NORTHWESTERN UNIVERSITY

Topics in Meta-Analysis with Few Studies

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Statistics

By

Rrita Zejnullahi

EVANSTON, ILLINOIS

September 2021

ABSTRACT

Topics in Meta-Analysis with Few Studies

Rrita Zejnullahi

This dissertation consists of three papers on methods for meta-analysis with few studies. These papers are concerned with proper inference from meta-analysis models that combine data from a small number of studies using fixed and random-effects models. Chapter 1 provides an introduction to meta-analysis, the motivation for this work and an overview of each substantive paper. Chapters 2 to 4 contain the substantive papers. The title and a brief summary of each follow.

Chapter 2. Fixed and random-effects meta-analysis of randomized trials when the outcome is continuous and the number of studies is small.

In this chapter, I begin with a review of the technical and conceptual components of fixed-effect and random-effects models. Although a fixed-effect model is not directly the focus, it nevertheless plays an important role in motivating the random-effects model. I explore the feasibility of applying commonly-used random-effects methods to a small collection of studies (i.e. $k < 10$) under various configurations of unbalancedness that are of interest to empirical practice. The focus of this chapter is the Standardized Mean Difference effect size, which is the most frequently used effect size when the outcome in a study is continuous. Good statistical methods will produce unbiased and precise estimates, and the construction of confidence intervals will give proper coverage of the summary treatment effect. I demonstrate that commonly used random-effects procedures, however, can result in confidence intervals that may be too

narrow, depending on the configuration of the within-study sample sizes, amount of heterogeneity and number of studies. The chapter considers the application of robust standard errors to the meta-analytic context and shows that these serve as good alternatives to estimate the variance of the summary estimate. It further provides adjusted degrees of freedom for the t-test statistics that make these methods more suitable for small meta-analyses.

Chapter 3. Random-effects meta-analysis of randomized clinical trials when the outcome is binary and the number of studies is small.

In this chapter, I extend the results from Chapter 2 to effect sizes based on binary data, i.e. risk difference, log odds ratio and log risk ratio. The results from this chapter provide some empirical evidence about the accuracy of various random-effects methods - including ones that are common practice - and methods that utilize robust standard errors. The study reveals that methods that use robust standard errors are indeed promising under the scenarios I investigated. It is further demonstrated that one variant of robust standard errors leads to generally proper coverage across all scenarios considered.

Chapter 4. Considerations in heterogeneity variance estimation in small meta-analysis.

The challenges that arise from applying a random-effects model to a small number of studies is primarily due to the difficulty in obtaining an unbiased and efficient estimate of the heterogeneity variance parameter. In chapter 4, I investigate the properties of five heterogeneity variance estimators in terms of bias and efficiency. Bias and variance of the estimators are derived analytically, where possible to do so, and investigated empirically through Monte Carlo simulations. Recommendations for which estimator to use in practice depend on the goal of meta-analysis. If the goal is to quantify heterogeneity, then a point estimate should be provided along with its standard error. Guidance for choosing an estimator is primarily based on bias. It turns out that no one estimator is superior to the alternatives considered across all scenarios. Despite this fact, some generalizations are credible and presented in chapter 4.

Acknowledgements

I am indebted to many friends and colleagues for the guidance, support and encouragement throughout my time in graduate school.

First and foremost, a debt of gratitude to Larry Hedges, Elizabeth Tipton and Hongmei Jiang for their service on my dissertation committee. Thanks to my advisor, Larry Hedges, for guiding my growth as a young scholar and contributing generously to the ideas presented in this dissertation. My gratitude to Beth Tipton for some of the earlier directions of this work. Thanks to Larry and Beth for providing many opportunities to relate statistics to practice that have balanced my intellectual growth. And to Hongmei Jiang who has been supportive throughout, particularly in this past year.

I thank the following friends and colleagues for a positive impact on my experience in graduate school: Arend Kuyper, Katie Fitzgerald, Mena Whalen, Sarah Peko-Spicer, Abby Smith, Katie Coburn and Karina Diaz. Special thanks to Arend Kuyper for his invaluable advice on my research and unparalleled support throughout my time at Northwestern. To my amazing friends and cohort mates, Katie Fitzgerald and Mena Whalen, for their relentless encouragement and useful feedback. And to the administrative staff, Valerie Lyne and Kisa Kowal, for always being helpful with academic concerns. You have all enriched my graduate experience.

A heartfelt thank you to my family, Abdullah, Shkurte, Rreze and Guri Zejnullahi for their belief in my abilities and continuous encouragement in this pursuit. Finally, my love and appreciation to my incredible husband, Trim Gllogjani, for his kindness and patience that cannot be overstated throughout these past five years.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Preface

1.1. Introduction

Systematic reviews and meta-analyses are an essential element of the evidence-based movement in education, medicine and the social sciences. A systematic review is a review of the existing evidence on a key research question. It uses a systematic approach to identify, screen and evaluate studies, and report on the findings. A meta-analysis is defined as the quantitative synthesis of two or more studies to produce an estimate of treatment effectiveness with improved precision. A meta analysis is often but not always included as a component of a systematic review. The goal of meta-analysis in the context of systematic reviews is to summarize the existing evidence quantitatively for policy-makers and practitioners, with the overarching goal of promoting evidence-based decisions about education-relevant outcomes, human health and overall social good.

Methods for meta-analysis can be divided into procedures for combining statistical significance and procedures for combining treatment effects across studies (Hedges, 1985). The early foundations for the latter methods have been laid out by Hedges (1985) and DerSimonian and Laird (1986), building on the work of Cochran (1937), Yates and Cochran (1938), Cochran (1943), and GLASS (1976). Of the two types of statistical procedures, methods for combining treatment effects across studies have become the primary focus of research on meta-analysis because of their desirable statistical properties. Today, with the amount of evidence increasing at an accelerating rate, interest in meta-analysis and its applications is rapidly growing. A multitude of books have been written, with applications to the social sciences (Card, 2011; Cooper *et al.*, 2009) and medicine (Hartung, 2008; Schmidt and Hunter, 2014; Stangl *et al.*, 2000; Sutton, 2000). While advances to the methodology are ongoing, to date, meta-analyses with few studies have received little attention, which brings us to the motivation for this thesis.

1.2. Motivation

Meta-analyses of randomized control trials are considered to be the gold standard for establishing which treatments are effective, thus are a key component in informing policy and practice. In this mindset, the U.S. Department of Education's Institute of Education Sciences initiated the What Works Clearinghouse (WWC), a research clearinghouse whose role is to determine what works in education. To do so, they review policies, programs and practices to help policy-makers and practitioners make informed decisions based on the existing evidence. In medicine, a similar central source is the Cochrane Collaboration which maintains a database of over $7,500$ systematic reviews to inform decisions about human health. When more than one study investigated the same intervention, treatment effects from these individual studies can be combined using meta-analysis methods. In meta-analyses performed by research clearinghouses, however, the number of studies eligible for synthesis is quite small. For example, in meta-analyses performed by the WWC, the median number of studies is two. In the Cochrane Library, in a sample of $22,453$ meta-analyses, the median number of studies was found to be three, and in approximately 75% of the meta-analyses, the number of studies was five or less (Davey *et al.*, 2011). While traditional meta-analysis methods are considered to be the best approach to evidence synthesis, their strength is limited to large numbers of studies. With this in mind, this thesis considers a subset of the problems in small meta-analyses, which we describe in the next section.

1.3. Overview of thesis

A limitation to traditional meta-analysis methods is their basis on large sample theory. As indicated previously, however, empirical researchers typically work in areas that are constrained to a small sample. In such circumstances, they may have to settle for using methods that may not be intended to answer their key question but have better small sample properties. A preferable option would be to use methods that are flexible, simple and appropriate to answer their research question, yet suitable for small samples. In this body of work, I use a systematic approach to investigate several aspects of small meta-analyses.

By and large, experiments are designed to include an equal number of individuals per treatment group, however, it is highly unlikely that the number of individuals across studies will be equal. If a

meta-analysis contains studies with variable numbers of individuals per study, it then becomes important to investigate what effect do variable sample sizes have on the summary estimate and its precision, in particular when meta-analyses contain a small number of studies. The second chapter addresses these problems. Specifically, it considers the application of two routinely used models in meta-analysis, notably the fixed-effect and random-effects model. There is extensive research on fixed-effect and random-effects meta-analysis methods for combining a handful of studies. Their statistical properties are well documented in the literature for large numbers of studies, however, the application of these models to meta-analyses containing a small number of sources with variable sample sizes has received limited attention.

In Chapter 2, I first consider the role that fixed-effect and random-effects models play in drawing reliable conclusions from small meta-analyses. Additionally, a fixed-effect model is used to motivate a random-effects framework, which I argue is more appropriate in the meta-analytic setting. The focus of this chapter is the Standardized Mean Difference effect size, a commonly used effect measure for continuous data. The chapter compares the performance of commonly used random-effects methods using both theory and Monte Carlo simulations in a systematic way. While random-effects methods produce unbiased and efficient estimates, and confidence intervals for the summary treatment effect have proper coverage when the number of studies is sufficiently large ($k > 20$), I demonstrate that these methods result in confidence intervals that are not wide enough when k is small. This depends on the configuration of sample sizes across studies, degree of true heterogeneity and number of studies. I further discuss a method that applies robust variance estimators to the small sample meta-analytic context and demonstrate that these estimators are good alternatives to the conventional estimator for estimating the variance of the summary estimate. Additionally, the chapter provides an alternative variance estimator with less bias than the standard estimator, particularly in cases where one study is potentially more influential than the rest. Furthermore, I adjust the degrees of freedom for the t-test statistics that make these methods better suited for small meta-analysis.

In Chapter 3, I shift focus to effect sizes that are functions of proportions, i.e. risk difference, odds ratio and risk ratio. Effect sizes based on binary data typically arise in randomized clinical trials, where the results are reported as the number successes and failures. Because the odds ratio and risk ratio

sample statistics converge to the normal distribution faster on the log scale, I work with the log odds ratio and log risk ratio instead. The emphasis of this chapter is entirely on random-effects methods. While randomized clinical trials are the best method to determine whether a treatment is effective, these are not designed to answer questions that are of interest to meta-analysis. Hence, heterogeneity among findings can reasonably be expected, and it is imperative to quantify this heterogeneity and formally incorporate it into the analysis. Random-effects models are particularly attractive for this very reason and are therefore considered throughout Chapter 3. Beyond comparing common practice random-effects methods, I extend the results from Chapter 2 and provide empirical evidence about the feasibility of both common practice methods and ones that utilize robust variance estimators. It turns out that one alternative of the robust estimators is particularly well suited in this context too, leading to typically nominal coverage probabilities across all scenarios.

Chapters 2 and 3 are primarily focused on obtaining an unbiased and precise estimate of the summary treatment effect and constructing confidence intervals that give proper coverage. While arguably this is the main goal of meta-analysis, another issue remains. In random-effects models, heterogeneity plays a crucial role in the analysis and needs to be properly quantified. In Chapter 4, I investigate the properties of five heterogeneity variance estimators in terms of bias and efficiency. Bias and variance of the estimators are derived analytically, where possible to do so, and investigated empirically through Monte Carlo simulations. Recommendations for which estimator to use in practice depend on the goal of meta-analysis. If the goal is to quantify heterogeneity, then a point estimate should be provided along with its standard error. If, on the other hand, the goal is to obtain an unbiased and efficient estimate of the summary treatment effect, the choice of heterogeneity variance estimator has little impact on providing confidence intervals with proper coverage. This is true only when one uses robust estimators to estimate the variance of the summary estimate, where the total variance of an estimate is replaced by its squared sample residual.

Guidance for choosing an estimator for the first goal is primarily based on bias. It turns out that no one estimator is superior to the alternatives considered across all scenarios. Despite this fact, some generalizations are credible. These generalizations are useful when interest lies in obtaining a unbiased estimate of the heterogeneity variance itself. When studies contain an equal number of individuals per

study, it is shown that four of the five estimators considered are identical. All estimators exhibit better behavior when within-study sample sizes are sufficiently large. When meta-analyses contain studies that are unbalanced, a precision weighted method of moments estimator is superior to the alternatives in the presence of small to moderate degrees of heterogeneity, while a method of moments estimator that assigns equal weights is superior in the presence of a large degree of heterogeneity. Restricted maximum likelihood, obtained as a result of maximizing the log likelihood function which adjusts for the loss in degrees of freedom, performs similarly to the precision weighted method of moments estimator, and is a compromise between the equally weighted method of moments estimator and the precision weighted method of moments estimator. When there are only two studies in meta-analysis, it turns out that the two method of moments and the restricted and approximate restricted maximum likelihood estimators coincide, thus any of these four can be used in such a situation.

CHAPTER 2

Fixed and random-effects meta-analysis of randomized trials when the outcome is continuous and the number of studies is small

2.1. Introduction

Meta-analysis is considered to be the gold standard for evidence synthesis. It involves combining data from multiple independent sources to produce a summary estimate with improved precision. Traditionally, meta-analysis methods have been applied to a large collection of studies, and past research efforts have indicated its numerous strengths (Hedges, 1985; Hedges and Vevea, 1998). Working groups such as the What Works Clearinghouse in education and the Cochrane Collaboration in medicine, for instance, use meta-analysis methods to synthesize evidence from randomized control trials (RCTs) to produce the highest level of evidence for policymakers and practitioners. Increasingly, however, we observe that the number of studies eligible for synthesis is quite small (i.e. less than 10), and it is not uncommon to combine information from only two studies. In situations where the goal is to promote evidence-based decisions, formally addressing the problems we encounter in small meta-analyses is critical. These problems arise from multiple sources and no statistical procedure is entirely defensible.

Although the statistical properties of fixed-effect and random-effects procedures have been previously investigated by numerous authors (Brockwell and Gordon, 2001; DerSimonian and Laird, 2015; Hedges, 1985; Hedges and Vevea, 1998; Hunter and Schmidt, 2000), to date meta-analyses with few studies have received little attention. In the few instances some of the problems were addressed (Friede *et al.*, 2017; Röver *et al.*, 2015; Seide *et al.*, 2019), the focus was solely on random effects procedures. While the assumptions under a random effects model are more plausible in the meta-analytic context, the uncritical

application to a small collection of studies has implications that are both technical and conceptual in nature. Technical difficulties arise in estimating the heterogeneity variance, where estimates are both biased and highly unstable. This necessarily affects the precision of the summary estimate and the distribution of test statistics. To avoid this problem, empirical researchers might instead use fixed effects procedures. While this is a reasonable and practical approach, it changes the questions asked and alters the population of inference.

In this chapter, our aim is threefold. First, we begin by reviewing the conceptual framework of fixed and random-effects procedures and discuss the advantages and disadvantages of each in the context of meta-analyses with few studies. Second, we compare the performance of the most common random-effects methods used in practice, which have typically been discussed in isolation. We consider performance by looking at properties of the weighted mean, properties of various variance estimators and the coverage probability of $100(1-\alpha)\%$ confidence intervals. Third, because the performance of random-effects procedures depends largely on the ability to estimate the variance of the summary estimate well, we investigate the use of robust standard errors in this small sample meta-analytic context. There are no problems estimating the summary effect; the problem is with estimating its variance, which matters for inference about the summary effect.

The chapter is outlined as follows. Section 2.2 introduces the fixed and random-effects models, followed by a discussion of the conceptual framework of the two inference procedures in Section 2.3. Section 2.4 discusses the most commonly used methods for forming confidence intervals and performing hypothesis tests. Section 2.5 presents the robust standard errors and several modifications that make it more suitable to use in a small meta-analysis, particularly when the number of observations varies across trials. Section 2.6 provides evidence from a simulation study. Section 4.7 summarizes our findings and concludes with recommendations for empirical researchers.

2.2. Notation

In this section, we introduce the fixed-effect and random-effects models, assumptions and notation used throughout the chapter. We restrict our focus to the Standardized Mean Difference (SMD) effect size, which is the most commonly used measure of effect when the outcome of interest in a study is continuous.

2.2.1. Standardized mean difference

Suppose the data come from k independent studies that compare a treatment to a control. Let Y_{ij}^T and Y_{ij}^C denote the jth observation in the ith study in the treatment and control groups, respectively, where $j = 1, ..., n_i$ observations, $i = 1, ..., k$ studies, and suppose that the realizations in both groups are normally distributed with $\mu_{Y,i}^T$ and $\mu_{Y,i}^C$ means and common variances σ_i^2, $i = 1, ..., k$. The SMD effect size is defined by $\delta_i = (\mu_{Y,i}^T - \mu_{Y,i}^C)/\sigma_i$ and can be estimated by $d_i = (\bar{y}_i^T - \bar{y}_i^C)/s_i$, where $\bar{y}$ and s are the usual sample mean and sample standard deviation estimators. It is well known that $d_i \sim AN(\delta_i, \sigma_{d_i}^2)$, where the variance is defined by

$$\sigma_{d_i}^2 = \frac{n_i^T + n_i^C}{n_i^T n_i^C} + \frac{\delta_i^2}{2(n_i^T + n_i^C - 2)} \tag{2.2.1}$$

(Hedges, 1985).

How accurate is the asymptotic approximation to the distribution of the SMD estimator is a reasonable concern. Hedges (1985) has shown that sample sizes exceeding 10 observations per group are sufficient for the normal approximation to hold, when effect sizes are between 0.25 and 1.5. Note here that the variance estimator includes the true parameter, δ_i, which in practice is replaced by its estimate, d_i.

2.2.2. Fixed-effect model

A statistical model commonly used in meta-analysis is the fixed-effect model. Since we have assumed that the estimates to be combined follow the normal distribution, we have the following linear model

$$d_i = \delta + \epsilon_i, \tag{2.2.2}$$

with $\epsilon_i \sim N(0, \sigma^2_{d_i})$, $i = 1, ..., k$, where δ is the common mean - a fixed but unknown constant - and $\sigma^2_{d_i}$ is the within-study variance.

The goal in meta-analysis is to obtain an estimate of this true mean, δ, and find its standard error. The estimator widely used in practice is the precision weighted mean, i.e.

$$\bar{d}_+ = \sum_{i=1}^{k} \beta_i d_i \tag{2.2.3}$$

which has variance $var(\bar{d}_+) = 1/W$, where $\beta_i = w_i/W$, $w_i = 1/\hat{\sigma}^2_{d_i}$ and $W = \sum_{i=1}^{k} w_i$. Because the usual assumption in meta-analysis is to treat the within-study variances, $\hat{\sigma}^2_{d_i}$, as known parameters, $\sigma^2_{d_i}$, we drop the hat notation from now on with the understanding that there is a source of uncertainty that we have not accounted for in our analysis method. Inference in fixed-effect analysis is then based on the normal distribution, see Section 2.4.

2.2.3. Random-effects model

If the data is obtained as a random sample from a population of $K > k$ studies, one can use the random effects approach to combine information from independent studies. The random-effects linear model is

$$\begin{aligned} d_i &= \delta_i + \epsilon_i \\ &= \delta + (\delta_i - \delta) + \epsilon_i \\ &= \delta + \xi_i + \epsilon_i, \end{aligned} \tag{2.2.4}$$

where δ is the mean of the distribution of effects, $\xi_i = (\delta_i - \delta)$ is the bias from study i, and ϵ_i is the sampling error associated with study i (DerSimonian and Laird, 1986; Hedges, 1983). Study biases and sampling errors are assumed to be independent normal with 0 means and variances τ^2 and $\sigma^2_{d_i}$, respectively, and independent of each other; that is, $\epsilon_i \sim N(0, \sigma^2_{d_i})$, $\xi_i \sim N(0, \tau^2)$ and $Cov(e_i, \xi_i) = 0$.

The goal under a random-effects approach remains the same; that is, we want to combine information from multiple independent sources to provide a summary estimate of δ and estimate its variance. Note, however, that in this situation δ takes a different meaning: *It is the mean of a distribution of effects.* To

estimate δ, a common estimator to use is

$$\bar{d}_+ = \sum_{i=1}^{k} \hat{\beta}_i d_i, \tag{2.2.5}$$

with variance $var(\bar{d}_+) = 1/\widehat{W}$, where $\hat{w}_i = 1/(\sigma^2_{d_i} + \hat{\tau}^2)$. The convention is to treat within-study variances as known, while the heterogeneity variance, τ^2, is estimated from the available data using a wide range of techniques. In this article, we use the DerSimonian and Laird method of moments estimator, $\hat{\tau}^2_{DL}$, defined as

$$\hat{\tau}^2_{DL} = \begin{cases} \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}} & Q > k-1 \\ 0 & Q \leq k-1 \end{cases} \tag{2.2.6}$$

(DerSimonian and Laird, 1986). For a more extensive discussion on alternative heterogeneity variance estimators and their merits in small meta-analyses, see Chapter 4. Inference in random effects meta-analysis can then be based on either the normal distribution or the t-distribution. Section 2.4 provides further details.

2.3. Fixed versus random-effects conceptual framework

In fixed-effect models, we assume the summary effect is an unknown but *fixed* constant to be estimated from the available data, where the only source of variation arises from sampling observations within studies. The goal here is to infer to the set of studies included in the meta-analysis only. In contrast, random-effects procedures assume there is a distribution of true effects; here we estimated the mean of this distribution. Random-effects models incorporate two sources of random variation: sampling error and true heterogeneity between findings. Additionally, the results obtained from a random-effects model are presumed to be generalisable to other similar studies.

Whether to consider the effects of studies as fixed or random is not immediately obvious, particularly when constrained to only a handful of studies. When results from different studies are not heterogeneous beyond what we expect from sampling error, then the fixed-effect model is adequate. However, this can

be a particularly difficult assumption to justify, because studies synthesized in meta-analysis are often not functionally identical. If we assume no heterogeneity, then $\tau^2 = 0$, which means that the only variance incorporated into the precision weights are the within-study variances. This allows the studies with the smallest standard errors to dominate, a feature that could be undesirable. Furthermore, if heterogeneity is expected, there does not exist a simple way of incorporating it. For this reason, it is preferred to use a more flexible procedure that formally considers between-study heterogeneity.

The random-effects model has been the method of choice for empirical practice, and although it is more flexible compared to the fixed-effect approach specifically because one formally incorporates between-study heterogeneity, it by no means can be applied to all situations. First, estimates of the between-study variance parameter are biased and variable in small samples which create challenges in obtaining nominal coverage probabilities of confidence intervals for the summary effect and controlling Type I error rate for hypothesis testing. In addition, confidence intervals tend to be wider compared to fixed-effect, and in the case of a small number of studies, this could mean confidence intervals that are practically uninformative. Finally, a random-effects model does not help us understand possible systematic errors.

2.4. Inference about δ: Confidence intervals and hypothesis tests

2.4.1. Fixed-effect inference

We can conduct hypothesis tests and obtain $100(1-\alpha)\%$ confidence intervals for δ using the normal distribution as a reference distribution. Remember that $d_i \sim AN(\delta, \sigma^2_{d_i})$. It then follows that

$$\bar{d}_+ \sim N(\delta, 1/W), \tag{2.4.1}$$

and one can construct a $100(1-\alpha)\%$ confidence interval (δ_L, δ_U) for δ by

$$\delta_L = \bar{d}_+ - z_{\alpha/2}\sigma_{\bar{d}_+}, \quad \delta_U = \bar{d}_+ + z_{\alpha/2}\sigma_{\bar{d}_+}, \tag{2.4.2}$$

where $z_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution. The hypothesis that the mean is different from a pre-specified value, i.e. $H_0 : \delta = \delta_0$ vs $H_a : \delta \neq \delta_0$, is tested using the test statistic

$$Z = \frac{\bar{d}_+ - \delta_0}{\sigma_{\bar{d}_W}}. \tag{2.4.3}$$

Under the null hypothesis, Z is a standard normal random variable, hence we reject the null hypothesis at level α if $|Z| > z_{\alpha/2}$.

