weight-function model (Equation 3.4). I include the code for both random- and mixed-effects analyses.

```

#Mean-only random-effects model
denfxn <- function(yy,vv,a2,b2,vc2,mu2) {
  pp <- 1-pnorm(yy/sqrt(vv))
  pp <- ifelse(pp < .00001,.00001,pp)
  pp <- ifelse(pp > .99999,.99999,pp)
  part1 <- pp^(a2-1)
  part2 <- (1-pp)^(b2-1)
  part3 <- dnorm(yy,mu2,sqrt(vv+vc2))
  return(part1*part2*part3) }

like <- function(pars) {
  a2 <- pars[1]

```



```

b2 <- pars[2]
vc2 <- pars[3]
mu2 <- pars[4]
num <- sum((a2-1)*log(p) +
           (b2-1)*log(1-p) -
           1/2*log(v+vc2) - 1/2*((y-mu2)^2/(v+vc2)))
total <- rep(0,length(y))
for(i in 1:length(y)) {
  vv = v[i]
  yy = y[i]
  answer <- integrate(denfxn,-Inf,Inf,vv=vv,a2=a2,b2=b2,
                     vc2=vc2,mu2=mu2)
  bottom <- .00001*denfxn(-Inf,vv,a2,b2,vc2,mu2)
  top <- .00001*denfxn(Inf,vv,a2,b2,vc2,mu2)
  total[i] <- log( answer$value + bottom + top ) }
den <- sum(total)
return(num-den) }

#Moderator mixed-effects model
denfxn <- function(yy,vv,XX,a2,b2,vc2,mu2,slope2) {
  pp <- 1-pnorm(yy/sqrt(vv))
  pp <- ifelse(pp < .00001,.00001,pp)
  pp <- ifelse(pp > .99999,.99999,pp)
  part1 <- pp^(a2-1)
  part2 <- (1-pp)^(b2-1)
  part3 <- dnorm(yy,XX*%c(mu2,slope2),sqrt(vv+vc2))
  return(part1*part2*part3) }

```

```

like <- function(pars) {
  a2 <- pars[1]
  b2 <- pars[2]
  vc2 <- pars[3]
  mu2 <- pars[4]
  slope2 <- pars[5]
  num <- sum((a2-1)*log(p) +
             (b2-1)*log(1-p) -
             1/2*log(v+vc2) - 1/2*((y-X%%c(mu2,slope2))^2/(v+vc2)))
  total <- rep(0,length(y))
  for(i in 1:length(y)) {
    vv = v[i]
    yy = y[i]
    XX = X[i,]
    answer <- integrate(denfxn,-Inf,Inf,vv=vv,XX,a2=a2,b2=b2,
                        vc2=vc2,mu2=mu2,slope2=slope2)
    bottom <- .00001*denfxn(-Inf,vv,XX,a2,b2,vc2,mu2,slope2)
    top <- .00001*denfxn(Inf,vv,XX,a2,b2,vc2,mu2,slope2)
    total[i] <- log( answer$value + bottom + top ) }
  den <- sum(total)
  return(num-den) }

```

In order to run the model using *R*'s *nlminb* built-in optimization algorithm, one must simply input the data (effects are represented by vector $\mathbf{y}$ and variances are denoted by vector $\mathbf{v}$), define a set of starting values for the parameters, and optimize the negative of the above log-likelihood function. Remember to place lower-bounds on a , b and τ^2 in order to keep the estimates from diverging to nonsensical (or non-estimable) values.

```
nlminb(pars,neglike,lower=c(0,0,0,-Inf))
```

Running the model using the MH algorithm is also not practically challenging. The following code implements the MH algorithm for a mean-only random-effects analysis. However, the code can easily be modified to include as many moderators as needed. It may be useful to look at the description of the algorithm (in Chapter 3's section *Bayesian Estimation*) while looking over the code in order to better understand the code.

```
#Define the acceptance rates.
acca <- 0
accb <- 0
accvc <- 0
accmu <- 0

#Define how many iterations of each chain will be run.
iterations <- 10000

#Set starting values for the parameters.
a <- rep(1,iterations)
b <- rep(1,iterations)
vc <- rep(.1,iterations)
mu <- rep(.5,iterations)

#Truncated normal distribution functions. These are included to
adjust for the asymmetry that is imposed on the parameters
when a lower- or upper-bound restriction is placed.
truncnum <- function(last,cand) {
  part1 <- (1/(sqrt(2*pi)*SD))*exp(-1/2*((last-cand)/SD)^2)
  part2 <- (pnorm(Inf,cand,SD) - pnorm(0,cand,SD))^-1
  return(part1*part2) }
```

```
truncden <- function(cand,last) {
  part1 <- (1/(sqrt(2*pi)*SD))*exp(-1/2*((cand-last)/SD)^2)
  part2 <- ( pnorm(Inf,last,SD) - pnorm(0,last,SD) )^-1
return(part1*part2) }
```

```
#Run the MH algorithm
```

```
for(i in 2:iterations) {
  #Update a.
  SD <- .1
  last = a[i-1]
  cand = rnorm(1,last,SD)
  a[i] = cand
  pars = c(cand,b[i-1],vc[i-1],mu[i-1])
  lastpars = c(last,b[i-1],vc[i-1],mu[i-1])
  if( like(pars) + log(dchisq(10-10*cand,2)) +
    log(truncnum(last,cand)) - like(lastpars) -
    log(dchisq(10-10*last,2)) - log(truncden(cand,last))
    < log(runif(1)) ) a[i] = a[i-1]
  else acca = acca + 1

  #Update b.
  last = b[i-1]
  cand = rnorm(1,last,.25)
  b[i] = cand
  apar <- a[i]
  pars = c(apar,cand,vc[i-1],mu[i-1])
  lastpars = c(apar,last,vc[i-1],mu[i-1])
  if( like(pars) + log(dnorm(cand,1,.4)) - like(lastpars) -
```

```

log(dnorm(last,1,.4)) < log(runif(1)) ) b[i] = b[i-1]
else accb = accb + 1

#Update vc.
SD <- .1
last = vc[i-1]
cand = rnorm(1,last,SD)
vc[i] = cand
apar = a[i]
bpar = b[i]
pars = c(apar,bpar,cand,mu[i-1])
lastpars = c(apar,bpar,last,mu[i-1])
if( cand < 0 || like(pars) + log(truncnum(last,cand)) -
like(lastpars) - log(truncden(cand,last))
< log(runif(1)) ) vc[i] = vc[i-1]
else accvc = accvc + 1

#Update mu.
last = mu[i-1]
cand = rnorm(1,last,.22)
mu[i] = cand
apar = a[i]
bpar = b[i]
vcpar = vc[i]
pars = c(apar,bpar,vcpar,cand)
lastpars = c(apar,bpar,vcpar,last)
if( like(pars) - like(lastpars)

```

```
< log(runif(1)) ) mu[i] = mu[i-1]  
else accmu = accmu + 1 }
```

Note that the positive of the log-likelihood function is run in the MH algorithm (not the negative log-likelihood as in *R*'s *nlminb* minimizer).