2.4.2. Random-effects inference using the normal distribution as a reference distribution

One standard method to conduct hypothesis tests and obtain $100(1-\alpha)\%$ confidence intervals for δ is to use the normal distribution as a reference distribution. Recall that $d_i \sim AN(\delta, \sigma_{d_i}^2 + \tau^2)$. It immediately follows that

$$\bar{d}_+ | \tau^2 \sim N(\delta, 1/W), \tag{2.4.4}$$

where the weights are inversely proportional to the variances of the effect estimates. Note here that the distributional assumption is based on an asymptotic argument as k approaches infinity.

A $100(1-\alpha)\%$ confidence interval (δ_L, δ_U) for δ is given by

$$\delta_L = \bar{d}_+ - z_{\alpha/2}\sigma_{\bar{d}_+},\ \delta_U = \bar{d}_+ + z_{\alpha/2}\sigma_{\bar{d}_+}, \tag{2.4.5}$$

where $z_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution. To test the hypothesis that the mean is different from a pre-specified value δ_0, i.e. $H_0 : \delta = \delta_0$ vs $H_a : \delta \neq \delta_0$, compute the test statistic

$$Z = \frac{\bar{d}_+ - \delta_0}{\sigma_{\bar{d}_+}}. \tag{2.4.6}$$

Under the null hypothesis, Z is a standard normal random variable, hence we reject the null hypothesis of no effect at level α if $|Z| > z_{\alpha/2}$.

A vast amount of literature has documented the inadequacy of the normal distribution as a reference distribution in the random effects framework when the number of studies is small (i.e. less than 20) (Hartung, 1999a; Hartung and Knapp, 2001a; IntHout *et al.*, 2014). To deal with these shortcomings, alternative methods have been proposed. The most widely used in empirical practice is the Hartung-Knapp adjustment (Hartung and Knapp, 2001a), also proposed by Sidik and Jonkman (2002a). Next, we turn to these methods.

2.4.3. Random effects inferences using the t-distribution as a reference distribution: Knapp-Hartung method

The use of the standard normal distribution as a reference distribution for hypothesis tests and confidence intervals has been previously shown to perform quite poorly when k is small (say, less than 20), leading to a high rate of liberal decisions. Hartung (1999b) and Sidik and Jonkman (2002b) proposed an alternative procedure that utilized the t-distribution which performs better compared to the standard method when $k < 20$. The procedure is as follows.

Let $\beta_i = w_i/W$, $i = 1, ..., k$. Define

$$\lambda(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{k} \beta_i^2}{1 - \sum_{i=1}^{k} \beta_i^2}, \text{ and } \psi_i(\boldsymbol{\beta}) = \beta_i + \frac{\beta_i^2 - \beta_i}{1 - \sum_{i=1}^{k} \beta_i^2}, \tag{2.4.7}$$

$i = 1, ..., k$, and consider the following sum of squares function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{k} \beta_i (d_i - \bar{d}_+)^2. \tag{2.4.8}$$

In the original work, it is shown that

(1) $WS(\boldsymbol{\beta}) \sim \chi^2_{k-1}$,

(2) $\bar{d}_+$ and $S(\boldsymbol{\beta})$ are stochastically independent, and

(3) $S(\boldsymbol{\beta})/(k-1)$ is an unbiased estimator of $var(\bar{d}_+) = 1/W$,

hence it follows that the test statistic

$$T_1 = \frac{\bar{d}_+ - \delta_0}{\hat{\sigma}_{\bar{d}_+}} \dot{\sim} t_{k-1}, \tag{2.4.9}$$

where $var(\bar{d}_+) = S(\boldsymbol{\beta})/(k-1)$. The test then rejects the null hypothesis ($H_0 : \delta = \delta_0$) at level α if $|T| > t_{k-1,\alpha/2}$. A $100(1-\alpha)\%$ confidence interval (δ_L, δ_U) for δ is computed by

$$\delta_L = \bar{d}_+ - t_{k-1,\alpha/2}\hat{\sigma}_{\bar{d}_+}\,, \delta_U = \bar{d}_+ + t_{k-1,\alpha/2}\hat{\sigma}_{\bar{d}_+}, \tag{2.4.10}$$

where $t_{k-1,\alpha/2}$ is the two-tailed critical value of the t-distribution with $(k-1)$ degrees of freedom.

The theory is based on the justification that the weights are known. In practice, however, the weights are estimated from the data, and when the data consist of a small number of sources, there is substantial undesired variability in the weights. This in turn leads to methods that are typically more liberal than we would like. Hartung (1999b) has already pointed out that the preceding procedure is, in fact, sensitive to variable weights and provides a more general approach that is less sensitive to this variability in the weights. Next, we turn to this more general approach.

2.4.4. Random-effects inference using the t-distribution as a reference distribution: Hartung method

First, consider the following sum of squares function

$$Q(\boldsymbol{\beta}) = \lambda(\boldsymbol{\beta})S(\boldsymbol{\beta}) + \sum_{i=1}^{k} \psi_i(\boldsymbol{\beta})\sigma^2_{d_i}, \tag{2.4.11}$$

where $\lambda(\boldsymbol{\beta})$ and $\psi_i(\boldsymbol{\beta})$ are defined by (2.4.7) and $S(\boldsymbol{\beta})$ is given by (2.4.8). When the weights are equal, i.e. $\beta_i = \beta$ for all $i = 1, ..., k$, $Q(\boldsymbol{\beta})$ reduces to $S(\boldsymbol{\beta})/(k-1)$, the variance estimator proposed by both Hartung & Knapp and Sidik & Jonkman.

On the other hand, when the weights, $\beta_1, ..., \beta_k$, take different values - which is typically the case in empirical practice - then $Q(\boldsymbol{\beta})$ can become negative with positive probability and therefore a new

truncated estimator is defined by

$$q_{L(\boldsymbol{\beta})}(\boldsymbol{\beta}) = L(\boldsymbol{\beta})Q(\boldsymbol{\beta}) + \big\{1 - L(\boldsymbol{\beta})\big\}R(\boldsymbol{\beta}), \tag{2.4.12}$$

where

$$L(\boldsymbol{\beta}) = min\Bigg\{1, max\Bigg\{0, \frac{(Q(\boldsymbol{\beta})/R(\boldsymbol{\beta})) - A}{B - A}\Bigg\}\Bigg\} \tag{2.4.13}$$

and

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{k} \beta_i^2 \sigma_{d_i}^2. \tag{2.4.14}$$

Under the null hypothesis, the test statistic

$$T_2 = \frac{\bar{d}_+ - \delta_0}{\sqrt{\hat{q}}} \dot{\sim} t_{\hat{f}}, \tag{2.4.15}$$

with $\hat{f} = 2\hat{q}^2/\widehat{var}(\hat{q})$ where $\widehat{var}(\hat{q}) = L(\beta)^2(1/\tau)^2 2(k-1)\big\{\lambda(\beta)^2\big\}$. It is worthwhile to note here that we have omitted knowledge about the uncertainty of $Var(\hat{\sigma}_{d_i}^2)$ because we began with the assumption that the within-study variances are true parameters and not estimates, therefore the values to use for A and B are fixed values. The choice of values for A and B in the neighborhood of 1 has little impact on the procedure; we therefore use values originally chosen by Hartung, $A = 0.95$ and $B = 1.05$, in our forthcoming simulation study.

The preceding two methods performed similarly in simulations as done by Hartung, therefore, it was recommended that the simpler approach be adopted in practice. Friede *et al.* (2017), however, revisited some of these issues for the KnHa method; they show that when sample sizes vary considerably across studies, the coverage probability of confidence intervals for δ is below nominal, leading to a higher rate of liberal decision. For this reason, we revisit the more general method to delineate the conditions under which it is worthwhile to use in practice.

2.5. Robust variance estimation

In this section, we present several variants of robust variance estimators and approximations for the distributions of t-test statistics under the null hypothesis that can be used to conduct hypothesis tests and compute confidence intervals in random-effects meta-analysis. Ideally, a variance estimator is consistent for the true variance and able to handle heteroscedasticity. Building on the work of Eicker (1967), Eicker (1963) and Huber (1967), White (1980) introduced robust variance estimators. The theoretical justification for these estimators and associated intervals relies on large samples. In small samples, these methods do not perform as satisfactorily as we would like. For this reason, other alternative estimators have been proposed - most of which make an adjustment to the squared sample residuals. Next, we present three of these alternatives.

Given constant weights, the variance of the summary estimate is given by

$$\widehat{var}(\bar{d}_+) = \sum_{i=1}^{k} \beta_i^2(\sigma_{d_i}^2 + \tau^2). \tag{2.5.1}$$

If the weights are know, this reduces to $1/W$, the typical variance used in performing a standard random-effects analysis. In practice, the weights are estimated instead, hence, one can define a new variance for the summary estimate in terms of the squared sample residuals, i.e.

$$\widehat{var}(\bar{d}_+)_{HC0} = \sum_{i=1}^{k} \beta_i^2(d_i - \bar{d}_+)^2. \tag{2.5.2}$$

This robust variance estimator has been used in a variety of applications and is widely known in the literature as HC0 (White, 1982). The HC0 estimator is biased downward in small samples, resulting in too liberal decisions from hypothesis testing and confidence intervals. Though its properties have been evaluated in other contexts and the general consensus is to not use it in a small sample context, it is nevertheless useful to discuss it for completeness, considering most other alternatives build on this estimator.

It is particularly instructive to consider the case of equal sized studies, i.e. $n_i = n$ for $i = 1, ..., k$. In this scenario, HC0 reduces to $HC0 \approx \sum_{i=1}^{k}(d_i - \bar{d})^2/k^2$. In the extreme case of $k = 2$ studies, this leads

to severe underestimation (by approximately 50%). The bias decreases as the number of studies increases, however, even when $k = 10$, HC0 underestimates the true variance by about 10%. It is therefore not recommended for practice; instead one adjusts the squared sample residuals by some factor to decrease its bias. The next three estimators presented are all alternatives to HC0 and all adjust this estimator by some constant factor.

One way of reducing the bias of HC0 is to multiply the squared sample residuals by a factor of $k/(k-1)$; this new estimator is referred to as HC1 in the literature (MacKinnon and White, 1985) and is

$$\widehat{var}(\bar{d}_+)_{HC1} = \frac{k}{k-1}\sum_{i=1}^{k}\beta_i^2(d_i - \bar{d}_+)^2. \tag{2.5.3}$$

This adjustment is obtained by considering equal weights, see A for more details. When $n_i = n$, $i = 1, ..., k$, $HC1 = \sum_{i=1}^{k}(d_i - \bar{d})^2/[k(k-1)]$, which is an unbiased estimator of $1/W$. This is also equivalent to the variance estimator proposed by Hartung and Knapp (2001b) and Sidik and Jonkman (2002a). In a retrospective meta-analysis, however, it is rarely the case that studies have equal sample sizes, and consequently, equal weights, hence the next estimator, HC2, is investigated.

HC2 builds on HC0 by adjusting the squared sample residuals by $(1-\beta_i)^{-1}$ (Horn *et al.*, 1975; MacKinnon and White, 1985) and is given by

$$\widehat{var}(\bar{d}_+)_{HC2} = \sum_{i=1}^{k}\beta_i^2(1-\beta_i)^{-1}(d_i - \bar{d}_+)^2. \tag{2.5.4}$$

Appendix A provides a proof of the bias of HC0 when the weights vary. In the case of equal sample sizes across studies, this estimator reduces to HC1 and is exactly unbiased for $1/W$. When there is unbalancedness, this estimator is slightly biased, particularly when synthesizing a small number of studies, however, the bias decreases as the number of studies, k, increases.

This estimator has also been previously suggested by Sidik and Jonkman (2006). Their investigation consisted of 10 or more studies and their recommendation was to use HC2 along with a t-distribution with $(k-1)$ degrees of freedom to perform hypothesis tests and compute confidence intervals since it provides protection against the *working weights*. However, they also show that the coverage probability

of confidence intervals using HC2 together with a t-distribution with $(k-1)$ degrees of freedom was practically the same as the coverage resulting from the Knapp-Hartung method. When $k < 10$, however, we show in Section 2.6 that using HC2 leads to undercoverage if we do not adjust the degrees of freedom for the t reference distribution. In the next section, we use a *Box* transformation to find the degrees of freedom adjustment for HC2.

Another estimator that has shown promise in other applied areas is HC3, which further penalizes the squared sample residuals (Davidson and MacKinnon, 1993). HC3 gives an extra penalty to values that could potentially be more influential by weighting the squared sample residuals by $1/(1-\beta_i)^2$, and is defined as

$$\widehat{var}(\bar{d}_+)_{HC3} = \sum_{i=1}^{k} \beta_i^2 (1-h_i)^{-2} (d_i - \bar{d}_+)^2. \tag{2.5.5}$$

When $n_i = n$, $i = 1, ..., k$, HC3 reduces to $\sum_{i=1}^{k} \frac{1}{(k-1)^2}(d_i - \bar{d}_+)^2$, and is positively biased for $1/W$ when k is small. This is supported by results in Section 2.6. The bias is more severe when $k = 2$, but decreases rapidly as the number of studies increases.

2.5.1. Inference using these estimators

Any of these variance estimators can be used to compute confidence intervals and perform hypothesis testing. One can use the t-distribution as a reference distribution with $(k-1)$ degrees of freedom along with one of: HC1, HC2 or HC3. Using the t-distribution with $(k-1)$ degrees of freedom and HC2 does not keep the coverage probability of confidence intervals for the summary effect, δ, at the nominal level, therefore, an adjustment for the degrees of freedom is suggested next.

We use a transformation by Box (1954) to derive the new degrees of freedom. That is, to find $g = 2[E(\widehat{Var}(\bar{d}_+))]^2 / Var(\widehat{Var}(\bar{d}_+))$, match the first two moments of the random variable Q to that of a χ^2 random variable. To find the expected value and variance of HC2, we use a general result by Rao *et al.*

(1981), which is restated in Appendix C. The expectation and variance are then found to be

$$E(\widehat{Var}(\bar{d}_+)_{HC2}) = \sum_{i=1}^{k} \frac{w_i}{1-\beta_i} - \frac{1}{\sum_{i=1}^{k} w_i} \sum_{i=1}^{k} \frac{w_i^2}{(1-\beta_i)}, \tag{2.5.6}$$

and

$$\begin{aligned} Var(\widehat{Var}(\bar{d}_+)_{HC2}) = & \frac{1}{\left(\sum_{i=1}^{k} w_i\right)^2} \sum_{i=1}^{k} \frac{w_i^2}{(1-\beta_i)^2} \\ & - \frac{2}{\left(\sum_{i=1}^{k} w_i\right)^3} \sum_{i=1}^{k} \frac{w_i^3}{(1-\beta_i)^2} \\ & + \frac{1}{\left(\sum_{i=1}^{k} w_i\right)^4} \left(\sum_{i=1}^{k} \frac{w_i^2}{(1-\beta_i)^2} \right)^2, \end{aligned} \tag{2.5.7}$$

respectively, where $w_i = 1/(\sigma_{d_i}^2 + \tau^2)$ and $\beta_i = w_i / \sum_{i=1}^{k} w_i$.

Hypothesis testing and confidence intervals can then be performed using any of these variance estimators along with a t-distribution as a reference distribution and associated degrees of freedom. Because we are working with quite small samples, it turns out that using HC3 with a t-reference distribution with $(k-1)$ degrees of freedom keeps the coverage probability of confidence intervals approximately at the nominal value across all scenarios covered in this article. HC2 with the adjusted degrees of freedom also performs well, but only for meta-analyses of at least four studies.

For completeness, a $100(1-\alpha)\%$ confidence interval (δ_L, δ_U) for δ is computed as

$$\delta_L = \bar{d}_+ - t_{k-1,\alpha/2} \sqrt{\widehat{var}(\bar{d}_+)_{HC3}}\,, \delta_U = \bar{d}_+ + t_{k-1,\alpha/2} \sqrt{\widehat{var}(\bar{d}_+)_{HC3}}. \tag{2.5.8}$$

The test statistic is computed as

$$T_{HC3} = \frac{\bar{d}_+ - \delta_0}{\sqrt{\widehat{Var}(\bar{d}_+)_{HC3}}}, \tag{2.5.9}$$

and rejects the null hypothesis at level α if $|T| > t_{k-1,\alpha/2}$.

2.6. Simulation evidence

The results presented in this article are derived from large sample approximations instead of exact small sample results. We therefore perform a simulation study to assess the properties of the summary estimate and its variance alternatives. We evaluate the coverage probability of 95% confidence intervals when the number of studies is small under several configurations of unbalancedness that are of interest to empirical researchers. Note here that simulations are based on random-effects procedures only. Fixed-effect procedures keep coverage rates at the nominal level because we assume that the within-study variances are known parameters.

2.6.1. Simulation design

The simulation study was based on two arm experiments with an equal number of treatment and control group observations. The meta-analyses consisted of both balanced and unbalanced configurations. We investigated analyses of all equal sized studies, one large study, one small study, and half-half large and small studies. The ratio of the study sample sizes ranged from 1 - which means that all studies are of equal size, to 10. We limited our simulations to small k ranging from 2 to 10. These values are typically found in practice, specifically in meta-analyses performed by research clearinghouses.

The data generating process for the random-effects procedures includes the following steps:

(1) First, SMD effect sizes, δ_i, are simulated from the normal distribution with variance τ^2. Without loss of generality, we set the average of the study true means equal to 0.5, corresponding to a medium effect (Cohen, 1988). The heterogeneity parameter, τ^2, can be defined in several ways. One approach is to set it equal to the simple arithmetic average of the within-study variances, $\tau^2 = r\bar{\sigma}^2$. We let r range from 0 to 2, representing several levels of true heterogeneity that are typical in practice.

(2) The second step is to simulate SMD effect estimates, d_i, from the normal distribution with mean δ_i and variance $\sigma^2_{d_i}$ defined by

eqrefeq:cohdVar.

For each combination of the parameters, we simulated $10,000$ meta-analyses and performed random-effects analyses as described in Sections 2.4 and 2.5.1. Although the combinations of parameters do not exhaust all the possibilities, they do cover scenarios that closely resemble practice so that the reader can get an adequate understanding of the issues present.

2.6.2. Simulation results

The first set of results report on the performance of the robust variance estimators in terms of bias. When synthesizing equal sized studies, all variance estimators are accurate across different degrees of heterogeneity, with the exception of $k = 2$. Here, the standard and Hartung variance estimators overestimate the empirical variance by 47% when $\tau^2 = 0$, but are quite accurate when heterogeneity increases. HC3, on the other hand, is severely positively biased when $k = 2$; this bias decreases as heterogeneity increases, however, even when $\tau^2 = 2$, HC3 overestimates the empirical variance by 91%. KnHa, HC1 and HC2 are all accurate. When there is unbalancedness across studies, most estimators underestimate the empirical variance, except for HC3. Underestimation is more severe in meta-analyses of half-half large and small studies or one large study, but even in the worst case scenario, the ratio of the model variance to the empirical variance is approximately 88%. HC3 exhibits positive bias across all scenarios when $k = 2$ and $\tau^2 = 0$, but shows better performance overall in cases of medium to large heterogeneity and $k > 2$, outperforming all other alternatives.

Next, we investigated the performance of confidence intervals using the standard Z method, Knapp-Hartung, Hartung, and the robust variance variants. In Figure 2.1, columns correspond to different levels of heterogeneity, while the rows show various scenarios of unbalancedeness. Each panels shows coverage rates for various numbers of studies ($k = 2$ to $k = 10$). When the studies in meta-analysis are balanced, the KnHa method yields superior coverage rates - corresponding to exactly the nominal value of 0.95. Using the Hartung method works well in cases of small to moderate heterogeneity ($\tau^2 = 0, 0.5$), but coverage decreases as heterogeneity increases when $k \leq 4$. This is because the estimate of the between-study variance component becomes more unstable as τ^2 increases, leading to slight to moderate undercoverage.

We see similar patterns in different scenarios of unbalancedeness - the KnHa method yields coverage rates that are closer to the nominal value for $k \leq 4$ in all cases of heterogeneity, however, when $k > 4$, the Hartung method keeps coverage approximately at the nominal value and is superior to the KnHa method. Similar results for the KnHa method have previously been reported in the literature (see Röver *et al.* (2015)). Because in a retrospective meta-analysis one would not expect balance, the Hartung method is recommended when $k \geq 5$.

The second set of simulations reports coverage rates resulting from using the robust estimators along with a t-distribution with $(k-1)$ degrees of freedom. Figure 2.2 presents these results. Similar to Figure 2.1, results are shown for different levels of true heterogeneity and different scenarios of unbalancedeness. When the studies are balanced, using any of the robust variance estimators to estimate the variance of the summary estimate along with a t-distribution with $(k-1)$ degrees of freedom yields nominal coverage rates. Confidence intervals and tests that use HC3 produce coverage rates that are slightly above the nominal value of 0.95 in the case of equal sized studies, however, in all other scenarios, HC3 is superior to all other methods and keeps coverage rates at approximately the nominal value.

Figure 2.3 presents a plot of the best performing methods in terms of coverage for easier comparison: Hartung, KnHa and HC3. We note here that KnHa performs extremely well in balanced settings. In unbalanced settings, its performance varies depending on the configuration of sample sizes across studies. The Hartung method performs quite well when $k > 5$ across all scenarios. HC3 is conservative when $k = 2$, but performs extremely well otherwise.

Finally, we include results from using HC2 with a corrected degrees of freedom, $\hat{g}$, in the following tables. We include coverage rates for HC2 without the adjustment and HC3 for comparison. Using an adjustment for HC2 leads to coverage rates that are closer to nominal across all scenarios considered, except for $k = 2$. Overall, HC2a and HC3 either perform similarly to or outperform most other methods.

2.7. Comment on the choice of analysis method

All the methods discussed in this article are appropriate for use in meta-analyses of a large number of studies. In small meta-analysis, however, investigators are faced with multiple challenges, including how to estimate the variance of the summary estimate well and what distribution to use for forming confidence intervals and performing hypothesis testing.

The standard method that uses the normal distribution as a reference distribution is not recommended when $k \leq 10$. In simulations not presented here, we find that coverage rates are far below the nominal value, particularly when there is unbalancedness, leading to a high rate of liberal decisions. This has been well documented in the literature. The Knapp-Hartung method has typically been recommended for use with small samples. We show, however, that this method leads to decreased coverage rates in cases of unbalancedness; these results agree with Röver *et al.* (2015). As a solution to this problem, Hartung initially proposed an alternative procedure that works better when weights are more variable, however, in simulations we find that the method is superior to Knapp-Hartung only when $k > 4$. A potential advantage of the Hartung method is that it allows for the incorporation of the uncertainty of the within-study variances. When data consists of SMD estimates, however, the gain is not substantial if within-study sample sizes are substantially large (i.e. $n > 30$) and hence not necessary to incorporate.

The previous two methods have shortcomings when $k < 10$. We therefore investigated the utility of applying robust standard errors to small meta-analyses. We found that both HC2 with an adjusted degrees of freedom and HC3 kept coverage rates approximately at the nominal level across most scenarios, with the exception of $k = 2$. In this case, HC3 resulted in overcoverage, while HC2a resulted in undercoverage.

2.8. Summary and discussion

Meta-analyses of a small number of studies are quite common in practice, yet little evidence is available about best practices for use of meta-analysis methods in this context. One of the goals of this chapter was to contribute to this gap in the literature by exploring the feasibility of applying random-effects methods to a handful of studies. In Section 2.2, we described the fixed and random-effects models. Section 2.3 discussed the conceptual differences between the two inference procedures. Section 2.4 considered common practice methods for statistical inference, while Section 2.5 presented the robust variance estimator and various modifications that make this estimator more appropriate to use in small meta-analyses. Section 2.6 presented a simulation study examining common practice methods and methods that use robust variance estimators. Our emphasis throughout was on the Standardized Mean Difference effect size which is frequently used when data in primary studies is continuous.

We find that common random-effects methods result in confidence intervals that are not wide enough, depending on the number of studies, degree of heterogeneity and whether meta-analyses contain equal or unequal numbers of observations per study. In cases of unbalancedness, coverage rates decrease, with the decrease depending on the configuration of the sample sizes of primary studies. Techniques that use robust variance estimators instead, show substantial promise in these scenarios. Specifically, it is shown that one of these estimators is least biased and leads to coverage rates near 95% in most scenarios investigated, except for meta-analyses containing two primary studies. When meta-analyses consist of only two studies, none of the methods discussed are adequate. Put another way, meta-analyses of two studies do not lend well to drawing conclusive findings in a random-effects framework, irrespective of the statistical approach used.

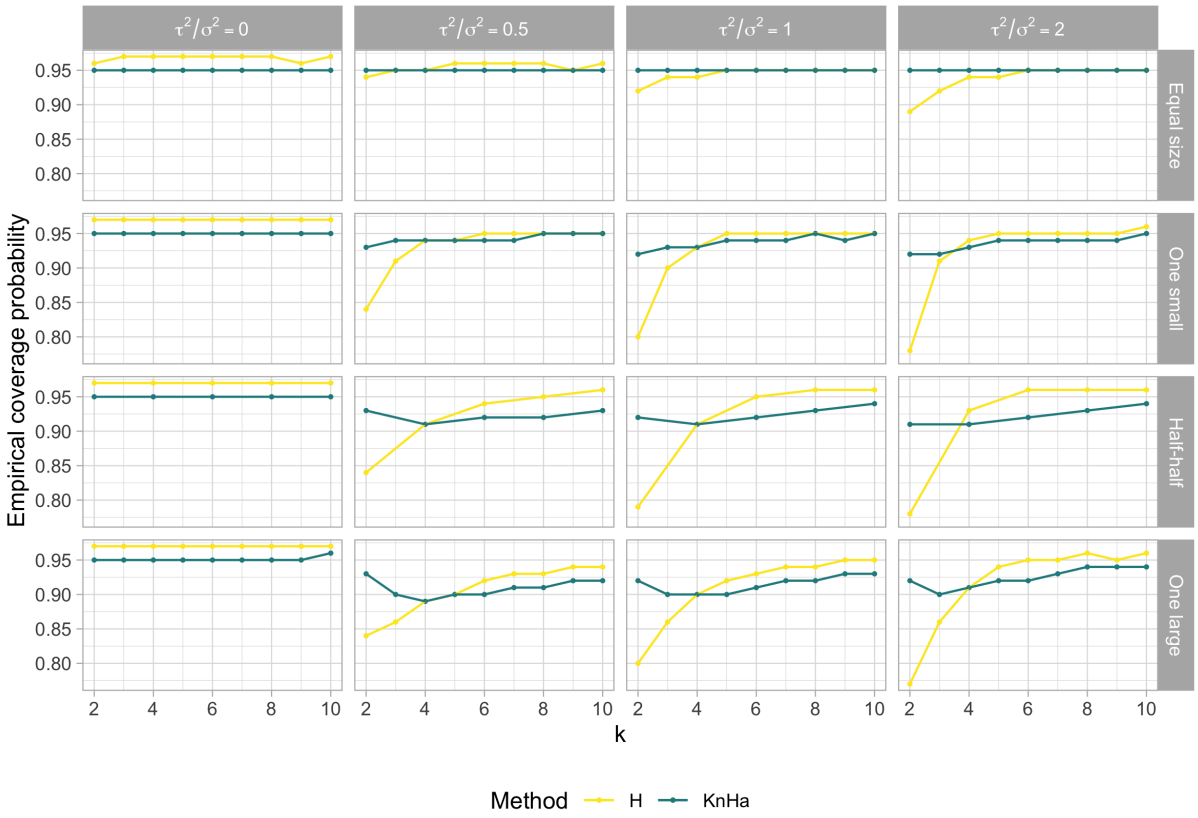

Figure 2.1. Empirical Coverage Probability of 95% Confidence Intervals for δ. This plot displays the coverage probability of the Hartung and KnHa 95% confidence intervals for the true effect ($\delta = 0.5$) using the t-distribution as a reference distribution. The columns represent different levels of heterogeneity. The rows represent balanced and unbalanced configurations. The number of studies ranges from $k = 2$ to $k = 10$.

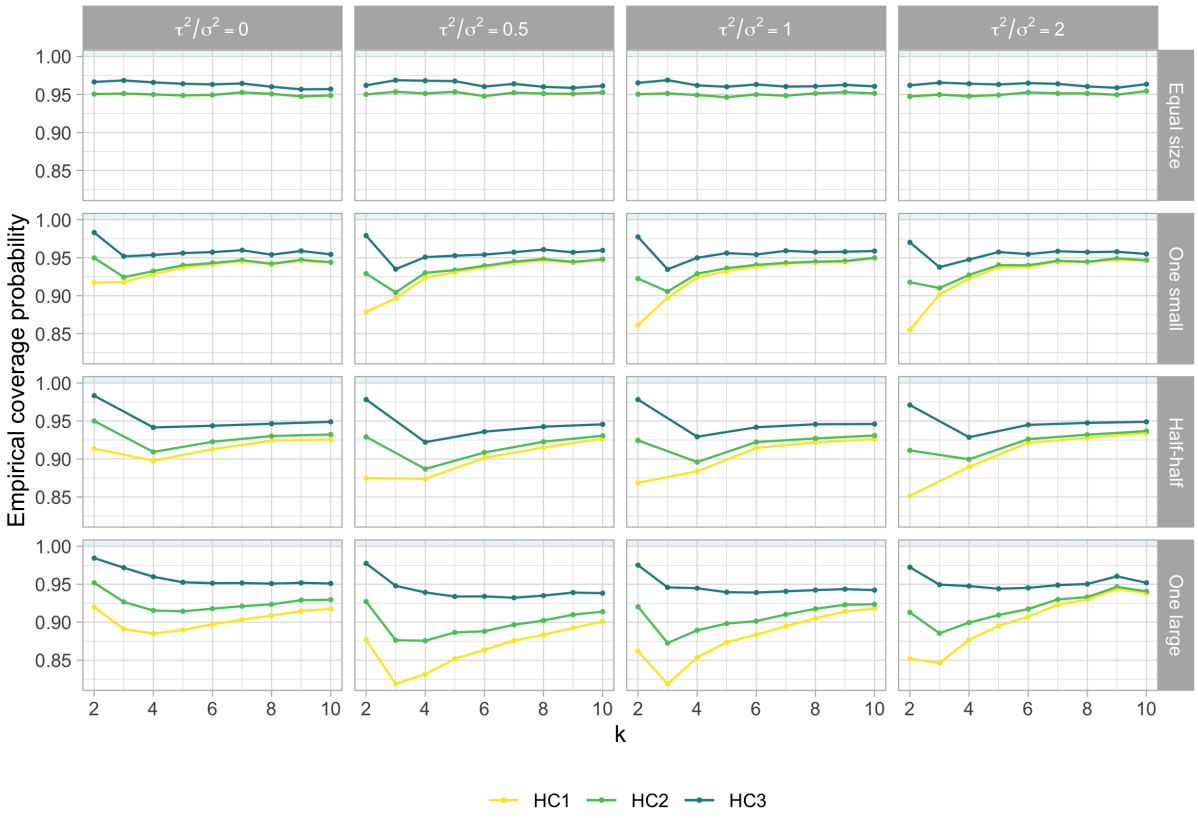

Figure 2.2. Empirical Coverage Probability of 95% Confidence Intervals for δ. This plot displays the coverage probability of 95% confidence intervals for the true effect ($\delta = 0.5$) using HC1, HC2 and HC3 for variance estimation along with a t-distribution as a reference distribution with $k - 1$ degrees of freedom. The columns represent different levels of heterogeneity. The rows represent balanced and unbalanced configurations. The number of studies ranges from $k = 2$ to $k = 10$.

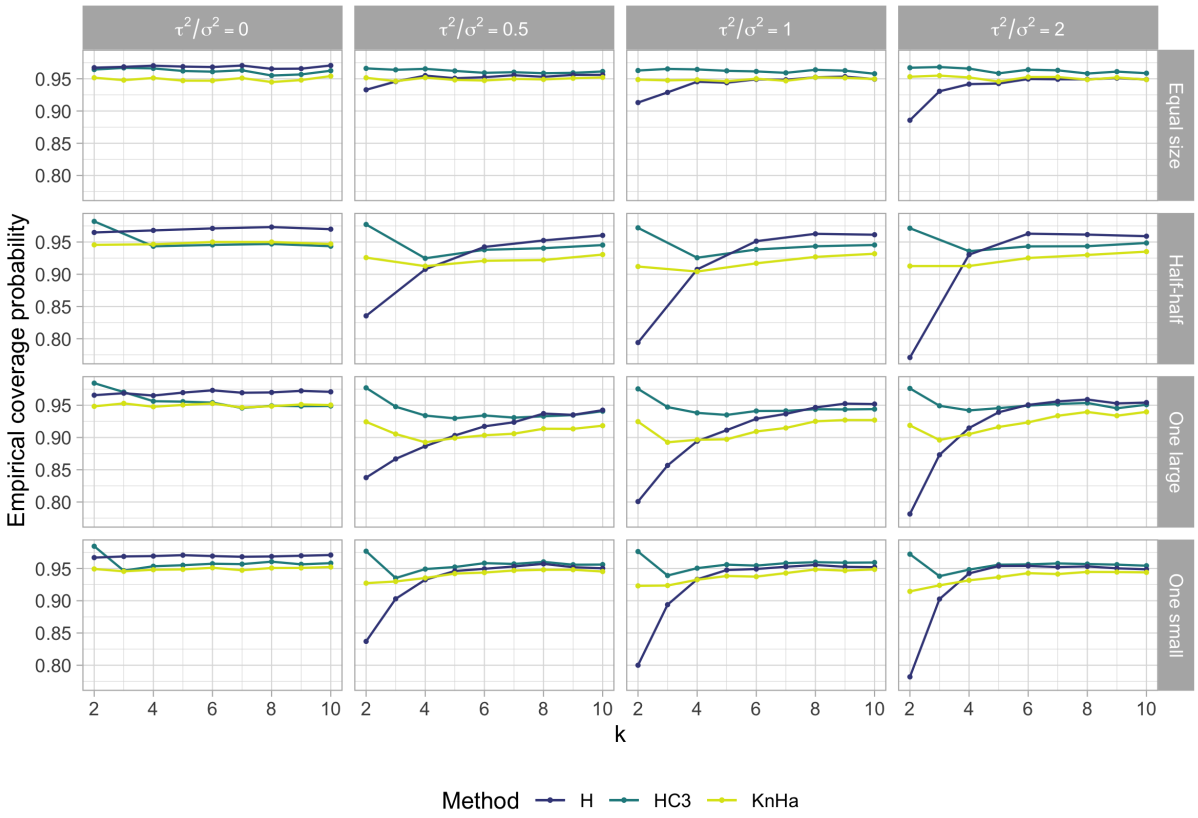

Figure 2.3. Empirical Coverage Probability of 95% Confidence Intervals for δ. This plot displays the coverage probability of the best performing methods of 95% confidence intervals for the true effect ($\delta = 0.5$). The columns represent different levels of heterogeneity. The rows represent balanced and unbalanced configurations. The number of studies ranges from $k = 2$ to $k = 10$.

Table 2.1. Ratio of Model Variance to Empirical Variance for the Standardized Mean Difference: Equal Size Studies

Pattern	τ^2/σ^2	k	Z	KnHa	H	HC1	HC2	HC3
Equal size	0.0	2	1.47	1.00	1.47	1.01	1.01	2.03
Equal size	0.0	4	1.28	1.00	1.28	1.01	1.01	1.35
Equal size	0.0	6	1.20	0.99	1.20	1.02	1.02	1.22
Equal size	0.0	8	1.21	0.99	1.21	1.00	1.00	1.14
Equal size	0.0	10	1.17	0.98	1.17	0.99	0.99	1.10
Equal size	0.5	2	1.27	1.02	1.27	0.99	0.99	1.97
Equal size	0.5	4	1.15	1.00	1.15	0.99	0.99	1.32
Equal size	0.5	6	1.04	0.99	1.04	0.98	0.98	1.18
Equal size	0.5	8	1.05	1.01	1.06	1.01	1.01	1.16
Equal size	0.5	10	1.05	1.00	1.05	1.03	1.03	1.14
Equal size	1.0	2	1.18	0.99	1.18	1.02	1.02	2.05
Equal size	1.0	4	1.09	1.01	1.09	0.99	0.99	1.33
Equal size	1.0	6	1.07	1.00	1.07	0.98	0.98	1.18
Equal size	1.0	8	1.03	0.99	1.03	1.01	1.01	1.16
Equal size	1.0	10	1.00	1.00	1.00	1.01	1.01	1.12
Equal size	2.0	2	1.08	0.97	1.08	0.96	0.96	1.91
Equal size	2.0	4	1.06	1.00	1.06	0.96	0.96	1.28
Equal size	2.0	6	1.01	1.00	1.01	1.00	1.00	1.20
Equal size	2.0	8	1.00	0.99	1.00	0.99	0.99	1.13
Equal size	2.0	10	0.96	0.98	0.96	1.01	1.01	1.12

Note:
This table displays the ratio of the model variance to the empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC1), 5. Heteroscedasticity-consistent variance (HC2) and 6. Heteroscedasticity-consistent variance (HC3). Columns correspond to the different methods. Rows correspond to the number of studies included in a meta-analysis for four levels of true heterogeneity.

Table 2.2. Ratio of Model Variance to Empirical Variance for the Standardized Mean Difference: One Small Study

Pattern	τ^2/σ^2	k	Z	KnHa	H	HC1	HC2	HC3
One small	0.0	2	1.64	1.36	1.64	1.16	1.38	4.56
One small	0.0	4	1.36	1.16	1.48	0.97	1.03	1.46
One small	0.0	6	1.27	1.08	1.31	0.97	0.99	1.23
One small	0.0	8	1.22	1.02	1.24	0.96	0.98	1.13
One small	0.0	10	1.18	1.03	1.20	0.97	0.98	1.10
One small	0.5	2	1.11	1.02	1.11	0.92	1.02	2.85
One small	0.5	4	1.03	0.99	1.07	0.94	0.99	1.38
One small	0.5	6	1.02	1.00	1.04	0.97	0.99	1.21
One small	0.5	8	1.04	1.02	1.05	0.99	1.01	1.16
One small	0.5	10	1.01	0.99	1.01	1.00	1.01	1.13
One small	1.0	2	1.06	0.96	1.06	0.85	0.92	2.37
One small	1.0	4	0.97	0.95	0.99	0.94	0.97	1.34
One small	1.0	6	0.98	0.98	0.99	0.97	0.98	1.20
One small	1.0	8	0.98	1.02	0.99	0.98	0.98	1.14
One small	1.0	10	1.01	0.97	1.01	1.02	1.03	1.15
One small	2.0	2	1.01	0.94	1.01	0.91	0.96	2.22
One small	2.0	4	0.98	0.92	0.97	0.90	0.92	1.25
One small	2.0	6	0.95	1.01	0.94	0.94	0.95	1.15
One small	2.0	8	0.98	1.01	0.98	0.99	0.99	1.14
One small	2.0	10	1.02	0.99	1.02	0.98	0.98	1.10

Note:
This table displays the ratio of the model variance to the empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC1), 5. Heteroscedasticity-consistent variance (HC2) and 6. Heteroscedasticity-consistent variance (HC3). Columns correspond to the different methods. Rows correspond to the number of studies included in a meta-analysis for four levels of true heterogeneity.

Table 2.3. Ratio of Model Variance to Empirical Variance for the Standardized Mean Difference: Half-half Large and Small Studies

Pattern	τ^2/σ^2	k	Z	KnHa	H	HC1	HC2	HC3
Half-half	0.0	2	1.67	1.36	1.67	1.16	1.38	4.55
Half-half	0.0	4	1.38	1.29	1.63	0.94	1.05	1.63
Half-half	0.0	6	1.30	1.22	1.48	0.93	1.01	1.33
Half-half	0.0	8	1.24	1.20	1.37	0.95	1.02	1.25
Half-half	0.0	10	1.26	1.17	1.37	0.93	0.99	1.16
Half-half	0.5	2	1.15	1.02	1.15	0.91	1.02	2.85
Half-half	0.5	4	0.96	0.95	1.01	0.82	0.88	1.28
Half-half	0.5	6	0.94	0.93	0.96	0.88	0.92	1.17
Half-half	0.5	8	0.95	0.96	0.97	0.89	0.93	1.11
Half-half	0.5	10	0.94	0.94	0.96	0.95	0.98	1.12
Half-half	1.0	2	1.07	0.96	1.07	0.87	0.95	2.43
Half-half	1.0	4	0.93	0.89	0.94	0.84	0.87	1.24
Half-half	1.0	6	0.93	0.91	0.93	0.90	0.92	1.15
Half-half	1.0	8	0.92	0.94	0.92	0.91	0.93	1.09
Half-half	1.0	10	0.96	0.93	0.95	0.92	0.94	1.06
Half-half	2.0	2	0.97	0.94	0.97	0.90	0.94	2.21
Half-half	2.0	4	0.95	0.91	0.94	0.82	0.84	1.16
Half-half	2.0	6	0.97	0.93	0.95	0.91	0.92	1.13
Half-half	2.0	8	0.95	0.97	0.93	0.92	0.93	1.08
Half-half	2.0	10	0.96	0.95	0.94	0.94	0.94	1.06

Note:
This table displays the ratio of the model variance to the empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC1), 5. Heteroscedasticity-consistent variance (HC2) and 6. Heteroscedasticity-consistent variance (HC3). Columns correspond to the different methods. Rows correspond to the number of studies included in a meta-analysis for four levels of true heterogeneity.

Table 2.4. Ratio of Model Variance to Empirical Variance for the Standardized Mean Difference: One Large Study

Pattern	τ^2/σ^2	k	Z	KnHa	H	HC1	HC2	HC3
One large	0.0	2	1.64	1.36	1.64	1.13	1.34	4.41
One large	0.0	4	1.35	1.21	1.44	0.95	1.12	2.21
One large	0.0	6	1.28	1.12	1.35	0.93	1.06	1.69
One large	0.0	8	1.24	1.10	1.30	0.94	1.06	1.50
One large	0.0	10	1.21	1.09	1.26	0.93	1.03	1.36
One large	0.5	2	1.11	1.02	1.11	0.95	1.06	2.91
One large	0.5	4	0.96	0.88	0.98	0.80	0.87	1.43
One large	0.5	6	0.91	0.91	0.93	0.84	0.90	1.23
One large	0.5	8	0.90	0.88	0.92	0.83	0.87	1.10
One large	0.5	10	0.89	0.90	0.90	0.86	0.90	1.06
One large	1.0	2	1.04	0.96	1.04	0.86	0.94	2.43
One large	1.0	4	0.91	0.91	0.92	0.85	0.89	1.34
One large	1.0	6	0.92	0.92	0.93	0.88	0.90	1.16
One large	1.0	8	0.93	0.96	0.93	0.89	0.92	1.09
One large	1.0	10	0.95	0.93	0.94	0.89	0.90	1.03
One large	2.0	2	0.99	0.94	0.99	0.90	0.95	2.20
One large	2.0	4	0.93	0.92	0.93	0.89	0.91	1.29
One large	2.0	6	0.95	0.93	0.95	0.90	0.91	1.12
One large	2.0	8	0.96	0.96	0.95	0.95	0.96	1.11
One large	2.0	10	0.99	0.94	0.99	0.97	0.97	1.08

Note:
This table displays the ratio of the model variance to the empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC1), 5. Heteroscedasticity-consistent variance (HC2) and 6. Heteroscedasticity-consistent variance (HC3). Columns correspond to the different methods. Rows correspond to the number of studies included in a meta-analysis for four levels of true heterogeneity.

Table 2.5. Empirical Coverage Probability of 95% Confidence Intervals for the Standardized Mean Difference: Equal Size Studies

Pattern	τ^2/σ^2	k	HC2	HC2a	HC3
Equal size	0.0	2	0.952	0.952	0.964
Equal size	0.0	4	0.951	0.951	0.966
Equal size	0.0	6	0.947	0.947	0.961
Equal size	0.0	8	0.945	0.945	0.955
Equal size	0.0	10	0.954	0.954	0.963
Equal size	0.5	2	0.952	0.952	0.966
Equal size	0.5	4	0.952	0.952	0.965
Equal size	0.5	6	0.947	0.947	0.959
Equal size	0.5	8	0.949	0.949	0.959
Equal size	0.5	10	0.952	0.952	0.961
Equal size	1.0	2	0.949	0.949	0.963
Equal size	1.0	4	0.949	0.949	0.965
Equal size	1.0	6	0.950	0.950	0.961
Equal size	1.0	8	0.952	0.952	0.964
Equal size	1.0	10	0.950	0.950	0.958
Equal size	2.0	2	0.953	0.953	0.967
Equal size	2.0	4	0.952	0.952	0.966
Equal size	2.0	6	0.953	0.953	0.964
Equal size	2.0	8	0.949	0.949	0.958
Equal size	2.0	10	0.949	0.949	0.959

Note:
Confidence intervals were computed using the following methods: 1. Heteroscedasticity-consistent variance HC2, 2. Heteroscedasticity-consistent variance HC2 with adjusted degrees of freedom, and 3. Heteroscedasticity-consistent variance HC3. Columns correspond to the three methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for four levels of heterogeneity.

Table 2.6. Empirical Coverage Probability of 95% Confidence Intervals for the Standardized Mean Difference: One Small Study

Pattern	τ^2/σ^2	k	HC2	HC2a	HC3
One small	0.0	2	0.949	0.949	0.985
One small	0.0	4	0.933	0.959	0.954
One small	0.0	6	0.943	0.952	0.957
One small	0.0	8	0.949	0.954	0.961
One small	0.0	10	0.950	0.953	0.958
One small	0.5	2	0.927	0.927	0.977
One small	0.5	4	0.928	0.953	0.949
One small	0.5	6	0.944	0.952	0.958
One small	0.5	8	0.949	0.954	0.960
One small	0.5	10	0.946	0.949	0.956
One small	1.0	2	0.923	0.923	0.976
One small	1.0	4	0.930	0.953	0.951
One small	1.0	6	0.938	0.946	0.955
One small	1.0	8	0.950	0.954	0.960
One small	1.0	10	0.949	0.952	0.959
One small	2.0	2	0.914	0.914	0.972
One small	2.0	4	0.927	0.948	0.948
One small	2.0	6	0.942	0.948	0.956
One small	2.0	8	0.945	0.949	0.957
One small	2.0	10	0.945	0.947	0.954

Note:
Confidence intervals were computed using the following methods: 1. Heteroscedasticity-consistent variance HC2, 2. Heteroscedasticity-consistent variance HC2 with adjusted degrees of freedom, and 3. Heteroscedasticity-consistent variance HC3. Columns correspond to the three methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for four levels of heterogeneity.

Table 2.7. Empirical Coverage Probability of 95% Confidence Intervals for the Standardized Mean Difference: Half-half Large and Small Studies

Pattern	τ^2/σ^2	k	HC2	HC2a	HC3
Half-half	0.0	2	0.946	0.946	0.982
Half-half	0.0	4	0.913	0.985	0.943
Half-half	0.0	6	0.920	0.967	0.946
Half-half	0.0	8	0.931	0.957	0.947
Half-half	0.0	10	0.929	0.950	0.944
Half-half	0.5	2	0.926	0.926	0.977
Half-half	0.5	4	0.893	0.969	0.925
Half-half	0.5	6	0.914	0.953	0.938
Half-half	0.5	8	0.923	0.946	0.940
Half-half	0.5	10	0.931	0.947	0.945
Half-half	1.0	2	0.912	0.912	0.972
Half-half	1.0	4	0.893	0.966	0.926
Half-half	1.0	6	0.916	0.949	0.938
Half-half	1.0	8	0.927	0.945	0.943
Half-half	1.0	10	0.933	0.945	0.945
Half-half	2.0	2	0.913	0.913	0.971
Half-half	2.0	4	0.906	0.967	0.936
Half-half	2.0	6	0.924	0.946	0.943
Half-half	2.0	8	0.930	0.940	0.944
Half-half	2.0	10	0.934	0.942	0.949

Note:
Confidence intervals were computed using the following methods: 1. Heteroscedasticity-consistent variance HC2, 2. Heteroscedasticity-consistent variance HC2 with adjusted degrees of freedom, and 3. Heteroscedasticity-consistent variance HC3. Columns correspond to the three methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for four levels of heterogeneity.

Table 2.8. Empirical Coverage Probability of 95% Confidence Intervals for the Standardized Mean Difference: One Large Study

Pattern	τ^2/σ^2	k	HC2	HC2a	HC3
One large	0.0	2	0.948	0.948	0.984
One large	0.0	4	0.915	0.981	0.956
One large	0.0	6	0.919	0.984	0.954
One large	0.0	8	0.923	0.978	0.949
One large	0.0	10	0.928	0.974	0.949
One large	0.5	2	0.924	0.924	0.977
One large	0.5	4	0.871	0.963	0.934
One large	0.5	6	0.890	0.964	0.934
One large	0.5	8	0.904	0.957	0.933
One large	0.5	10	0.919	0.955	0.941
One large	1.0	2	0.925	0.925	0.975
One large	1.0	4	0.883	0.963	0.938
One large	1.0	6	0.904	0.961	0.941
One large	1.0	8	0.920	0.953	0.944
One large	1.0	10	0.927	0.948	0.944
One large	2.0	2	0.919	0.919	0.976
One large	2.0	4	0.897	0.959	0.942
One large	2.0	6	0.921	0.954	0.950
One large	2.0	8	0.937	0.953	0.954
One large	2.0	10	0.938	0.946	0.951

Note:
Confidence intervals were computed using the following methods: 1. Heteroscedasticity-consistent variance HC2, 2. Heteroscedasticity-consistent variance HC2 with adjusted degrees of freedom, and 3. Heteroscedasticity-consistent variance HC3. Columns correspond to the three methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for four levels of heterogeneity.

CHAPTER 3

Random-effects meta-analysis of randomized clinical trials when the outcome is binary and the number of studies is small

3.1. Introduction

Systematic reviews and meta-analyses of randomized clinical trials are quite common in the medical literature. If the data consist of 2x2 tables, measures of treatment effectiveness include the risk difference, risk ratio and odds ratio. Traditional meta-analysis methods then pool the data across studies to produce a summary estimate for the gain from treatment. While methods for combining effect sizes from binary data exist, these require a large collection of studies for satisfactory performance. Frequently, however, empirical researchers work in settings with limited data. For example, the majority of meta-analyses in the Cochrane Library - a research clearinghouse whose purpose is to combine the medical evidence and summarize it for policy-makers and practitioners - include only a handful of studies. This small sample problem leaves systematic reviewers with two options: (1) apply readily available meta-analysis methods to the limited data, or (2) conduct a qualitative review only. The issues that arise in such settings suggest a need for further investigation of current methods for evidence synthesis and considerations for improvements that make certain methods more suitable for application to a small collection of studies.

In this article, we are concerned with random-effects meta-analysis methods of randomized clinical trials that synthesize effect sizes based off of binary data. While clinical trials are considered to be the gold standard for determining treatment effectiveness, these are not necessarily designed to answer future research questions that are of interest in meta-analysis. It is therefore reasonable to expect considerable heterogeneity between findings. This could be due to a multitude of reasons, including differences in, but not limited to, the sampled population, how the intervention is delivered and the setting where the experiment was conducted. Random-effects models are more flexible in that these

methods formally incorporate heterogeneity among expected effect sizes, summarized by the heterogeneity variance. Additionally, these models allow us to generalize the results to future populations that differ among themselves in the same manner as those in the primary studies.

The random-effects framework, described in more detail in Chapter 2, is directly applicable to the synthesis of effect sizes based on binary data. While the general framework still holds, it does not immediately follow that the methods commonly used will have desirable statistical properties, particularly when meta-analyses contain small numbers of studies and variable numbers of individuals per study. This chapter addresses these problems by investigating the empirical coverage probabilities of $(1-\alpha)100\%$ confidence intervals for the summary effect, for both common practice random-effects methods and the methods introduced in Chapter 2 that utilize alternative estimators for estimating the variance of the average effect estimate.

The chapter is organized as follows. Section 3.2 provides the theory on the risk difference, odds ratio, and risk ratio effect sizes. Section 3.3 summarizes the variance estimators considered throughout. More details on these estimators can be found in Chapter 2. Section 3.5 reports on evidence from a simulation study. Section 3.6 illustrates the various methods on data from a systematic review published by the Cochrane Library, which focuses on interventions intended to increase registration for organ donation. Finally, Section 3.7 summarizes our findings and concludes.

3.2. Effect sizes in clinical trials with binary outcomes

When the outcome of interest in a primary study is binary, the effect size of interest could be the risk difference, risk ratio or odds ratio. In this section, we provide the theory associated with each of these effects sizes. Suppose we have data from K independent randomized control trials, each having two experimental groups, $i = t, c$, with associated sample sizes n_{tk} and n_{ck}, where $k = 1, ..., K$ studies. These data are displayed in 2x2 contingency tables, where the rows are the experimental groups and the columns are the categories of the binary response variable, Y (such as success and failure). Within each experimental group, the response, $p_{i=k}$, is a binomial proportion. Denote the true proportion by π_{ik}

which is estimated using the sample statistic p_{ik}. It follows that

$$n_{tk}p_{tk} \sim bin(n_{tk}, \pi_{tk}) \tag{3.2.1}$$

$$n_{ck}p_{ck} \sim bin(n_{ck}, \pi_{ck}), \tag{3.2.2}$$

where π_{ik} is the true probability of success and $p_{ik} = m_{ik}/n_{ik}$ with variance $Var(p_{ik}) = \pi_{ik}(1 - \pi_{ik})/n_{ik}$.

3.2.1. Risk difference

The true difference in risks between treatment and control group is defined by

$$RD = \pi_{tk} - \pi ck \tag{3.2.3}$$

and estimated by

$$\widehat{RD} = p_{tk} - p_{ck}. \tag{3.2.4}$$

This estimator is unbiased for RD and its variance is given by

$$Var\widehat{RD} = \frac{\pi_{tk}(1 - \pi_{tk})}{n_{tk}} + \frac{\pi_{ck}(1 - \pi_{ck})}{n_{ck}}. \tag{3.2.5}$$

To estimate the variance, one simply plugs in the sample estimates p_{ik} for π_{ik}. Note that when $m_{ik} = 0$ or $m_{ik} = n_{ik}$, the sample proportion is 0 or 1, and the variance estimate of the risk difference estimate is undefined. To avoid this problem, one simply adds a 0.5 count to the underlying 2x2 table.

3.2.2. Risk ratio

Another parameter of interest based on binary data is the risk ratio (RR), defined by

$$RR = \frac{\pi_{tk}}{\pi_{ck}}, \tag{3.2.6}$$

which can be estimated by the sample statistic $\widehat{RR} = p_{tk}/p_{ck}$. Since the sample relative risk converges faster to the normal distribution on the logarithmic scale, the estimator commonly used is, in fact,

$$log(\widehat{RR}) = log\left(\frac{p_{tk}}{p_{ck}}\right). \tag{3.2.7}$$

Its asymptotic variance is given by

$$Var(log(\widehat{RR})) = \frac{1-\pi_{tk}}{n_{tk}\pi_{tk}} + \frac{1-\pi_{ck}}{n_{ck}\pi_{ck}}, \tag{3.2.8}$$

and is estimated by substituting the sample statistic, p_{ik}, for π_{ik}. As is the case with the risk difference, when p_{ik} is 0 or 1, we add 0.5 to the counts in the 2x2 tables and define a refined estimator

$$log(\widehat{RR}) = log\left[\frac{(m_{tk}+0.5)/(n_{tk}+0.5)}{(m_{ck}+0.5)/(n_{ck}+0.5)}\right] \tag{3.2.9}$$

(PETTIGREW *et al.*, 1986). Its variance is estimated using its asymptotic variance by plugging in p_{ik} for π_{ik}.

3.2.3. Odds ratio

The last effect size considered in this Chapter is the odds ratio (OR), defined by

$$OR = \frac{\pi_{tk}/(1-\pi_{tk})}{\pi_{ck}/(1-\pi_{ck})}, \tag{3.2.10}$$

where $\pi_{ik}/(1-\pi_{ik})$ is the odds of success. An estimate is obtained by substituting sample statistics, p_{ik}, for π_{ik}, however, this estimator is undefined if p_{ik} is either 0 or 1. To overcome this problem, Gart and Zweifel (1967) and Haldane (1956) proposed an amended estimator

$$log(\widehat{OR}) = log\left[\frac{m_{tk}+0.5}{n_{tk}-m_{tk}+0.5}\frac{n_{ck}-m_{ck}+0.5}{m_{ck}+0.5}\right], \tag{3.2.11}$$

derived by adding half counts to the 2x2 table. The variance of the estimate of the $log(OR)$ is

$$Var(log(\widehat{OR})) = \frac{1}{n_{tk}\pi_{tk}(1-\pi_{tk})} + \frac{1}{n_{ck}\pi_{ck}(1-\pi_{ck})}, \tag{3.2.12}$$

and is estimated by substituting p_{ik} for π_{ik}.

3.3. Summary of variance estimators

This section provides an overall summary of the five variance estimators considered for application to the small-sample setting in meta-analysis when the effect size of interest is a risk difference, log risk ratio or log odds ratio. Table 3.1 gives their algebraic form and includes the reference distribution that is used with each estimator to compute confidence intervals for the summary treatment effect. We highlight the fact that the variance given by the Hartung method is more complex, thus we include its general form only. For a more thorough understanding of this estimator, and the Hartung method more generally, see Chapter 2. It is worth emphasizing that this method takes into consideration the variability in weights (i.e. when there are different numbers of observations per study), hence we expect it to perform well in unbalanced situations.

All the estimators assign weights that are inversely proportional to the total variance of the effect estimates, that is, $\hat{w}_i = 1/(\sigma_i^2 + \hat{\tau}^2)$, where $\hat{\tau}^2$ is the estimated heterogeneity variance. Without loss of generality, throughout this chapter we used the DerSimonian and Laird estimator to estimate the between-study variance. While we did examine other alternatives, properties of confidence intervals remained unchanged irrespective of the estimator used to obtain an estimate of τ^2. We suspect that this phenomenon will likely not hold when combining small studies (i.e. $n < 30$), albeit this scenario was not investigated.

Table 3.1. Summary of Variance Estimators. This table gives a summary of the estimators considered in Chapter 2 and reference distributions that were used with each estimator for computing confidence intervals for the summary effect.

Variance estimator	Reference distribution	Acronym
$\frac{1}{\sum_{i=1}^{k} \hat{w}_i}$	Z	Z
$\frac{\sum_{i=1}^{k} \hat{w}_i (y_i - \bar{y}_+)^2}{(k-1)\sum_{i=1}^{k} \hat{w}_i}$	t_{k-1}	KnHa
$L(\beta)Q(\beta) + (1 - L(\beta))R(\beta)$	$t_{\hat{f}}$	H
$\frac{\sum_{i=1}^{k} \hat{w}_i{}^2 (y_i - \bar{y}_+)^2 [1-(\hat{w}_i / \sum_{i=1}^{k} \hat{w}_i)]^{-1}}{(\sum_{i=1}^{k} \hat{w}_i)^2}$	t_{k-1}	HC2
$\frac{\sum_{i=1}^{k} \hat{w}_i{}^2 (y_i - \bar{y}_+)^2 [1-(\hat{w}_i / \sum_{i=1}^{k} \hat{w}_i)]^{-2}}{(\sum_{i=1}^{k} \hat{w}_i)^2}$	t_{k-1}	HC3

3.4. Quantifying heterogeneity: I^2 parameter

A common way of representing true heterogeneity is through the I^2 parameter, defined as

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2}. \tag{3.4.1}$$

I^2 represents the proportion of the between-study variance to the total variance and is particularly appealing because it is independent of the effect size parameter, thus allowing for easier comparisons between the properties of random-effects methods based on the three different effect sizes. For this reason, throughout this chapter, we use I^2 instead of τ^2.

3.5. Simulation evidence

This simulation study assesses the performance of five random-effects methods for all three binary effect sizes, i.e. risk difference, log risk ratio and log odds ratio. In particular, it evaluates the empirical coverage probability of 95% confidence intervals for the summary estimate in meta-analyses containing a small collection of studies and under different configurations of unbalancedeness that reflect the difficulties research clearinghouses face in practice. The simulation study was carried out in R. The reported coverage probabilities were based on $10,000$ replications of the corresponding methods. Hence, the simulation error is approximately $\sqrt{\frac{0.95(1-0.95)}{100}} = 0.0022$.

3.5.1. Simulation design

The data generation process according to a random effects procedure included two steps. First, we generated the true study specific probabilities, π_{ik}, for $i = c, t$ and $k = 1, ..., K$. Next, we generated the observed probabilities, p_{ik}, for both experimental groups in all K studies and calculated the effect size estimates and their variances according to the formulas presented in Section 3.2. In the first stage, we drew the study specific probabilities from the Beta distribution which ensures that the values are between 0 and 1 (Emerson *et al.*, n.d.; Tipton, 2013). In the second stage, we drew the observed probabilities from the binomial distribution with parameters n_{ik} and π_{ik}, using the *rbinom* function in R. Because the goal in primary studies is to design balanced experiments, we made the simplifying assumption that $n_{ck} = n_{tk} = n_k$ for all k. Note, however, that in practice there is substantial variability in sample sizes across studies. To reflect this, the meta-analyses included balanced and unbalanced configurations of within-study sample sizes.

One set of simulations included analyses where sample sizes are equal across studies. Another set of simulations included analyses where one study was large and the rest small, one study was small and the rest large, and half-half large and small studies. The ratio of study sample sizes in the unbalanced setting was 10, that is, one study was 10 times smaller (or larger) than the rest of the studies in the analysis. The number of studies ranged from 2 to 10, reflecting typical values found in meta-analyses conducted by research clearinghouses.

We focused on three levels of the heterogeneity parameter, $I^2 = (0, 0.33, 0.50)$, and three pairs of true underlying probabilities, $(\pi_c, \pi_t) = (0.10, 0.10), (0.50, 0.50)$ and $(0.30, 0.40)$, that have been previously used in other simulation studies (Emerson *et al.*, n.d.; Hartung and Knapp, 2001a; Tipton, 2013). For each combination of the parameters, we simulated $10,000$ meta-analyses and performed the random-effects analyses discussed in Chapter 2 and summarized in Section 3.3.

3.5.2. Simulation results

The results of the simulation study are separated based on effect size. Table 3.1, 3.2 and 3.3 report the empirical coverage probabilities of 95% confidence intervals for the risk difference effect size, the log risk

ratio, and the log odds ratio, respectively. In these tables, we first note the different patterns (equal sized studies, one small study, one large study, and half-half large and small studies). Within each pattern, we reported the results for three levels of the heterogeneity parameter ($I^2 = (0, 0.33, 0.50)$ and for k ranging from 2 to 10. Note also that the columns are separated by the true underlying probabilities considered.

For each pair of true rates, empirical coverage probabilities are given for the standard Z method, the Hartung-Knapp method - which uses the t-distribution instead of the normal distribution for constructing confidence intervals, the more general method developed by Hartung that takes into account the variability in weights across studies, and HC2 and HC3 which correspond to robust variance estimators for estimating the variance of the summary estimate. Remember that the only difference between HC2 and HC3 is the adjustment made to the squared sample residuals. Based on the analytic results from Chapter 2 on robust variance estimators, we expect confidence intervals based on HC3 to be wider compared to the ones that use HC2 to estimate the variance of the average effect.

3.5.3. Results for the risk difference effect size

3.5.3.1. Case of $(\pi_c, \pi_t) = (0.10, 0.10)$ and $(\pi_c, \pi_t) = (0.50, 0.50)$. First, we report the results of the simulations for the risk difference effect size. We considered true probabilities $(\pi_c, \pi_t) = (0.10, 0.10)$ and $(\pi_c, \pi_t) = (0.50, 0.50)$ which captures both small and moderate rates and corresponds to a true risk difference of 0. Let us first discuss the equal sized studies case. When there is no heterogeneity ($I^2 = 0$) - where, theoretically, the fixed effects approach is appropriate - all the methods produced coverage rates close to the nominal value of 0.95 or lead to slight overcoverage (by about 1 percentage point). As heterogeneity increases, coverage rates for the standard Z method decrease, a point that has previously been noted in the literature. The KnHa and HC2 lead to nominal coverage probabilities when there is moderate to high heterogeneity and for k as small as 2, whereas HC3 is slightly conservative, but by only 1 percentage point.

When there is only one small study while the rest are large and there is no heterogeneity, all but HC2 lead to nominal coverage rates or slight overcoverage, while HC2 leads to slight undercoverage when $k = 4$, but tends to the nominal value as the number of studies increases. As heterogeneity increases, it

has already been observed that the standard Z method performs poorly, and we observe this phenomenon here as well; however, as k increases, coverage tends to 0.95. The KnHa and HC2 methods are quite similar and both lead to slight undercoverage. In the extreme case of $k = 2$, the undercoverage is by about 2 percentage points. The Hartung method, performs poorly when $k = 2$. We suspect this is because estimates of the between-study variance parameter are unstable in this case. Coverage does tend to the nominal value as k increases. By $k = 4$, coverage is already roughly 0.93. HC3, on the other hand, performs well in all scenarios, leading to nominal coverage in cases of small to large heterogeneity for all k except for the extreme case of $k = 2$. In this instance, the method is conservative.

When the pattern consists of one large study or half-half large and small studies, we expect coverage rates to be further from the nominal value. We do observe this phenomenon in this simulation study, except when I^2 is 0. When $I^2 = 0$, all but HC2 perform well. HC3 performs best when the pattern consists of one large study, leading to nominal coverage rates in all cases except for the case of $k = 2$, which is conservative by about 3 percentage points. When half the studies are large and half are small, HC3 and the Hartung method perform similarly when k is 4 or greater, and outperform all other methods. When $k = 2$, HC3 is conservative while the Hartung method leads to more severe undercoverage (by at most 9 percentage points when $k = 2$ and when there is a large degree of heterogeneity).

3.5.3.2. Case of $(\pi_c, \pi_t) = (0.30, 0.40)$. When the true underlying probabilities are $(\pi_c, \pi_t) = (0.30, 0.40)$, corresponding to a risk difference of 0.10, we observe similar patterns when we synthesize equal sized studies. The Hartung method leads to undercoverage when $k = 2$, but yields approximately nominal values by $k = 4$. KnHa and HC2 produce coverage rates that are approximately nominal for all levels of I^2 and all k, while HC3 leads to slight overcoverage, by only about 1 percentage point.

When a meta-analysis includes one small study, all methods perform similarly except for the case of $k = 2$, where HC3 is more conservative (by about 3 percentage points). As heterogeneity increases, all but HC3 lead to undercoverage. KnHa and HC2 perform similarly for all levels of heterogeneity and for k ranging from 2 to 10. These two methods produce coverage rates that are approximately nominal, but are anticonservative (by about 2 percentage points) in the $k = 2$ case, while HC2 is conservative (by approximately 3 percentage points). By $k = 4$, HC3 leads to nominal coverage rates.

The most problematic case is when the pattern consists of one large study (and the rest are small) for moderate to high heterogeneity. When $I^2 = 0$, most methods perform well. HC2 leads to undercoverage even when there is no true heterogeneity. As heterogeneity increases, however, HC3 outperforms all other methods, particularly when $k \leq 6$, but performs similarly to the Hartung method when k is larger. In the case of half-half large and small studies, KnHa and HC3 perform similarly and outperform the rest, except in the case of $k = 2$, where KnHa leads to undercoverage while HC3 leads to overcoverage.

3.5.4. Results for the log risk ratio effect size

3.5.4.1. Case of $(\pi_c, \pi_t) = (0.10, 0.10)$ and $(\pi_c, \pi_t) = (0.50, 0.50)$. Results for the log risk ratio effect size show similar trends compared to the risk difference. Not surprisingly, the standard Z method performs the least favorable for moderate to high degrees of heterogeneity. This trend is similar across both pairs of true probabilities and for all study patterns. There is slight improvement in coverage in the case of equal sized studies and one small study patterns compared to half-half large and small studies and one large study when I^2 is moderate to large and $k > 2$, by approximately 5 percentage points. HC2 is comparable to KnHa for equal sized studies and one small study patterns, however, in the case of half-half large and small studies and one large study, KnHa is superior to HC2. The Hartung method, on the other hand, performs poorly when k is small (i.e. less than 4), but tends to the nominal value as the number of studies increases. HC3 is comparable to KnHa in many of the scenarios considered, and superior to it in the case of half-half large and small and one large study patterns.

3.5.4.2. Case of $(\pi_c, \pi_t) = (0.30, 0.40)$. When the true probabilities are $(\pi_c, \pi_t) = (0.30, 0.40)$, similar patterns are observed in the case of equal sized study patterns. Once again, the standard Z method performs poorly across scenarios considered, with the exception of $I^2 = 0$. Coverage decreases when the weights are variable across studies, with more problematic scenarios being half-half large and small and one large study patterns. The Hartung method leads to undercoverage for very small k, but converges to the nominal value as k increases. KnHa produces nominal coverage for equal weights, however, coverage decreases as weights become more variable. This is more evident when synthesizing half-half large and small studies, and in particular, when meta-analyses contain one large study. In these scenarios, HC3

outperforms all other alternatives and produces approximately nominal coverage. When the studies are balanced, HC3 leads to slight overcoverage, but by only 1 percentage point.

3.5.5. Results for the log odds ratio effect size

3.5.5.1. Case of $(\pi_c, \pi_t) = (0.10, 0.10)$ and $(\pi_c, \pi_t) = (0.50, 0.50)$. When the true underlying probabilities are equal across experimental groups, coverage probabilities for the log odds ratio effect size are similar to both the risk difference and log risk ratio effect sizes. In particular, it is observed that the standard Z method performs poorly across all study size patterns and when $I^2 > 0$. When $I^2 = 0$, this method leads to nominal coverage for all configurations of sample sizes. The Hartung method performs poorly when $k = 2$, but by $k = 4$, the method leads to the nominal value of 0.95, except in the case of one large study. KnHa performs well when the weights are equal across studies, however, coverage decreases as weights become more variable, particularly in cases of one large study. HC2 performs similarly to the other methods with equal weights, but once again coverage decreases when studies are unbalanced. On the other hand, HC3 leads to approximately nominal coverage across all scenarios. In the case of equal sized studies and one small study, HC3 is comparable to KnHa, but in meta-analyses containing one large study or half-half large and small studies, HC3 is superior. This is in line with the previous observations on the risk difference and log risk ratio effect sizes.

3.5.5.2. Case of $(\pi_c, \pi_t) = (0.30, 0.40)$. When the true rates differ between the experimental groups, the standard Z method performs poorly across most scenarios, as is expected, except in the case of no heterogeneity. KnHa shows satisfactory performance in situations where the studies are balanced, but tend to perform less satisfactorily as studies become more unbalanced. This is particularly the case when meta-analyses contain one large study. Across all scenarios considered, HC3 is comparable or superior to all other methods. This is especially true in one large study patterns. When the studies are balanced, HC3 leads to slight overcoverage, but by only 1 percentage point. When $k = 2$, HC3 is conservative throughout by about 3 percentage points.

3.6. Application to a meta-analysis of interventions aimed at increasing organ donor registration

In this section, we consider an application of the various random-effects methods to the synthesis of studies that investigated the effectiveness of any intervention intended to increase organ donor registration, carried out by Li *et al.* (2021) as part of the Cochrane Collaboration library of systematic reviews. The primary outcome was *registration behavior*. The type of intervention and mode of delivery varied across studies, as did the setting and sample sizes in the experimental groups. Heterogeneity is therefore anticipated and is appropriately quantified.

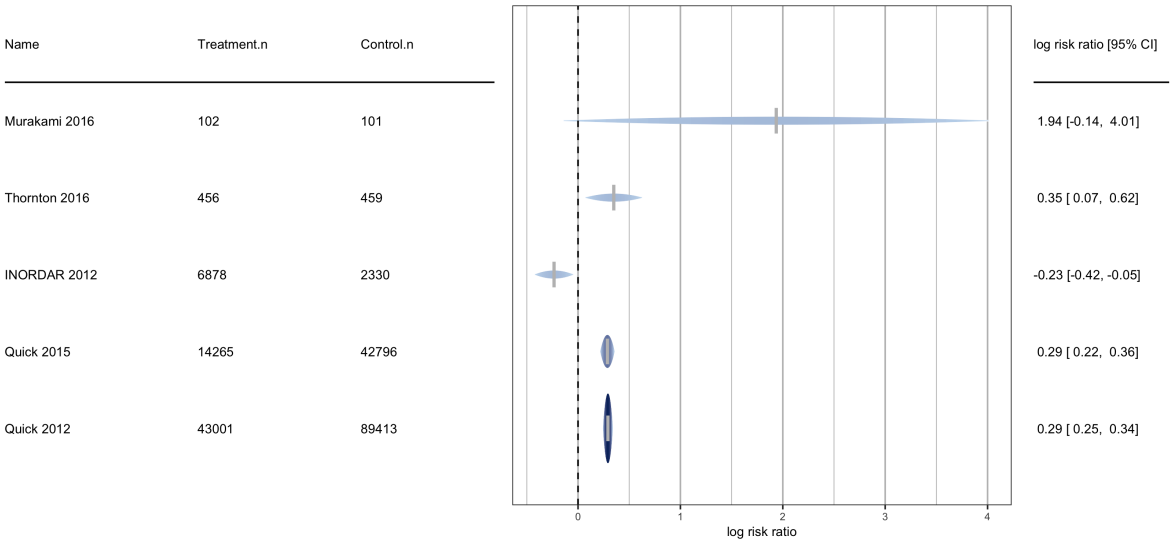

Figure 3.1. Data on Interventions to Increase Organ Donor Registration. This rain forest plot shows data from five studies reporting the effectiveness of interventions intended to increase organ donor registration versus a control. It reports the study specific log risk ratio estimates, $log(\widehat{OR})$, and associated 95% confidence intervals.

We analyze part of the data from Analysis 1.1 of the report which presents risk ratio estimates and their associated 95% confidence intervals, comparing treatments intended to increase organ donor registration to a control. In this example, we include randomized trials only and exclude any other study designs. While it is certainly possible to include different study designs (such as cluster randomized trials and quasi-randomized experiments), the nuances they introduce are not the purpose of this example.

The rain forest plot containing the log risk ratio estimates and 95% confidence intervals is displayed in Figure 3.1. It also includes sample sizes of experimental groups for all studies. Because there is quite a wide range of total sample sizes across studies, we expect the different random-effects methods to produce different results.

We performed random-effects analyses on the data using the following five methods: Z, KnHa, H, HC2 and HC3. The results are summarized in Table 3.2. First, note that the estimate of the heterogeneity variance was estimated as $\hat{\tau}^2 = 0.0206$ with an estimated standard error of $SE = 0.0258$. The estimated I^2 was 87.53%, indicating substantial heterogeneity among findings. Turning to the results obtained, if we use the standard Z method, which uses the normal distribution as a reference distribution, we obtain the narrowest interval that does not include 0 (favoring interventions aimed at increasing organ donor registration). This approach, however, is not acceptable due to its poor coverage rate.

Table 3.2. 95% Confidence Intervals for μ.This table displays 95% confidence intervals for μ using the data on interventions aiming to increase organ donor registration. The rows correspond to the methods used to compute confidence intervals.

			Confidence interval	
Method	Distribution (df)	Summary(SE)	LB	UB
Z	Normal	0.19(0.0810)	0.0340	0.3514
KnHa	t_{k-1}	0.19(0.1299)	-0.1680	0.5534
H	$t_{\hat{f}}$	0.19(0.1455)	-0.1102	0.4956
HC2	t_{k-1}	0.19(0.1219)	-0.1459	0.5313
HC3	t_{k-1}	0.19(0.1393)	-0.1939	0.5793

i. $\hat{\tau}^2_{DL} = 0.0206$, $SE = 0.0258$
ii. $\hat{I}^2 = 87.53\%$.

The other possibilities utilize the t-distribution as a reference distribution and give confidence intervals that are larger in width compared to the standard Z method. They differ moderately in width, with HC3 producing the widest confidence interval. In this particular example, we are able to draw the same conclusion irrespective of the method we use to compute confidence intervals (if we use the t-distribution as a reference distribution). This, however, does not necessarily always follow.

3.7. Summary and discussion

In this Chapter, we provided further insight into various random-effects meta-analysis methods that are common practice when effect sizes are based on binary data, including the risk difference, log risk ratio, and log odds ratio. A special focus was on understanding what effect do variable weights have on properties of the summary estimate, its precision and whether differences in weights substantially impact coverage probabilities of confidence intervals for the average effect. In Section 3.2, we described the three effect sizes considered, along with their point estimators and asymptotic variances. In Section 3.4, we defined the I^2 parameter which we used to quantify the degree of heterogeneity in the simulation study. In Section 3.5, we conducted extensive simulations for the three different effect sizes, under a range of scenarios that are useful for practice. Section 3.6 illustrated the methods on an example using published data from the Cochrane Library.

The results from this chapter provide empirical evidence about the utility of many random-effects methods as applied to only a handful of studies, including methods proposed in Chapter 2 that use robust variance estimators for performing hypothesis testing and computing confidence intervals for the summary treatment effect. As the example illustrates, using the different procedures can be consequential for drawing conclusions from meta-analysis, particularly when sample sizes per study are significantly different. While a method that considers the variability in sample sizes has previously been proposed by Hartung, this chapter shows that in cases of small meta-analyses (i.e. small k), the method performs unsatisfactorily. HC3, on the other hand, is either comparable to all other methods considered, or outperforms them. Specifically, with the exception of a very small number of studies (i.e. $k = 2$) where this method leads to overcoverage, in all other scenarios, it generally leads to approximately nominal coverage.

Table 3.3. Empirical Coverage Probability of 95% Confidence Intervals for the Risk Difference: Equal Size Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Equal size	0.00	2	**0.9613**	0.9437	**0.9659**	0.9437	**0.9597**	**0.9630**	**0.9545**	**0.9677**	**0.9545**	**0.9670**	**0.9612**	**0.9509**	**0.9661**	**0.9509**	**0.9644**
		4	**0.9621**	0.9442	**0.9677**	0.9438	**0.9621**	**0.9594**	0.9449	**0.9662**	0.9448	**0.9610**	**0.9634**	0.9481	**0.9703**	0.9482	**0.9611**
		6	**0.9640**	**0.9511**	**0.9701**	**0.9509**	**0.9640**	**0.9619**	**0.9508**	**0.9697**	**0.9508**	**0.9624**	**0.9620**	**0.9509**	**0.9689**	**0.9509**	**0.9635**
		8	**0.9607**	**0.9534**	**0.9676**	**0.9536**	**0.9630**	**0.9609**	**0.9505**	**0.9668**	**0.9505**	**0.9611**	**0.9605**	0.9481	**0.9660**	0.9482	**0.9590**
		10	**0.9602**	0.9486	**0.9661**	0.9490	**0.9593**	**0.9605**	0.9481	**0.9660**	0.9481	**0.9568**	**0.9648**	**0.9535**	**0.9707**	**0.9535**	**0.9620**
	0.33	2	0.9237	0.9495	0.9349	0.9495	**0.9649**	0.9207	0.9450	0.9314	0.9450	**0.9621**	0.9242	0.9470	0.9356	0.9470	**0.9606**
		4	0.9346	0.9481	**0.9543**	0.9477	**0.9640**	0.9316	**0.9508**	**0.9514**	**0.9507**	**0.9661**	0.9329	**0.9518**	**0.9547**	**0.9519**	**0.9672**
		6	0.9312	0.9495	**0.9511**	0.9495	**0.9620**	0.9327	**0.9509**	**0.9538**	**0.9509**	**0.9616**	0.9348	0.9494	**0.9543**	0.9494	**0.9615**
		8	0.9379	**0.9518**	**0.9558**	**0.9517**	**0.9614**	0.9349	0.9487	**0.9540**	0.9487	**0.9593**	0.9333	0.9486	**0.9534**	0.9486	**0.9582**
		10	0.9381	0.9497	**0.9547**	0.9496	**0.9592**	0.9303	0.9448	**0.9505**	0.9449	**0.9553**	0.9386	0.9485	**0.9544**	0.9486	**0.9596**
	0.50	2	0.8914	0.9476	0.9116	0.9476	**0.9638**	0.8935	**0.9506**	0.9133	**0.9506**	**0.9644**	0.8899	0.9455	0.9111	0.9455	**0.9608**
		4	0.9109	**0.9502**	0.9440	0.9498	**0.9642**	0.9139	0.9480	0.9453	0.9480	**0.9655**	0.9139	**0.9509**	0.9455	**0.9508**	**0.9641**
		6	0.9179	0.9499	**0.9506**	0.9491	**0.9621**	0.9167	0.9475	0.9492	0.9475	**0.9617**	0.9198	0.9486	0.9497	0.9487	**0.9621**
		8	0.9232	**0.9511**	**0.9509**	**0.9511**	**0.9616**	0.9243	**0.9506**	**0.9511**	**0.9506**	**0.9606**	0.9244	0.9487	0.9485	0.9486	**0.9609**
		10	0.9316	**0.9523**	**0.9530**	**0.9514**	**0.9602**	0.9292	**0.9507**	**0.9530**	**0.9508**	**0.9608**	0.9257	0.9497	**0.9510**	0.9498	**0.9577**

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.4. Empirical Coverage Probability of 95% Confidence Intervals for the Risk Difference: One Small Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One small	0.00	2	**0.9652**	**0.9528**	**0.9665**	**0.9528**	**0.9862**	**0.9648**	**0.9550**	**0.9660**	**0.9550**	**0.9886**	**0.9685**	**0.9504**	**0.9696**	**0.9504**	**0.9869**
		4	**0.9651**	**0.9508**	**0.9707**	0.9339	**0.9529**	**0.9636**	0.9497	**0.9702**	0.9312	**0.9516**	**0.9618**	0.9479	**0.9669**	0.9328	**0.9541**
		6	**0.9614**	**0.9503**	**0.9685**	0.9422	**0.9577**	**0.9601**	0.9476	**0.9677**	0.9398	**0.9555**	**0.9626**	0.9465	**0.9690**	0.9404	**0.9539**
		8	**0.9630**	**0.9547**	**0.9708**	**0.9502**	**0.9604**	**0.9608**	0.9487	**0.9673**	0.9430	**0.9558**	**0.9612**	**0.9504**	**0.9687**	0.9476	**0.9591**
		10	**0.9611**	0.9475	**0.9684**	0.9447	**0.9569**	**0.9596**	0.9457	**0.9668**	0.9418	**0.9544**	**0.9609**	**0.9502**	**0.9685**	0.9495	**0.9586**
	0.33	2	0.9044	0.9351	0.9092	0.9351	**0.9819**	0.8970	0.9368	0.9027	0.9368	**0.9811**	0.8992	0.9317	0.9054	0.9317	**0.9795**
		4	0.9218	0.9429	0.9443	0.9278	**0.9517**	0.9221	0.9408	0.9434	0.9270	0.9484	0.9236	0.9422	0.9440	0.9328	**0.9538**
		6	0.9319	0.9464	**0.9578**	0.9423	**0.9575**	0.9331	0.9467	**0.9574**	0.9457	**0.9605**	0.9340	0.9471	**0.9552**	0.9438	**0.9584**
		8	0.9296	0.9457	**0.9515**	0.9448	**0.9578**	0.9336	0.9451	**0.9555**	0.9438	**0.9570**	0.9359	0.9484	**0.9574**	0.9463	**0.9597**
		10	0.9386	0.9493	**0.9558**	0.9487	**0.9597**	0.9378	0.9486	**0.9561**	0.9483	**0.9579**	0.9364	0.9478	**0.9553**	0.9484	**0.9594**
	0.50	2	0.8554	0.9351	0.8664	0.9351	**0.9779**	0.8562	0.9307	0.8673	0.9307	**0.9807**	0.8565	0.9310	0.8658	0.9310	**0.9772**
		4	0.9001	0.9408	0.9341	0.9299	**0.9545**	0.8969	0.9423	0.9318	0.9313	**0.9528**	0.9024	0.9385	0.9377	0.9306	**0.9534**
		6	0.9146	0.9445	0.9491	0.9412	**0.9563**	0.9129	0.9414	**0.9501**	0.9423	**0.9599**	0.9111	0.9434	0.9496	0.9417	**0.9567**
		8	0.9191	0.9440	0.9483	0.9434	**0.9567**	0.9187	0.9431	**0.9509**	0.9447	**0.9566**	0.9208	0.9446	**0.9519**	0.9458	**0.9590**
		10	0.9242	0.9441	**0.9507**	0.9445	**0.9535**	0.9235	0.9466	**0.9512**	0.9452	**0.9568**	0.9236	0.9457	**0.9518**	0.9463	**0.9590**

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.5. Empirical Coverage Probability of 95% Confidence Intervals for the Risk Difference: Half-half Large and Small Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Half-half	0.00	2	**0.9643**	0.9496	**0.9655**	0.9496	**0.9862**	**0.9627**	0.9486	**0.9635**	0.9486	**0.9854**	**0.9656**	0.9492	**0.9673**	0.9492	**0.9854**
		4	**0.9639**	0.9491	**0.9700**	0.9139	0.9455	**0.9645**	**0.9511**	**0.9696**	0.9139	0.9441	**0.9675**	0.9496	**0.9726**	0.9175	0.9456
		6	**0.9653**	**0.9519**	**0.9722**	0.9228	0.9450	**0.9634**	**0.9523**	**0.9700**	0.9296	**0.9505**	**0.9642**	0.9495	**0.9723**	0.9202	0.9453
		8	**0.9662**	**0.9542**	**0.9732**	0.9293	0.9469	**0.9660**	**0.9524**	**0.9712**	0.9345	0.9496	**0.9632**	**0.9506**	**0.9713**	0.9313	0.9487
		10	**0.9621**	**0.9509**	**0.9712**	0.9334	0.9490	**0.9662**	**0.9545**	**0.9729**	0.9330	0.9490	**0.9656**	**0.9561**	**0.9747**	0.9350	**0.9507**
	0.33	2	0.9025	0.9389	0.9084	0.9389	**0.9840**	0.8937	0.9406	0.8998	0.9406	**0.9828**	0.9035	0.9389	0.9081	0.9389	**0.9822**
		4	0.9110	0.9319	0.9317	0.8983	0.9324	0.9082	0.9279	0.9277	0.8940	0.9292	0.9098	0.9294	0.9278	0.8972	0.9319
		6	0.9198	0.9330	0.9441	0.9184	0.9425	0.9172	0.9335	0.9458	0.9195	0.9433	0.9205	0.9320	0.9442	0.9190	0.9417
		8	0.9186	0.9278	**0.9502**	0.9180	0.9361	0.9193	0.9318	**0.9521**	0.9233	0.9419	0.9197	0.9337	**0.9539**	0.9270	0.9447
		10	0.9212	0.9311	**0.9544**	0.9278	0.9423	0.9239	0.9308	**0.9548**	0.9285	0.9445	0.9196	0.9309	**0.9523**	0.9287	0.9438
	0.50	2	0.8548	0.9308	0.8642	0.9308	**0.9797**	0.8521	0.9341	0.8638	0.9341	**0.9798**	0.8513	0.9308	0.8617	0.9308	**0.9772**
		4	0.8773	0.9176	0.9143	0.8893	0.9262	0.8805	0.9160	0.9141	0.8933	0.9289	0.8796	0.9166	0.9112	0.8938	0.9274
		6	0.8931	0.9221	0.9389	0.9096	0.9362	0.8941	0.9194	0.9378	0.9108	0.9365	0.8904	0.9212	0.9373	0.9143	0.9387
		8	0.9033	0.9265	**0.9502**	0.9244	0.9418	0.9059	0.9289	**0.9521**	0.9277	0.9459	0.9022	0.9269	**0.9517**	0.9230	0.9428
		10	0.9046	0.9292	**0.9562**	0.9292	0.9467	0.9060	0.9274	**0.9544**	0.9270	0.9422	0.9092	0.9307	**0.9585**	0.9330	0.9484

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.6. Empirical Coverage Probability of 95% Confidence Intervals for the Risk Difference: One Large Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One large	0.00	2	**0.9658**	0.9486	**0.9680**	0.9486	**0.9862**	**0.9604**	**0.9503**	**0.9618**	**0.9503**	**0.9842**	**0.9649**	0.9499	**0.9663**	0.9499	**0.9850**
		4	**0.9678**	**0.9533**	**0.9721**	0.9217	**0.9641**	**0.9655**	0.9493	**0.9701**	0.9174	**0.9586**	**0.9655**	**0.9531**	**0.9697**	0.9200	**0.9594**
		6	**0.9649**	0.9487	**0.9707**	0.9194	**0.9530**	**0.9636**	**0.9502**	**0.9713**	0.9209	**0.9535**	**0.9624**	0.9483	**0.9706**	0.9133	0.9485
		8	**0.9658**	**0.9522**	**0.9729**	0.9289	**0.9536**	**0.9603**	0.9468	**0.9696**	0.9199	0.9467	**0.9623**	0.9490	**0.9699**	0.9205	0.9491
		10	**0.9604**	**0.9505**	**0.9685**	0.9270	0.9486	**0.9627**	**0.9526**	**0.9719**	0.9271	0.9483	**0.9597**	0.9486	**0.9693**	0.9253	0.9475
	0.33	2	0.9010	0.9386	0.9072	0.9386	**0.9825**	0.9036	0.9395	0.9094	0.9395	**0.9802**	0.9038	0.9366	0.9098	0.9366	**0.9804**
		4	0.8963	0.9142	0.9131	0.8791	0.9376	0.8932	0.9118	0.9097	0.8771	0.9393	0.8938	0.9116	0.9109	0.8797	0.9417
		6	0.9005	0.9101	0.9235	0.8900	0.9326	0.8981	0.9123	0.9230	0.8918	0.9361	0.9008	0.9112	0.9255	0.8875	0.9301
		8	0.8951	0.9065	0.9263	0.8968	0.9335	0.9079	0.9172	0.9351	0.9051	0.9374	0.9018	0.9100	0.9311	0.8986	0.9332
		10	0.9047	0.9167	0.9360	0.9095	0.9368	0.9099	0.9161	0.9415	0.9065	0.9346	0.9074	0.9140	0.9383	0.9072	0.9335
	0.50	2	0.8571	0.9345	0.8686	0.9345	**0.9801**	0.8562	0.9266	0.8664	0.9266	**0.9792**	0.8536	0.9309	0.8642	0.9309	**0.9792**
		4	0.8644	0.9001	0.8910	0.8744	0.9360	0.8677	0.9042	0.8956	0.8733	0.9392	0.8624	0.8982	0.8926	0.8712	0.9351
		6	0.8751	0.8977	0.9158	0.8845	0.9311	0.8785	0.9023	0.9172	0.8902	0.9348	0.8750	0.9004	0.9155	0.8887	0.9308
		8	0.8904	0.9121	0.9354	0.9034	0.9318	0.8936	0.9131	0.9333	0.9074	0.9371	0.8900	0.9100	0.9337	0.9080	0.9370
		10	0.9011	0.9220	0.9469	0.9212	0.9417	0.9000	0.9209	0.9436	0.9194	0.9423	0.9005	0.9180	0.9417	0.9185	0.9410

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.7. Empirical Coverage Probability of 95% Confidence Intervals for the Log Risk Ratio: Equal Size Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Equal size	0.00	2	**0.9605**	0.9472	**0.9663**	0.9472	**0.9625**	**0.9590**	**0.9535**	**0.9640**	**0.9535**	**0.9675**	**0.9630**	**0.9514**	**0.9656**	**0.9514**	**0.9647**
		4	**0.9643**	**0.9501**	**0.9697**	0.9494	**0.9629**	**0.9619**	**0.9501**	**0.9684**	0.9497	**0.9636**	**0.9619**	0.9485	**0.9675**	0.9494	**0.9635**
		6	**0.9629**	0.9496	**0.9688**	0.9484	**0.9614**	**0.9621**	0.9478	**0.9694**	0.9475	**0.9587**	**0.9640**	**0.9513**	**0.9704**	**0.9513**	**0.9636**
		8	**0.9604**	**0.9504**	**0.9677**	0.9498	**0.9592**	**0.9645**	**0.9512**	**0.9693**	**0.9509**	**0.9622**	**0.9611**	**0.9504**	**0.9676**	0.9496	**0.9613**
		10	**0.9619**	**0.9504**	**0.9677**	0.9494	**0.9578**	**0.9597**	0.9470	**0.9661**	0.9468	**0.9573**	**0.9623**	0.9500	**0.9681**	0.9497	**0.9585**
	0.33	2	0.9258	0.9490	0.9384	0.9490	**0.9631**	0.9238	**0.9528**	0.9355	**0.9528**	**0.9665**	0.9286	0.9494	0.9388	0.9494	**0.9633**
		4	0.9334	0.9499	**0.9529**	0.9495	**0.9660**	0.9313	**0.9505**	**0.9509**	0.9500	**0.9643**	0.9333	**0.9543**	**0.9543**	**0.9535**	**0.9682**
		6	0.9401	**0.9521**	**0.9593**	**0.9512**	**0.9639**	0.9365	**0.9504**	**0.9558**	**0.9501**	**0.9633**	0.9348	0.9495	**0.9549**	0.9496	**0.9629**
		8	0.9401	**0.9539**	**0.9562**	**0.9529**	**0.9648**	0.9348	0.9498	**0.9553**	**0.9503**	**0.9599**	0.9359	0.9475	**0.9542**	0.9472	**0.9593**
		10	0.9348	0.9462	**0.9508**	0.9452	**0.9556**	0.9385	0.9493	**0.9533**	0.9490	**0.9587**	0.9383	0.9478	**0.9534**	0.9475	**0.9568**
	0.50	2	0.8947	**0.9537**	0.9120	**0.9537**	**0.9665**	0.8957	**0.9511**	0.9183	**0.9511**	**0.9650**	0.8958	**0.9528**	0.9133	**0.9528**	**0.9674**
		4	0.9125	0.9487	0.9436	0.9485	**0.9636**	0.9113	0.9493	0.9441	0.9490	**0.9644**	0.9153	**0.9533**	0.9477	**0.9535**	**0.9675**
		6	0.9202	0.9487	0.9480	0.9476	**0.9601**	0.9204	**0.9515**	**0.9508**	**0.9509**	**0.9625**	0.9194	0.9490	**0.9503**	0.9492	**0.9631**
		8	0.9314	**0.9574**	**0.9583**	**0.9569**	**0.9654**	0.9298	**0.9564**	**0.9561**	**0.9570**	**0.9655**	0.9255	**0.9522**	**0.9526**	**0.9522**	**0.9625**
		10	0.9270	0.9491	0.9490	0.9486	**0.9578**	0.9295	**0.9528**	**0.9539**	**0.9528**	**0.9614**	0.9296	**0.9529**	**0.9541**	**0.9532**	**0.9605**

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.8. Empirical Coverage Probability of 95% Confidence Intervals for the Log Risk Ratio: One Small Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One small	0.00	2	**0.9666**	**0.9512**	**0.9676**	**0.9512**	**0.9833**	**0.9615**	0.9498	**0.9628**	0.9498	**0.9846**	**0.9662**	0.9484	**0.9681**	0.9484	**0.9824**
		4	**0.9634**	0.9445	**0.9680**	0.9260	0.9490	**0.9684**	**0.9532**	**0.9737**	0.9350	**0.9570**	**0.9674**	**0.9523**	**0.9728**	0.9336	**0.9530**
		6	**0.9630**	**0.9507**	**0.9699**	0.9426	**0.9567**	**0.9605**	**0.9510**	**0.9692**	0.9432	**0.9572**	**0.9599**	0.9465	**0.9666**	0.9410	**0.9552**
		8	**0.9629**	0.9499	**0.9702**	0.9438	**0.9575**	**0.9604**	**0.9517**	**0.9663**	0.9466	**0.9572**	**0.9637**	**0.9506**	**0.9700**	0.9472	**0.9597**
		10	**0.9656**	**0.9540**	**0.9718**	**0.9528**	**0.9624**	**0.9597**	0.9491	**0.9665**	0.9442	**0.9561**	**0.9612**	0.9493	**0.9684**	0.9468	**0.9568**
	0.33	2	0.9018	0.9361	0.9059	0.9361	**0.9810**	0.9035	0.9375	0.9093	0.9375	**0.9815**	0.8995	0.9397	0.9056	0.9397	**0.9817**
		4	0.9228	0.9424	0.9440	0.9273	0.9494	0.9246	0.9413	0.9447	0.9273	**0.9501**	0.9266	0.9433	0.9449	0.9314	**0.9531**
		6	0.9310	0.9467	**0.9526**	0.9431	**0.9614**	0.9367	0.9487	**0.9598**	0.9447	**0.9590**	0.9305	0.9431	**0.9532**	0.9409	**0.9557**
		8	0.9361	0.9471	**0.9562**	0.9458	**0.9589**	0.9360	0.9474	**0.9566**	0.9465	**0.9583**	0.9340	0.9469	**0.9585**	0.9461	**0.9593**
		10	0.9364	0.9482	**0.9544**	0.9477	**0.9570**	0.9357	0.9484	**0.9536**	0.9475	**0.9571**	0.9372	0.9491	**0.9563**	0.9496	**0.9593**
	0.50	2	0.8586	0.9329	0.8695	0.9329	**0.9783**	0.8530	0.9374	0.8635	0.9374	**0.9797**	0.8629	0.9353	0.8745	0.9353	**0.9809**
		4	0.9022	0.9371	0.9354	0.9264	0.9498	0.8965	0.9420	0.9335	0.9309	**0.9534**	0.9004	0.9389	0.9353	0.9283	**0.9505**
		6	0.9145	0.9413	0.9496	0.9387	**0.9545**	0.9099	0.9440	**0.9508**	0.9418	**0.9583**	0.9157	0.9415	**0.9501**	0.9407	**0.9560**
		8	0.9212	0.9463	**0.9524**	0.9472	**0.9585**	0.9173	0.9441	**0.9503**	0.9445	**0.9566**	0.9234	0.9480	**0.9537**	0.9480	**0.9586**
		10	0.9233	0.9494	**0.9536**	0.9495	**0.9603**	0.9253	0.9493	**0.9520**	0.9483	**0.9578**	0.9230	0.9433	**0.9503**	0.9445	**0.9537**

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.9. Empirical Coverage Probability of 95% Confidence Intervals for the Log Risk Ratio: Half-half Large and Small Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Half-half	0.00	2	**0.9618**	0.9490	**0.9633**	0.9490	**0.9826**	**0.9671**	**0.9502**	**0.9682**	**0.9502**	**0.9851**	**0.9634**	0.9493	**0.9649**	0.9493	**0.9836**
		4	**0.9657**	**0.9512**	**0.9698**	0.9130	0.9456	**0.9651**	0.9482	**0.9689**	0.9121	0.9449	**0.9647**	0.9484	**0.9688**	0.9124	0.9430
		6	**0.9659**	**0.9534**	**0.9707**	0.9230	0.9452	**0.9643**	**0.9519**	**0.9707**	0.9215	0.9466	**0.9635**	0.9491	**0.9697**	0.9194	0.9421
		8	**0.9629**	0.9491	**0.9695**	0.9268	0.9459	**0.9625**	**0.9509**	**0.9693**	0.9296	0.9460	**0.9622**	**0.9501**	**0.9711**	0.9238	0.9441
		10	**0.9640**	**0.9519**	**0.9716**	0.9347	0.9492	**0.9655**	**0.9534**	**0.9721**	0.9342	0.9489	**0.9628**	**0.9515**	**0.9701**	0.9304	0.9474
	0.33	2	0.9079	0.9413	0.9127	0.9413	**0.9807**	0.8992	0.9397	0.9050	0.9397	**0.9802**	0.9012	0.9398	0.9078	0.9398	**0.9808**
		4	0.9151	0.9297	0.9311	0.9009	0.9300	0.9118	0.9293	0.9282	0.8982	0.9300	0.9082	0.9287	0.9267	0.9007	0.9347
		6	0.9205	0.9353	0.9462	0.9136	0.9390	0.9175	0.9298	0.9454	0.9144	0.9390	0.9158	0.9319	0.9425	0.9164	0.9415
		8	0.9202	0.9305	0.9490	0.9254	0.9440	0.9231	0.9354	**0.9560**	0.9259	0.9473	0.9255	0.9360	**0.9537**	0.9273	0.9471
		10	0.9206	0.9294	**0.9511**	0.9251	0.9409	0.9233	0.9352	**0.9553**	0.9284	0.9449	0.9158	0.9287	**0.9508**	0.9252	0.9417
	0.50	2	0.8579	0.9316	0.8677	0.9316	**0.9791**	0.8558	0.9298	0.8652	0.9298	**0.9778**	0.8565	0.9326	0.8653	0.9326	**0.9771**
		4	0.8803	0.9217	0.9099	0.8909	0.9257	0.8874	0.9199	0.9178	0.8966	0.9285	0.8811	0.9195	0.9153	0.8951	0.9288
		6	0.8988	0.9251	0.9411	0.9157	0.9389	0.8983	0.9235	0.9392	0.9158	0.9372	0.8910	0.9204	0.9362	0.9103	0.9334
		8	0.8981	0.9235	0.9471	0.9214	0.9407	0.9056	0.9298	**0.9540**	0.9246	0.9427	0.8978	0.9207	0.9493	0.9197	0.9394
		10	0.9101	0.9316	**0.9585**	0.9331	0.9482	0.9070	0.9295	**0.9561**	0.9286	0.9448	0.9090	0.9303	**0.9558**	0.9291	0.9451

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.10. Empirical Coverage Probability of 95% Confidence Intervals for the Log Risk Ratio: One Large Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One large	0.00	2	**0.9643**	0.9484	**0.9657**	0.9484	**0.9859**	**0.9664**	**0.9525**	**0.9680**	**0.9525**	**0.9834**	**0.9684**	**0.9558**	**0.9701**	**0.9558**	**0.9873**
		4	**0.9655**	**0.9556**	**0.9700**	0.9169	**0.9598**	**0.9657**	0.9493	**0.9698**	0.9157	**0.9601**	**0.9656**	0.9473	**0.9692**	0.9133	**0.9581**
		6	**0.9633**	0.9462	**0.9693**	0.9128	0.9494	**0.9635**	0.9456	**0.9692**	0.9129	0.9483	**0.9668**	**0.9529**	**0.9726**	0.9199	**0.9542**
		8	**0.9615**	0.9486	**0.9692**	0.9237	0.9493	**0.9658**	**0.9532**	**0.9732**	0.9285	**0.9518**	**0.9627**	0.9481	**0.9710**	0.9254	**0.9502**
		10	**0.9626**	**0.9505**	**0.9710**	0.9236	0.9484	**0.9631**	0.9496	**0.9713**	0.9253	0.9483	**0.9605**	0.9490	**0.9712**	0.9281	0.9489
	0.33	2	0.8998	0.9382	0.9059	0.9382	**0.9774**	0.9036	0.9370	0.9089	0.9370	**0.9795**	0.9022	0.9380	0.9078	0.9380	**0.9775**
		4	0.8974	0.9157	0.9149	0.8836	0.9421	0.9003	0.9214	0.9169	0.8867	0.9417	0.8962	0.9169	0.9147	0.8833	0.9403
		6	0.8994	0.9128	0.9226	0.8897	0.9347	0.8951	0.9045	0.9174	0.8827	0.9300	0.8985	0.9101	0.9217	0.8897	0.9338
		8	0.9025	0.9114	0.9324	0.8992	0.9356	0.9039	0.9136	0.9312	0.8965	0.9327	0.9046	0.9137	0.9334	0.9002	0.9334
		10	0.9124	0.9205	0.9420	0.9113	0.9387	0.9090	0.9150	0.9384	0.9103	0.9374	0.9078	0.9212	0.9406	0.9110	0.9376
	0.50	2	0.8568	0.9308	0.8671	0.9308	**0.9787**	0.8606	0.9306	0.8707	0.9306	**0.9777**	0.8545	0.9305	0.8664	0.9305	**0.9791**
		4	0.8691	0.9047	0.9006	0.8816	0.9418	0.8656	0.8994	0.8937	0.8737	0.9366	0.8658	0.9061	0.8950	0.8788	0.9424
		6	0.8756	0.9013	0.9158	0.8872	0.9295	0.8808	0.9083	0.9220	0.8919	0.9345	0.8799	0.9035	0.9188	0.8888	0.9341
		8	0.8873	0.9062	0.9316	0.9002	0.9345	0.8888	0.9118	0.9334	0.9053	0.9366	0.8881	0.9099	0.9311	0.9016	0.9345
		10	0.8993	0.9170	0.9396	0.9151	0.9381	0.9056	0.9227	0.9435	0.9186	0.9417	0.9032	0.9200	0.9447	0.9175	0.9411

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.11. Empirical Coverage Probability of 95% Confidence Intervals for the Log Odds Ratio: Equal Size Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Equal size	0.00	2	**0.9627**	0.9468	**0.9665**	0.9468	**0.9601**	**0.9621**	0.9482	**0.9661**	0.9482	**0.9645**	**0.9630**	0.9489	**0.9668**	0.9489	**0.9641**
		4	**0.9650**	**0.9510**	**0.9711**	**0.9506**	**0.9649**	**0.9629**	0.9474	**0.9696**	0.9474	**0.9635**	**0.9646**	0.9493	**0.9702**	0.9492	**0.9655**
		6	**0.9627**	**0.9515**	**0.9699**	**0.9514**	**0.9622**	**0.9622**	0.9496	**0.9697**	0.9496	**0.9628**	**0.9622**	0.9479	**0.9688**	0.9478	**0.9621**
		8	**0.9613**	**0.9523**	**0.9679**	**0.9510**	**0.9615**	**0.9643**	**0.9505**	**0.9695**	**0.9505**	**0.9614**	**0.9591**	0.9486	**0.9664**	0.9486	**0.9604**
		10	**0.9621**	0.9482	**0.9669**	0.9478	**0.9577**	**0.9581**	0.9467	**0.9644**	0.9466	**0.9570**	**0.9596**	0.9481	**0.9661**	0.9479	**0.9566**
	0.33	2	0.9245	0.9492	0.9353	0.9492	**0.9637**	0.9246	**0.9505**	0.9365	**0.9505**	**0.9644**	0.9260	**0.9505**	0.9376	**0.9505**	**0.9653**
		4	0.9335	0.9497	**0.9535**	0.9491	**0.9655**	0.9295	0.9495	**0.9514**	0.9495	**0.9651**	0.9324	**0.9519**	**0.9528**	**0.9516**	**0.9649**
		6	0.9351	0.9491	**0.9540**	0.9488	**0.9621**	0.9353	**0.9528**	**0.9564**	**0.9529**	**0.9655**	0.9424	**0.9544**	**0.9616**	**0.9541**	**0.9663**
		8	0.9346	0.9480	**0.9532**	0.9469	**0.9595**	0.9387	0.9493	**0.9559**	0.9493	**0.9603**	0.9383	**0.9512**	**0.9545**	**0.9505**	**0.9619**
		10	0.9371	0.9494	**0.9535**	0.9485	**0.9574**	0.9388	0.9491	**0.9541**	0.9491	**0.9580**	0.9387	**0.9538**	**0.9560**	**0.9541**	**0.9632**
	0.50	2	0.8904	0.9477	0.9119	0.9477	**0.9623**	0.8919	0.9468	0.9142	0.9468	**0.9609**	0.8926	**0.9506**	0.9121	**0.9506**	**0.9655**
		4	0.9104	0.9493	0.9428	0.9490	**0.9647**	0.9131	0.9495	0.9434	0.9495	**0.9651**	0.9128	**0.9535**	0.9473	**0.9533**	**0.9681**
		6	0.9210	0.9495	**0.9503**	0.9494	**0.9624**	0.9189	**0.9511**	0.9494	**0.9511**	**0.9649**	0.9204	0.9473	0.9474	0.9470	**0.9589**
		8	0.9261	**0.9544**	**0.9537**	**0.9538**	**0.9630**	0.9228	0.9498	**0.9516**	0.9498	**0.9617**	0.9210	0.9479	0.9497	0.9480	**0.9587**
		10	0.9250	0.9467	0.9481	0.9469	**0.9568**	0.9261	**0.9519**	**0.9526**	**0.9519**	**0.9627**	0.9288	**0.9530**	**0.9536**	**0.9532**	**0.9616**

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.12. Empirical Coverage Probability of 95% Confidence Intervals for the Log Odds Ratio: One Small Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One small	0.00	2	**0.9669**	**0.9537**	**0.9681**	**0.9537**	**0.9844**	**0.9684**	**0.9507**	**0.9693**	**0.9507**	**0.9867**	**0.9638**	**0.9517**	**0.9649**	**0.9517**	**0.9856**
		4	**0.9635**	0.9465	**0.9689**	0.9269	0.9487	**0.9621**	0.9475	**0.9680**	0.9286	**0.9511**	**0.9656**	**0.9518**	**0.9716**	0.9347	**0.9568**
		6	**0.9612**	0.9476	**0.9691**	0.9410	**0.9567**	**0.9591**	0.9490	**0.9668**	0.9429	**0.9584**	**0.9628**	**0.9507**	**0.9713**	0.9423	**0.9583**
		8	**0.9609**	0.9475	**0.9671**	0.9448	**0.9598**	**0.9586**	**0.9515**	**0.9669**	0.9457	**0.9573**	**0.9625**	0.9500	**0.9697**	0.9449	**0.9550**
		10	**0.9613**	0.9482	**0.9677**	0.9467	**0.9569**	**0.9623**	**0.9510**	**0.9685**	0.9481	**0.9576**	**0.9643**	**0.9504**	**0.9705**	0.9480	**0.9580**
	0.33	2	0.8988	0.9406	0.9030	0.9406	**0.9820**	0.9043	0.9347	0.9100	0.9347	**0.9814**	0.9004	0.9359	0.9056	0.9359	**0.9791**
		4	0.9291	0.9479	**0.9512**	0.9336	**0.9535**	0.9259	0.9450	0.9474	0.9321	**0.9523**	0.9199	0.9426	0.9420	0.9277	0.9486
		6	0.9321	0.9438	**0.9550**	0.9398	**0.9561**	0.9305	0.9430	**0.9535**	0.9371	**0.9557**	0.9312	0.9453	**0.9563**	0.9420	**0.9572**
		8	0.9371	0.9457	**0.9561**	0.9456	**0.9560**	0.9297	0.9416	**0.9516**	0.9414	**0.9544**	0.9319	0.9456	**0.9538**	0.9445	**0.9567**
		10	0.9350	0.9470	**0.9537**	0.9461	**0.9572**	0.9373	**0.9502**	**0.9553**	0.9485	**0.9570**	0.9335	0.9448	**0.9506**	0.9444	**0.9541**
	0.50	2	0.8574	0.9338	0.8672	0.9338	**0.9790**	0.8555	0.9351	0.8664	0.9351	**0.9812**	0.8554	0.9321	0.8641	0.9321	**0.9794**
		4	0.9041	0.9428	0.9348	0.9309	**0.9536**	0.8931	0.9381	0.9314	0.9290	**0.9506**	0.9048	0.9399	0.9383	0.9315	**0.9524**
		6	0.9162	0.9437	**0.9517**	0.9414	**0.9559**	0.9130	0.9425	**0.9515**	0.9426	**0.9586**	0.9145	0.9421	**0.9514**	0.9397	**0.9537**
		8	0.9189	0.9452	**0.9518**	0.9444	**0.9557**	0.9204	0.9445	**0.9519**	0.9440	**0.9560**	0.9165	0.9424	0.9492	0.9432	**0.9546**
		10	0.9296	0.9493	**0.9545**	0.9493	**0.9586**	0.9257	0.9487	**0.9533**	0.9499	**0.9601**	0.9235	0.9499	**0.9548**	0.9493	**0.9594**

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.13. Empirical Coverage Probability of 95% Confidence Intervals for the Log Odds Ratio: Half-half Large and Small Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Half-half	0.00	2	**0.9641**	0.9498	**0.9653**	0.9498	**0.9835**	**0.9688**	0.9482	**0.9702**	0.9482	**0.9848**	**0.9630**	0.9481	**0.9638**	0.9481	**0.9830**
		4	**0.9656**	0.9487	**0.9692**	0.9133	0.9419	**0.9650**	**0.9516**	**0.9698**	0.9148	0.9447	**0.9668**	0.9494	**0.9703**	0.9146	0.9439
		6	**0.9636**	**0.9524**	**0.9692**	0.9247	0.9452	**0.9599**	0.9482	**0.9677**	0.9209	0.9439	**0.9657**	**0.9502**	**0.9711**	0.9237	0.9460
		8	**0.9626**	0.9457	**0.9691**	0.9250	0.9437	**0.9644**	**0.9504**	**0.9717**	0.9260	0.9455	**0.9664**	**0.9515**	**0.9724**	0.9261	0.9454
		10	**0.9658**	**0.9516**	**0.9731**	0.9305	0.9452	**0.9584**	0.9471	**0.9672**	0.9333	0.9480	**0.9661**	**0.9528**	**0.9732**	0.9347	0.9484
	0.33	2	0.9004	0.9440	0.9042	0.9440	**0.9816**	0.9023	0.9395	0.9067	0.9395	**0.9828**	0.8982	0.9367	0.9037	0.9367	**0.9798**
		4	0.9197	0.9316	0.9357	0.8962	0.9306	0.9144	0.9340	0.9320	0.9015	0.9362	0.9167	0.9330	0.9351	0.8941	0.9307
		6	0.9229	0.9348	0.9468	0.9179	0.9401	0.9188	0.9293	0.9438	0.9175	0.9406	0.9179	0.9305	0.9442	0.9138	0.9375
		8	0.9204	0.9284	0.9491	0.9225	0.9411	0.9262	0.9348	**0.9537**	0.9279	0.9459	0.9190	0.9293	0.9480	0.9200	0.9396
		10	0.9209	0.9313	**0.9527**	0.9266	0.9427	0.9276	0.9378	**0.9565**	0.9338	0.9494	0.9276	0.9390	**0.9563**	0.9342	0.9477
	0.50	2	0.8621	0.9365	0.8729	0.9365	**0.9823**	0.8586	0.9342	0.8686	0.9342	**0.9793**	0.8590	0.9275	0.8688	0.9275	**0.9760**
		4	0.8769	0.9167	0.9098	0.8926	0.9249	0.8798	0.9219	0.9124	0.8970	0.9317	0.8820	0.9230	0.9170	0.8949	0.9271
		6	0.8960	0.9213	0.9371	0.9141	0.9360	0.8921	0.9187	0.9376	0.9100	0.9360	0.8938	0.9245	0.9407	0.9145	0.9386
		8	0.9006	0.9235	0.9472	0.9213	0.9409	0.8974	0.9240	0.9467	0.9192	0.9386	0.9013	0.9260	**0.9504**	0.9229	0.9420
		10	0.9099	0.9309	**0.9548**	0.9304	0.9476	0.9077	0.9290	**0.9549**	0.9286	0.9440	0.9112	0.9334	**0.9591**	0.9313	0.9469

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table 3.14. Empirical Coverage Probability of 95% Confidence Intervals for the Log Odds Ratio: One Large Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One large	0.00	2	**0.9632**	0.9499	**0.9641**	0.9499	**0.9853**	**0.9677**	0.9483	**0.9687**	0.9483	**0.9856**	**0.9628**	**0.9516**	**0.9640**	**0.9516**	**0.9833**
		4	**0.9665**	**0.9502**	**0.9706**	0.9122	**0.9573**	**0.9671**	**0.9502**	**0.9708**	0.9155	**0.9578**	**0.9681**	**0.9514**	**0.9710**	0.9166	**0.9591**
		6	**0.9618**	0.9484	**0.9675**	0.9188	**0.9510**	**0.9687**	**0.9539**	**0.9751**	0.9271	**0.9559**	**0.9639**	0.9476	**0.9704**	0.9181	0.9490
		8	**0.9620**	**0.9514**	**0.9703**	0.9240	0.9480	**0.9643**	**0.9506**	**0.9719**	0.9239	0.9496	**0.9632**	0.9494	**0.9711**	0.9248	**0.9501**
		10	**0.9636**	**0.9517**	**0.9717**	0.9289	0.9491	**0.9638**	**0.9509**	**0.9720**	0.9323	**0.9523**	**0.9642**	0.9489	**0.9743**	0.9299	**0.9506**
	0.33	2	0.9054	0.9389	0.9113	0.9389	**0.9830**	0.9092	0.9386	0.9135	0.9386	**0.9821**	0.9004	0.9412	0.9058	0.9412	**0.9824**
		4	0.8969	0.9136	0.9156	0.8836	0.9427	0.9006	0.9164	0.9158	0.8819	0.9424	0.8959	0.9165	0.9149	0.8832	0.9415
		6	0.9017	0.9106	0.9246	0.8883	0.9330	0.9008	0.9069	0.9225	0.8888	0.9353	0.8963	0.9074	0.9221	0.8866	0.9326
		8	0.9049	0.9135	0.9335	0.9026	0.9340	0.9027	0.9152	0.9343	0.9016	0.9381	0.9058	0.9138	0.9362	0.9013	0.9337
		10	0.9035	0.9124	0.9363	0.9056	0.9335	0.9063	0.9144	0.9388	0.9120	0.9391	0.9085	0.9207	0.9391	0.9102	0.9364
	0.50	2	0.8575	0.9294	0.8676	0.9294	**0.9771**	0.8545	0.9323	0.8659	0.9323	**0.9780**	0.8571	0.9344	0.8678	0.9344	**0.9786**
		4	0.8598	0.8971	0.8904	0.8690	0.9365	0.8639	0.9013	0.8924	0.8759	0.9387	0.8588	0.8946	0.8863	0.8723	0.9360
		6	0.8749	0.9007	0.9132	0.8857	0.9323	0.8756	0.9009	0.9149	0.8859	0.9293	0.8828	0.9071	0.9222	0.8951	0.9378
		8	0.8919	0.9106	0.9332	0.9057	0.9387	0.8929	0.9172	0.9353	0.9089	0.9417	0.8870	0.9081	0.9324	0.9019	0.9354
		10	0.9043	0.9202	0.9435	0.9141	0.9388	0.8977	0.9188	0.9439	0.9156	0.9378	0.9007	0.9179	0.9441	0.9156	0.9393

Note:
Confidence intervals were computed using the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods used to compute confidence intervals. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

CHAPTER 4

Considerations in heterogeneity variance estimation in small meta-analysis

4.1. Introduction

We consider the random-effects linear model in meta-analysis

$$y_i = \delta + \xi_i + \epsilon_i, \tag{4.1.1}$$

where ξ_i and ϵ_i are normal random variables with expectations 0 and variances τ^2 and σ_i^2, respectively, and $i = 1, ..., k$. This model is widely used in education, the social sciences and medical sciences for combining results from experiments. The parameters of interest are: the summary effect, δ, heterogeneity variance, τ^2, and sampling errors σ_i^2. Our interest in this article lies in unbiased and efficient estimation of the heterogeneity variance parameter, τ^2. We provide some evidence about the merits of five heterogeneity variance estimators with a particular focus on unbalanced data and small numbers of studies. By unbalance, we mean data that arises from studies containing different numbers of observations per study, and by small we mean $k < 10$.

A large number of heterogeneity variance estimators have been proposed in the literature. Some of the earlier proposed methods include an analogue to the random-effects analysis of variance, based on equal weights (Cochran, 1954; Hedges, 1985), and a precision weighted estimator, weighted by the inverse of the within-study variances (DerSimonian and Laird, 1986). Both estimators are results of one-step methods, thus are convenient in practice. Alternatives include maximum likelihood and restricted maximum likelihood (Rebecca J Hardy and Thompson, 1996; RJ Hardy and Thompson, 1996; Harville,

1977), and others which have been previously discussed in the literature (Langan *et al.*, 2019; Schmidt and Hunter, 2015; Sidik and Jonkman, 2005).

Our motivation for this work lies in a practical problem faced by many research clearinghouses. Research clearinghouses perform systematic reviews of randomized trials to provide evidence about treatment effectiveness to policy-makers and practitioners alike. When the number of studies is at least two, the findings from individual studies are typically combined using meta-analysis methods. In meta-analyses performed by research clearinghouses, however, the number of studies eligible for synthesis is quite small. For example, in the What Works Clearinghouse database, the median number of studies is two. In the Cochrane Library, the median number of studies was found to be three and in approximately 75% of the meta-analyses, the number of studies was five or less (Davey *et al.*, 2011). This presents challenges in using a random-effects model due to the difficulty in obtaining an unbiased and efficient estimate of the heterogeneity parameter.

In this article, we consider five heterogeneity variance estimators for investigation: two method of moments estimators (ANOVA and DL) and three estimators obtained by maximum likelihood methods (ML, REML and AREML). AREML motivates a new estimator that closeley resembles REML. Performance is assessed by examining bias and variance of the estimators for variations of the heterogeneity parameter, number of studies and within-study sample sizes. This work focuses on situations where the effect size of interest is the Standardized Mean Difference. Our goal is to provide useful guidelines in applying these estimators to small meta-analyses that reflect the aforementioned practical settings.

The chapter is structured as follows. Sections 4.2 and 4.3 describe the procedures for obtaining an estimate of the between-study variance, grouped into method of moments and maximum likelihood procedures. Section 4.4 investigates the analytic properties of the estimators, in cases where it is possible to do so. Section 4.5 discusses a small simulation study that complement the analytic results. Section 4.6 comments on the merits of each estimator. Finally, Section 4.7 summarizes our findings and concludes.

4.2. Heterogeneity variance estimators: Method of Moments

In this section and the following, we describe the five main estimators considered in this chapter, namely the ANOVA estimator, DL, ML, REML and AREML. While many more heterogeneity variance estimators have been proposed, we focus on these ones in particular because our ultimate goal is to understand how errors in estimating the between-study variance parameter propagate to the mean of the distribution of effects and its standard error, hence we find it unnecessary to be exhaustive. Next we discuss the method of moments, followed by maximum likelihood methods.

The estimators for the heterogeneity variance proposed by both Hedges and Olkin and DerSimonian Laird use the method of moments (MM) and can be obtained as a result of a unifying identity (Kacker, 2004). Denote $\bar{d}_+ = \sum_{i=1}^k a_i d_i / \sum_{i=1}^k a_i$, where a_i are positive constants. The Method of Moments identiy is given by

$$E\left[\sum_{i=1}^{k} a_i (d_i - \bar{d}_+)^2\right] = \left(\sum_{i=1}^{k} a_i \sigma_i^2 - \frac{\sum_{i=1}^k a_i^2 \sigma_i^2}{\sum_{i=1}^k a_i}\right) + \tau^2 \left(\sum_{i=1}^{k} a_i - \frac{\sum_{i=1}^k a_i^2}{\sum_{i=1}^k a_i}\right). \tag{4.2.1}$$

Equating $\sum_{i=1}^k a_i (d_i - \bar{d}_+)^2$ to its expected value yields the general method of moments (MM) estimator of the between-study variance,

$$\hat{\tau}^2_{MM} = \left[\sum_{i=1}^{k} a_i (d_i - \bar{d}_+)^2 - \left(\sum_{i=1}^{k} a_i \sigma_i^2 - \frac{\sum_{i=1}^k a_i^2 \sigma_i^2}{\sum_{i=1}^k a_i}\right)\right] \left(\sum_{i=1}^{k} a_i - \frac{\sum_{i=1}^k a_i^2}{\sum_{i=1}^k a_i}\right)^{-1} \tag{4.2.2}$$

where the $a_i's$ are positive constants. Method of moments estimates can then be found using different values for the constants. Two of these methods are summarized in the next two subsections: ANOVA and the DerSimonian and Laird estimator.

Method of moments estimators are particularly attractive in empirical research because they are simple and can be obtained in one-step. An additional advantage of their use is that they do not assume a distribution of the δ_i's, contrary to iterative methods such as maximum likelihood and restricted maximum likelihood.

4.2.1. ANOVA

We can obtain a between-study variance estimator by substituting $a_i = 1/(k-1)$ in equation (4.2.2). This estimator has first been introduced in the random-effects meta-analytic framework by Hedges (1985) and initially proposed by Cochran (1954). The estimator is

$$\hat{\tau}^2_H = \sum_{i=1}^{k} \frac{(d_i - \bar{d})^2}{k-1} - \frac{1}{k}\sum_{i=1}^{k} \sigma^2_{d_i}, \tag{4.2.3}$$

where $\bar{d}$ is the simple arithmetic average of the individual study effect estimates. If the estimator produces a negative value, it is set to 0. This estimator is exactly unbiased before truncation.

4.2.2. DerSimonian and Laird Method of Moments

By substituting $a_i = 1/\sigma_i^2$ in equation (4.2.2), we obtain the DerSimonian and Laird estimator. Its functional form is given by

$$\hat{\tau}^2_{DL} = \begin{cases} \dfrac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}} & Q > k-1 \\ 0 & Q \leq k-1 \end{cases} \tag{4.2.4}$$

where $Q = \sum_{i=1}^{k} w_i(d_i - \bar{d}_+)^2$, $\bar{d}_+$ is the weighted mean, and the weights assigned are proportional to the inverse of within-study variances, i.e. $a_i = w_i = 1/\sigma_i^2$. Note that the DL method can also produce a negative estimate for the heterogeneity variance, in which case the estimator is truncated at 0. When the weights are equal, so that $w_i = w$ for all $i = 1, ..., k$, the DL method reduces to the ANOVA method, and both produce identical estimates. The DL estimator is also exactly unbiased before truncation.

4.3. Heterogeneity variance estimators: Maximum likelihood methods

Maximum likelihood methods include the maximum likelihood estimator (ML), restricted maximum likelihood (REML) and approximate restricted maximum likelihood (AREML). As a preview, although REML and AREML are nearly identical, we include a discussion of AREML since it proves to be easier to work with for analytic comparisons to the other alternatives. Furthermore, we introduce another

estimator that makes a direct adjustment for the degrees of freedom which is closely related to REML and is motivated by AREML. The following estimators involve maximizing the log likelihood function and are iterative in nature. Next, we describe the details of these estimators.

4.3.1. Maximum likelihood

Recall that $d_i|\tau^2 \sim AN(\mu, \tau^2 + \sigma_i^2)$. Then the log likelihood function of the random-effects model is given by

$$logL(\delta, \tau^2) = -\frac{k}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{k} log(\tau^2 + \sigma_i^2) - \frac{1}{2}\sum_{i=1}^{k} \frac{(d_i - \delta)^2}{\tau^2 + \sigma_i^2}. \tag{4.3.1}$$

The maximum likelihood estimators for the summary effect and heterogeneity variance are found by maximizing the log likelihood function with respect to δ and τ^2, respectively. The estimators take the following form

$$\hat{\delta}_{ML} = \frac{\sum_{i=1}^{k} \hat{w}_i d_i}{\sum_{i=1}^{k} \hat{w}_i} \tag{4.3.2}$$

and

$$\hat{\tau}^2_{ML} = \frac{\sum_{i=1}^{k} \hat{w}_i^2[(d_i - \hat{\delta}_{ML})^2 - \sigma_i^2]}{\sum_{i=1}^{k} \hat{w}_i{}^2}, \tag{4.3.3}$$

where $\hat{w}_i = 1/(\hat{\tau}^2_{ML} + \sigma_i^2)$.

Values for $\hat{\delta}_{ML}$ and $\hat{\tau}^2_{ML}$ are found by iterating through expressions (4.3.2) and (4.3.3), with an initial estimate of $\hat{\tau}^2_{ML}$. The initial estimate can be set to 0, or alternatively one can use the ANOVA estimate or the one obtained from the DerSimonian and Laird method. If a negative value is obtained in an iterative step, the estimator is set equal to 0.

One drawback of the maximum likelihood method is that we assume δ is known. This assumption leads to bias in the estimate of the between-study variance parameter. The problem can be remedied by using the restricted maximum likelihood estimator, which we introduce next.

4.3.2. Restricted Maximum Likehood

The restricted maximum likelihood approach involves separating the log likelihood function into two log likelihoods, one that contains only the summary effect, δ, and another that involves only the heterogeneity variance parameter, τ^2. The reduced log likelihood is given by

$$logL(\delta, \tau^2) = -\frac{k}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{k} log(\tau^2 + \sigma_i^2) - \frac{1}{2}log\sum_{i=1}^{k}\frac{1}{\tau^2 + \sigma_i^2} - \frac{1}{2}\sum_{i=1}^{k}\frac{(d_i - \delta)^2}{\tau^2 + \sigma_i^2}. \tag{4.3.4}$$

We then maximize the log likelihood function with respect to τ^2 and find that

$$\hat{\tau}^2_{REML} = \frac{\sum_{i=1}^{k} \hat{w}_i^2[(d_i - \hat{\delta}_{REML})^2 + 1/\sum_{i=1}^{k} \hat{w}_i - \sigma_i^2]}{\sum_{i=1}^{k} \hat{w}_i{}^2}, \tag{4.3.5}$$

where $\hat{\delta}_{REML} = \frac{\sum_{i=1}^{k} \hat{w}_i d_i}{\sum_{i=1}^{k} \hat{w}_i}$, $\hat{\delta}_{REML} = \frac{\sum_{i=1}^{k} \hat{w}_i d_i}{\sum_{i=1}^{k} \hat{w}_i}$ and the weights assigned are $\hat{w}_i = 1/(\hat{\tau}^2_{REML} + \sigma_i^2)$. A solution is then obtained by iterating through the above expressions until convergence. As in the maximum likelihood approach, REML can also produce negative values in an iteration step, in which case it is set equal to 0.

4.3.3. Approximate Restricted Maximum Likelihood

One closely related estimator to the restricted maximum likelihood estimator is the approximate restricted maximum likelihood (AREML) (Morris, 1983; Thompson and Sharp, 1999; Veroniki *et al.*, 2016). The AREML estimator directly adjusts for the degrees of freedom (Veroniki *et al.*, 2016). The estimator is given by

$$\hat{\tau}^2_{AREML} = \frac{\sum_{i=1}^{k} \hat{w}_i^2(\frac{k}{k-1}(d_i - \hat{\delta}_{AREML})^2 - \sigma_i^2)}{\sum_{i=1}^{k} \hat{w}_i{}^2}, \tag{4.3.6}$$

where $\hat{\delta}_{AREML} = \frac{\sum_{i=1}^{k} \hat{w}_i d_i}{\sum_{i=1}^{k} \hat{w}_i}$ with weights $\hat{w}_i = 1/(\hat{\tau}^2_{AREML} + \sigma_i^2)$. A value for AREML is found in an iterative fashion, similar to ML and REML. If a step produces a negative value, the estimator is set to 0. When $w_i = w$ for $i = 1, ..., k$, the estimator is identical to REML.

AREML motivated us to make a direct adjustment for the loss in degrees of freedom when the weights vary. In this case, the adjustment $k/(k-1)$ is no longer optimal. An adjustment that is more suitable given variable weights can be made as follows

$$\hat{\tau}^2_{AREML2} = \frac{\sum_{i=1}^{k} \hat{w}_i^2 \left[\left(1 - \frac{\hat{w}_i}{\sum_{i=1}^{k} \hat{w}_i} \right)^{-1} (d_i - \hat{\delta}_{AREML2})^2 - \sigma_i^2 \right]}{\sum_{i=1}^{k} \hat{w}_i{}^2}, \tag{4.3.7}$$

where $\hat{\delta}_{AREML2} = \frac{\sum_{i=1}^{k} \hat{w}_i d_i}{\sum_{i=1}^{k} \hat{w}_i}$ with weights $\hat{w}_i = 1/(\hat{\tau}^2_{AREML2} + \sigma_i^2)$. This estimator is expected to be closely related to REML, we therefore only include a discussion of it for analytic purposes.

4.4. Analytic comparison of heterogeneity variance estimators

In this section, we compare the bias and precision of the five heterogeneity variance estimators algebraically. In the subsequent section, we present a small simulation study to complement the analytic results derived in this section.

4.4.1. Method of moments estimators

Deriving bias and mean square error of methods of moments estimators is straightforward. First remember the generalized Q statistic, i.e. $Q = \sum_{i=1}^{k} a_i (d_i - \bar{d}_+)^2$, where a_i are positive constants. According to Box (1954), we can approximate the distribution of Q in random-effects meta-analysis using a scaled, location shifted, χ^2 distribution, that is, $Q \sim g\chi^2_h$, where $g = \frac{Var(Q)}{2E(Q)}$ and $h = \frac{2[E(Q)]^2}{Var(Q)}$. The expectation and variance of Q can be derived by an application of a theorem in Searle (1971). A general result for the case of one-way random-effects was given in Rao *et al.* (1981). Biggerstaff and Tweedie (1997) presented the first two moments of Q when the weights are proportional to within-study variances. For a statement of the general result, see Appendix C.

Both DL and the ANOVA analogue are unbiased before truncation assuming within-study variances are estimated without error. Negative values can occur with non-zero probability; in this case the estimates are taken to be 0.

4.4.2. DL method

The variance of the DerSimonian and Laird estimator is obtained as follows. First, the first two moments of Q are

$$E(Q) = (k-1) + \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) \tau^2 \tag{4.4.1}$$

and

$$Var(Q) = 2(k-1) + 4 \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) \tau^2 + 2 \left(\sum_{i=1}^{k} w_i^2 - 2\frac{\sum_{i=1}^{k} w_i^3}{\sum_{i=1}^{k} w_i} + \frac{(\sum_{i=1}^{k} w_i^2)^2}{(\sum_{i=1}^{k} w_i)^2} \right) \tau^4. \tag{4.4.2}$$

From expressions (4.4.1) and (4.4.2), we observe that when the between-study variance parameter, τ^2, is 0, the expectation and variance of Q are $k-1$ and $2(k-1)$, respectively, and hence $Q \sim \chi^2_{k-1}$ - which is the distribution of Q under homogeneity.

Now, the variance of the *untruncated* DL method can be calculated directly,

$$Var(\hat{\tau}^2_{DL}) = \frac{Var(Q)}{\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right)^2}. \tag{4.4.3}$$

where $w_i = 1/\sigma_i^2$. To get some intuition about (4.4.3), consider the case of equal weights so that $w_i = w$ for all $i = 1, ..., k$. Then expression (4.4.3) reduces to $\frac{2(\tau^2+\sigma^2)^2}{(k-1)}$. Because we operationalize the between-study variance parameter as a function of within-study variances, we can further simplify the variance of the DL method as

$$Var(\hat{\tau}^2_{DL}) = \frac{2(r+1)^2\sigma^4}{k-1}, \tag{4.4.4}$$

where r typically ranges from 0 to 2 (i.e. $\tau^2 = r\sigma^2$). We can observe two points here. First, a more stable estimate of τ^2 is obtained when the number of studies is large. Second, if we fix the number of studies k, as true heterogeneity increases, the variance of this estimator grows larger. We expect a similar trend for the truncated version of DL.

Since the estimator obtained from the DL method is a truncated estimator, one can analyze its behavior either using numerical methods or simulations. The results presented in the next section are from a simulation study, however, it is instructive to consider the expected value and variance of the truncated DL estimator using numerical integration for a better understanding of why certain properties follow. Note that the expectation of $\hat{\tau}^2_{DL}$can be obtained by conditioning on the random variable Q and weighting by its density function. The conditional expected value is given by

$$E(\hat{\tau}^2) = \int_{k-1}^{\infty} \frac{q-(k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}} \frac{1}{g} f\left(\frac{q}{g}\right) dq, \tag{4.4.5}$$

where $f(\frac{q}{g})$ is the probability density of a scaled χ^2 random variable with h degrees of freedom. We can also derive its variance by noting that

$$Var(\hat{\tau}^2) = E[(\hat{\tau}^2)^2] - E(\hat{\tau}^2)^2. \tag{4.4.6}$$

In the case of $k = 2$,

$$Q = \frac{(d_i - d_2)^2}{\sigma_1^2 + \sigma_2^2 + 2\tau^2}, \tag{4.4.7}$$

which follows a χ^2 distribution with 1 degree of freedom, scaled by $\frac{\sigma_1^2+\sigma_2^2+2\tau^2}{\sigma_1^2+\sigma_2^2} = 1 + \frac{\tau^2}{(\sigma_1^2+\sigma_2^2)/2}$. Because the bias in the DL estimator is primarily due to truncation, we expect bias to decrease as true heterogeneity increases. For small degrees of heterogeneity, the DL method will be more biased, particularly when the number of studies is small. This is because the probability of truncation in these scenarios can be quite high. From expression (4.4.7), we also observe that a natural way to think about heterogeneity is via $\sum_{i=1}^{k} \sigma_i^2/k$. When $k = 2$, this is exactly true, and approximately so when k is larger than 2. If $\sigma_1^2 = \sigma_2^2$, then Q is scaled by $1 + \tau^2/\sigma^2$. If, on the other hand, $\sigma_2^2 = 10\sigma_1^2$, Q is scaled by $1 + (2/11)\tau^2/\sigma^2$ which indicates that the probability of truncation will be higher compared to studies with equal within-study variances. These observations are confirmed using simulations in the next section.

Figure 4.1 displays the probability of truncation for the distribution of Q when there is no heterogeneity for k ranging from 2 to 20. Note that the probability is quite high for small collections of studies

(approximately 68%). While this probability decreases as k increases, even when $k = 20$, the probability of truncation is remains high (approximately 54%). We therefore expect the DL estimator to be positively biased in this scenario, and, in fact, all estimators that limit variance estimates to non-negative values.

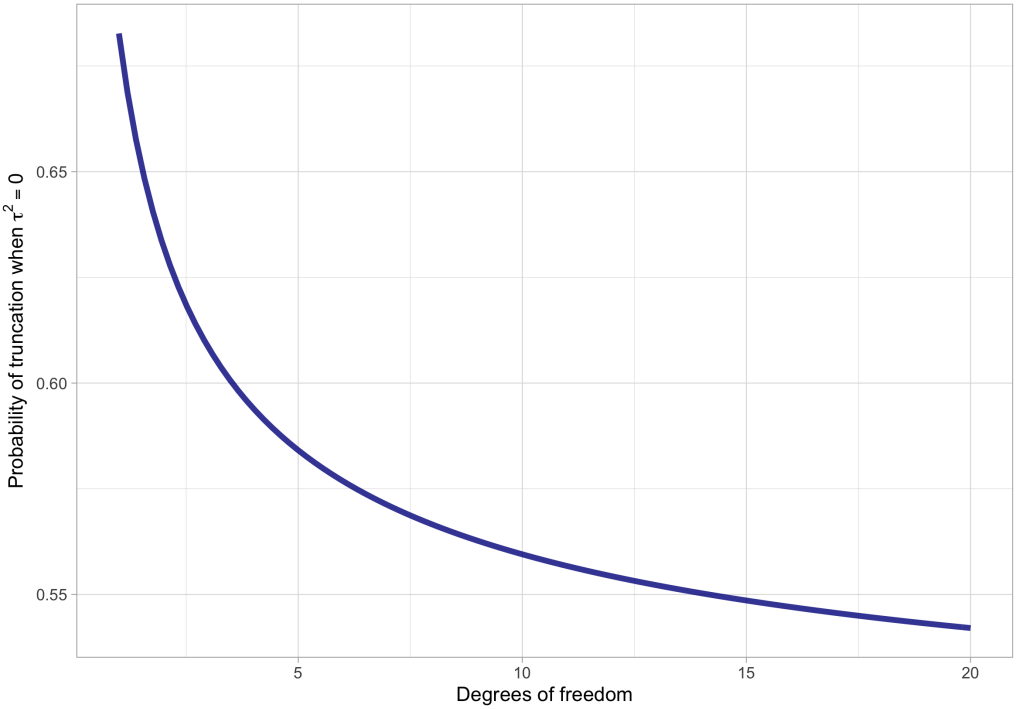

Figure 4.1. $\Pr[Q < (k-1)]$. This plot displays the probability of truncating the distribution of the Q random variable. It shows that even when k is large, the probability of truncation remains high.

In the case of $k = 2$ with unequal weights, the DL method reduces to

$$\hat{\tau}^2_{DL} = \frac{(d_1 - d_2)^2}{2} - \frac{(\sigma_1^2 + \sigma_2^2)}{2}, \tag{4.4.8}$$

which further turns out to be identical to ANOVA, REML and AREML.

4.4.3. ANOVA

The variance of the ANOVA estimator is

$$Var(\hat{\tau}_A^2) = \frac{2(k-2)}{k(k-1)^2}\sum_{i=1}^{k}\sigma_i^4 + \frac{2}{k^2(k-1)^2}\left(\sum_{i=1}^{k}\sigma_i^2\right)^2 + \frac{4}{k(k-1)}\tau^2\sum_{i=1}^{k}\sigma_i^2 + \frac{2}{(k-1)}\tau^4. \tag{4.4.9}$$

This has also been observed by Friedman (2000). The result is derived using a direct application of a general result by Rao *et al.* (1981). When the within-study variances are equal, that is, $\sigma_i^2 = \sigma^2$ for $i = 1, ..., k$, the above expression reduces to $\frac{2(\tau^2+\sigma^2)^2}{(k-1)}$ which is equal to $Var(\hat{\tau}_{DL}^2) = \frac{2(\tau^2+\sigma^2)^2}{(k-1)}$, and similar conclusions to the DL method follow immediately.

Furthermore, when $k = 2$ and $w_1 \neq w_2$, the ANOVA estimator reduced to

$$\hat{\tau}_A^2 = \frac{(d_1 - d_2)^2}{2} - \frac{(\sigma_1^2 + \sigma_2^2)}{2}, \tag{4.4.10}$$

which is exactly equal to DL. We therefore expect to see no differences in the simulation results in this scenario.

4.4.4. Maximum likelihood methods

The asymptotic variance of the Maximum Likelihood estimator can be obtained by using the fact that $d_i - \delta \sim N(0, \sigma_i^2 + \tau^2)$, and assuming the weights are known,

$$Var(\hat{\tau}_{ML}^2) = \frac{\sum_{i=1}^{k} w_i^4 Var[(d_i - \delta)^2 - \sigma_i^2]}{(\sum_{i=1}^{k} w_i^2)^2}. \tag{4.4.11}$$

Now the variance of $(d_i - \delta)^2 = E(d_i - \delta)^4 - [E(d_i - \delta)^2]^2$ can be obtained using the general expression of moments, i.e. $E(X^n) = \frac{d^n M(t)}{dt^n}$ evaluated at $t = 0$. The moments of $(d_i - \delta)$ are

$$E(d_i - \delta)^4 = 3(\sigma_i^2 + \tau^2)^2 \tag{4.4.12}$$

and

$$E(d_i - \delta)^2 = (\sigma_i^2 + \tau^2). \tag{4.4.13}$$

Putting everything together, it follows that

$$Var(\hat{\tau}^2_{ML}) = \frac{2}{\sum_{i=1}^k w_i^2}. \tag{4.4.14}$$

ML attains the Cramer-Rao bound for large k, therefore it follows that it is asymptotically efficient.

Though the maximum likelihood estimator has smallest variance among all estimators considered, i is also the most biased because we assumed the mean is known when deriving it. If the weights are equal across studies, then the ML estimator reduces to

$$\hat{\tau}^2_{ML} = \frac{\sum_{i=1}^k (d_i - \bar{d})}{k} - \sigma^2. \tag{4.4.15}$$

Note that the first term is divided by k instead of $k-1$ - this difference is equivalent to the difference in obtaining the typical sample variance via ML versus REML.

We can also obtain the bias of ML when the weights vary across studies. That is,

$$E(\hat{\tau}^2_{ML}) = \frac{\sum_{i=1}^k w_i^2 E(d_i - \hat{\delta})^2 - \sum_{i=1}^k w_i^2 \sigma_i^2}{\sum_{i=1}^k w_i^2}, \tag{4.4.16}$$

and

$$\begin{aligned} E(d_i - \hat{\delta}) &= E(d_i^2) - 2E(d_i \hat{delta}) + E(\hat{\delta}^2) \\ &= \delta^2 + (\sigma_i^2 + \tau^2) - 2\left[\delta^2 + \frac{w_i}{\sum_{i=1}^k w_i}(\sigma_i^2 + \tau^2)\right] + \left[\sum_{i=1}^k \frac{w_i^2}{(\sum_{i=1}^k w_i)^2}(\sigma_i^2 + \tau^2)\right] + \delta^2 \\ &= (\sigma_i^2 + \tau^2) + \frac{1}{\sum_{i=1}^k w_i}. \end{aligned} \tag{4.4.17}$$

Therefore, the bias of the ML estimator is

$$Bias(\hat{\tau}^2_{ML}) = E(\hat{\tau}^2) - \tau^2 = -\frac{1}{\sum_{i=1}^k w_i}. \tag{4.4.18}$$

Here we note that REML adjusts for this bias, i.e.,

$$\hat{\tau}^2_{REML} = \frac{\sum_{i=1}^k \hat{w}_i^2 [(d_i - \hat{\delta}_{REML})^2 - \sigma_i^2]}{\sum_{i=1}^k \hat{w}_i{}^2} + \frac{1}{\sum_{i=1}^k \hat{w}_i}, \tag{4.4.19}$$

and is therefore a bias reduced estimator in meta-analysis.

The asymptotic variance of REML is given by

$$Var(\hat{\tau}^2_{REML}) = \frac{2}{\sum_{i=1}^{k} w_i^2 - 2\frac{\sum_{i=1}^{k} w_i^3}{\sum_{i=1}^{k} w_i} + \frac{(\sum_{i=1}^{k} w_i^2)^2}{(\sum_{i=1}^{k} w_i)^2}} \tag{4.4.20}$$

(Viechtbauer, 2005). When $w_i = w$ for $i = 1, ..., k$, it is easier to work with AREML for algebraic purposes. In this case, AREML reduces to

$$\hat{\tau}^2_{AREML} = \frac{\sum_{i=1}^{k}(d_i - \bar{d})^2}{k-1} - \sigma^2 \tag{4.4.21}$$

and its variance is $\frac{2(\tau^2+\sigma^2)^2}{k-1}$. This estimator is identical to ANOVA and DL under the equal weights assumption. When the weights differ, as is common practice, differences result in the values obtained from these estimators.

4.5. Simulation evidence

4.5.1. Simulation design

In this article, we have selected the standardized mean difference (SMD) effect size for investigation. For an extensive simulation study considering effect sized based on binary data, see Langan *et al.* (2019). Although they too investigate the SMD effect size, their investigation is mostly restricted to simulations. The purpose of our simulation study is to accompany the analytic results given in the previous section.

The parameter values for this investigation were chosen to understand the effect of sample size patterns, number of studies, and increasing values of the heterogeneity parameter on the estimators' performance. We vary the parameters as follows:

- The number of studies, k, varied from 2 to 10.
- The pattern of sample sizes consisted of equal size studies and half-half large and small studies. The ratio of large to small was 5.

- The heterogeneity variance parameter was defined as a proportion of the simple arithmetic average of the within-study variances, i.e. $\tau^2 = r\sigma^2$, where $\sigma^2 = \sum_{i=1}^{k} \sigma_i^2/k$ and r varied from 0 - indicating no heterogeneity - to 2 - indicating a large degree of heterogeneity.
- Without loss of generality, δ was set to 0.5, indicating a medium effect.

Data were generated according to a random-effects model. This involves two steps: First, true effect sizes, δ_i, were generated from the normal distribution with expectation δ and variance τ^2. Second, SMD estimates, d_i, were generated from the normal distribution with expectation δ_i and within-study variances σ_i^2. To remove the small sample bias in d_i, we multiply each SMD estimate by $c(f) = 1 - \frac{3}{4f-1}$, where $f = N - 2 = 2(n-1)$ (given an equal number of observations in treatment and control groups), resulting in Hedges' g. We do not apply the bias correction factor to the within-study variances since doing so leads to underestimates of the true within-study variances (Hedges, 1985).

For each combination of the parameters, the process was repeated $10,000$ times. For each iteration, the four heterogeneity variance estimators were calculated. For each estimator, grouped by the number of studies, pattern of sample sizes and heterogeneity parameter, performance was assessed by computing bias, relative bias, variance, mean square error, proportional mean square error and proportion of zero estimates of τ^2.

For maximum likelihood methods, i.e. ML and REML, we used a similar procedure to Sidik and Jonkman (2007). That is, the number of iterations for the ML and REML estimators to converge was restricted to a maximum of 20. If an iteration step produced a negative value, it was set to 0. The stopping criteria for assessing convergence was

$$\frac{\hat{\tau}_{j+1}^2 - \hat{\tau}_j^2}{1 + \hat{\tau}_j^2} < 0.00001, \tag{4.5.1}$$

and in situations where convergence was not attained, all other estimates were discarded. As a preview, both iterative estimators converged rather fast if within-study sample sizes were equal across studies and for larger numbers of studies. In the case of unbalancedness, the smallest number of replicates was 7773 when k was 2 and $\bar{n}$ was 50, which means that approximately 22% of the values were discarded. This

improved with larger values of $\bar{n}$ and k. For example, when $k = 4$, only approximately 4% of the values were discarded. Note that when $k = 2$, REML is identical to ANOVA and DL, hence using it as a method to estimate the heterogeneity variance in this case seems unnecessary. The properties computed were based on the actual number of replications rather than $10,000$. Additionally, we set the initial value of τ^2 to 0.

4.5.2. Performance criteria

The performance criteria for the estimators include: absolute bias, relative bias, variance, mean square error, proportional mean square error and proportion of zero estimates of τ^2. These performance measures are in line with other studies considering estimation of the heterogeneity variance parameter. For ease of interpretation, note that a relative bias of 100% means that the estimate of τ^2 is two times larger than τ^2, on average. Similar interpretation can be used for proportional MSE, i.e. a proportional MSE of 100% means than the MSE of the estimator is equal to the true τ^2. These measures are used for easier comparison across different scenarios considered, as recommended by Langan *et al.* (2019).

4.5.3. Results of simulations

The empirical properties of four estimators are presented in the next few subsections. Note that the properties of AREML are very similar to REML, hence they are omitted from this simulation study. We present results for $\delta = 0.5$ only. This is because the magnitude of the effect size has little bearing on the results and thus the empirical properties presented are generalizable to $\delta = 0.2$ and $\delta = 0.8$. For easier comparisons, we split the properties into three subsections. Section 4.5.4 provides results for the proportion of zero estimates obtained from the methods. Section 4.5.5 provides the results of absolute and relative bias, while Section 4.5.6 presents the results for the mean square error and proportional mean square error of the estimators. Extra simulation results are provided in Appendix C.

4.5.4. Proportion of zero estimates

Figure 4.2 displays the empirical distribution of the DL estimator for $k = 2$, $k = 5$ and $k = 10$ when within-study sample sizes are $n = 30$ across all studies. Two points are observed here. First, as heterogeneity increases, the probability of truncation decreases for all k (going across columns for a given k). Second, as k increases, the probability of truncation decreases for various degrees of heterogeneity (going down a column for a given value of heterogeneity). Similar results can be shown for the other estimators. In addition, this plot confirms that when there is no to small heterogeneity, the probability of truncation remains high even when k is large, leading to considerable bias in the estimates obtained in this scenario.

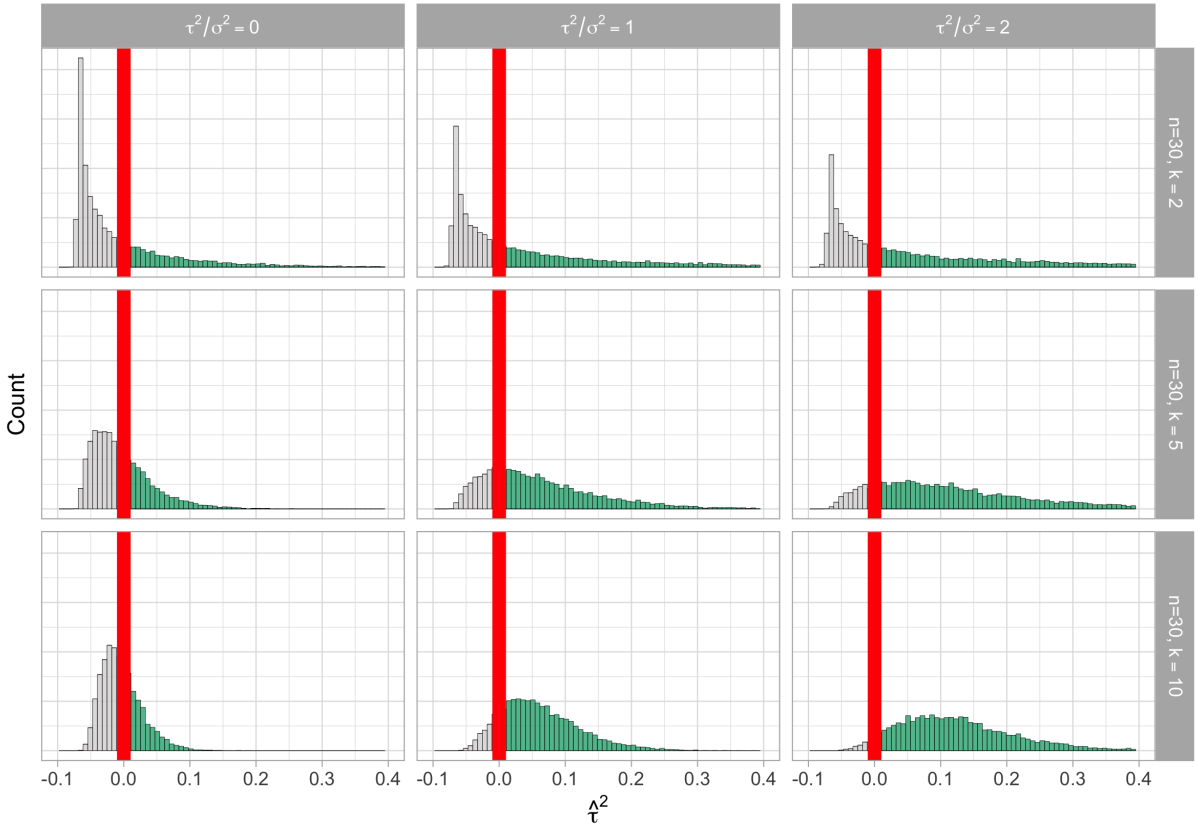

Figure 4.2. The Empirical Distribution of $\hat{\tau}^2_{DL}$. This plot displays the empirical distribution of the DL estimator of the heterogeneity variance parameter. The columns represent three levels of heterogeneity (none, medium, large). The rows represent meta-analyses of $2, 5$ and 10 studies, when $n = 30$.

Figures 4.3 and 4.4 show the proportion of zero estimates obtained using the four estimators for balanced sample size patterns (i.e. n of $10, 30, 50$ and 100) and unbalanced patterns (i.e. $\bar{n}$ of 50 and 100). For the unbalanced setting, we included half-half large and small studies, where the ratio of large to small was 5. It is apparent that the proportion of zero estimates is quite high for $k = 2$ for no heterogeneity - approximately 0.75 for ANOVA DL, and REML and roughly 0.80 for ML. The proportion decreases as heterogeneity increases, however, when there are only two studies in a meta-analysis, even in the case of a large degree of heterogeneity, the proportion remains hight at approximately 0.44. Furthermore, as k increases, when there is no heterogeneity, the proportion of zero estimates remains high for all within-study sample sizes considered.

4.5.5. Bias and relative bias

Figures 4.3 and 4.4 display the absolute and relative bias of the estimators in balanced meta-analyses, respectively, while 4.3 and 4.4 display the same in unbalanced settings. From the first figure, it is seen that the four estimators overestimate the heterogeneity variance when true $\tau^2 = 0$; this positive bias is more severe when synthesizing small studies (i.e. $n = 10$ across studies) and when k is small. As the number of studies increases, all estimators underestimate the heterogeneity variance for increasing τ^2. When n is at least 30, and k is sufficiently large (at least 6), absolute biases are essentially 0.

Relative biases are well over 50% when $k = 2$ and $\tau^2 = 0$ for all but the maximum likelihood estimator. ML severely underestimates the heterogeneity variance across all scenarios investigated, however, the severity decreases as k and τ^2 increase. It is apparent that when within-study sample sizes are small, all estimators perform poorly, but some exhibit better behavior than others. In particular, it can be seen that the ANOVA estimator is less downwardly biased, and is thus recommended when synthesizing small studies for all k. As n increases, differences between DL, REML and ANOVA become negligible, hence one can use any of these three estimators.

4.5.6. Proportional mean square error

As shown in Figures 4.9 and 4.10, ML has smallest mean square compared to the other three estimators across all scenarios, including balanced and unbalanced settings. This is due to its large negative bias - which is not necessarily a good property. Proportional mean square error of the ML estimator is larger in scenarios of small studies (i.e. $n = 10$), and decreases slightly in scenarios containing large studies (i.e. $n = 100$). DL and REML, on the other hand, have similar proportional MSE across all scenarios considered. ANOVA has slightly larger MSE for small heterogeneity compared to DL and REML. This is expected since ANOVA uses equal weights, while DL and REML assign weights inversely proportional to within-study variances and total variances of the SMD estimates, respectively. Noting that the main differences between the ANOVA method, DL and REML is in the weights they assign, it seems reasonable to expect similar results for their properties, particularly for DL and REML.

4.6. Comment on merits of estimators

Based on the entirety of our investigation, we make the following concluding remarks about the merits of each heterogeneity estimator when the effect size of interest is the Standardized Mean Difference and the number of studies in meta-analysis is small. Although the performance criteria included variance and mean square error for all the estimators considered, it appears that bias and relative bias may be better criteria for evaluating the performance of variance estimators. This is because all four estimators (ANOVA, DL, ML and REML) are restricted to non-negative values, which means that their sampling distributions are more concentrated, and therefore their variances are smaller leading to smaller mean square error (Sidik and Jonkman, 2007). Consequently, bias is perhaps more informative when evaluating heterogeneity variance estimators and the following recommendations reflect this.

First, we emphasize that ANOVA, DL and REML produce identical estimates when study sample sizes are equal across studies, with the exception of meta-analyses of small studies (i.e. $n < 30$). These three estimators are positively biased in scenarios of no heterogeneity, but negatively biased as heterogeneity increases in meta-analyses containing small studies. The least biased of these alternatives is ANOVA, which can be recommended for meta-analyses containing small studies, for all k. As within-study sample

sizes increase, however, any of ANOVA, DL or REML can be used to estimate the between-study variance parameter in a small sample meta-analytic setting (i.e. small k). When only two studies are synthesized, ANOVA, DL and REML are identical; consequently any of these three estimators can be used in this scenario. ML severely underestimates true heterogeneity across all scenarios, with bias being more pronounced when k is small. In view of this, ML is not recommended.

4.7. Summary and discussion

This chapter has provided details on two types of heterogeneity variance estimators, grouped into method of moments and maximum likelihood methods. In Sections 4.2 and 4.3, we provided details of the five estimators considered throughout that can be used to estimate the between-study variance parameter in meta-analysis. In Section 4.4, we compared the estimators analytically. Specifically, we derived the biases and variances of each estimator, where possible to do so, and provided insight into bias versus variance trade-offs, particularly when the number of studies is small. To complement the analytic results, we conducted a simulation study in Section 4.5. Throughout the chapter, our focus has been the Standardized Mean Difference effect size and Hedges' g to estimate the SMD. There were two questions that guided this chapter. First, given the numerous heterogeneity variance estimators available, how does one choose an estimator to estimate the heterogeneity variance parameter when the number of studies is small and sample sizes vary across studies? Second, do properties of these estimators propagate to the estimated mean and its standard error in random-effects models? To answer these questions, we made comparisons in a systematic way, both analytically and using simulations.

Providing recommendations on the choice of estimator requires considering both bias and variance of the estimators separately, across different scenarios. Although no one estimator can be recommended for all situations, some generalizations are possible. These generalizations are useful when interest lies in obtaining an estimate of τ^2 itself. If, however, interest lies in choosing an estimator that leads to unbiased estimation of the summary treatment effect with minimum variance, then the choice of estimator has little effect on the mean of the distribution of effects. Further, the variances that use squared sample residuals

to estimate the variance of the summary effect are robust to heterogeneity variance estimates, leading to coverage probabilities that are generally close to nominal.

When interest lies on reporting an estimate of heterogeneity, the following can be noted. First, both method of moments and maximum likelihood estimators are biased and unstable in small samples (i.e. small k). The estimators show better behavior as the number of studies increases and for larger degrees of heterogeneity. When there is no heterogeneity, however, bias of all estimators remains large even when k is large. This is primarily due to restricting estimates to be non-negative, an issue previously noted in the literature (Chung *et al.*, 2013; Viechtbauer, 2005). Of the five estimators, the maximum likelihood estimator has smallest mean square error, which is primarily due to its large negative bias. When study sample sizes are equal across studies, DL reduces to the ANOVA method, and both are identical to the REML estimator. All estimators show better behavior when n is sufficiently large ($n > 30$). When sample sizes vary across studies, the DL estimator is preferred over its alternatives in the presence of small to moderate heterogeneity, while the ANOVA estimator is favored when there is a large degree of heterogeneity. This recommendation is based on bias alone. The REML estimator, on the other hand, is a compromise between ANOVA and DL.

As a last note, although none of the estimators discussed perform satisfactorily in a small sample meta-analytic context, it it imperative to report a point estimate of the heterogeneity variance, in addition to its standard error. While standard errors will likely be large, they are nevertheless useful to inform the reader about the amount of information that is being provided. Finally, we emphasize that these properties are not generalizable to other effect sizes that are functions of proportions (i.e. risk difference, odds ratio and risk ratio). This is because the weights used in such situations are much more dependent on the underlying stochastic data, which pose further challenges in estimating the between-study variance parameter reliably. Further analytic work is needed to understand the applicability of heterogeneity variance estimators to estimate heterogeneity variance when the outcome in a study is binary.

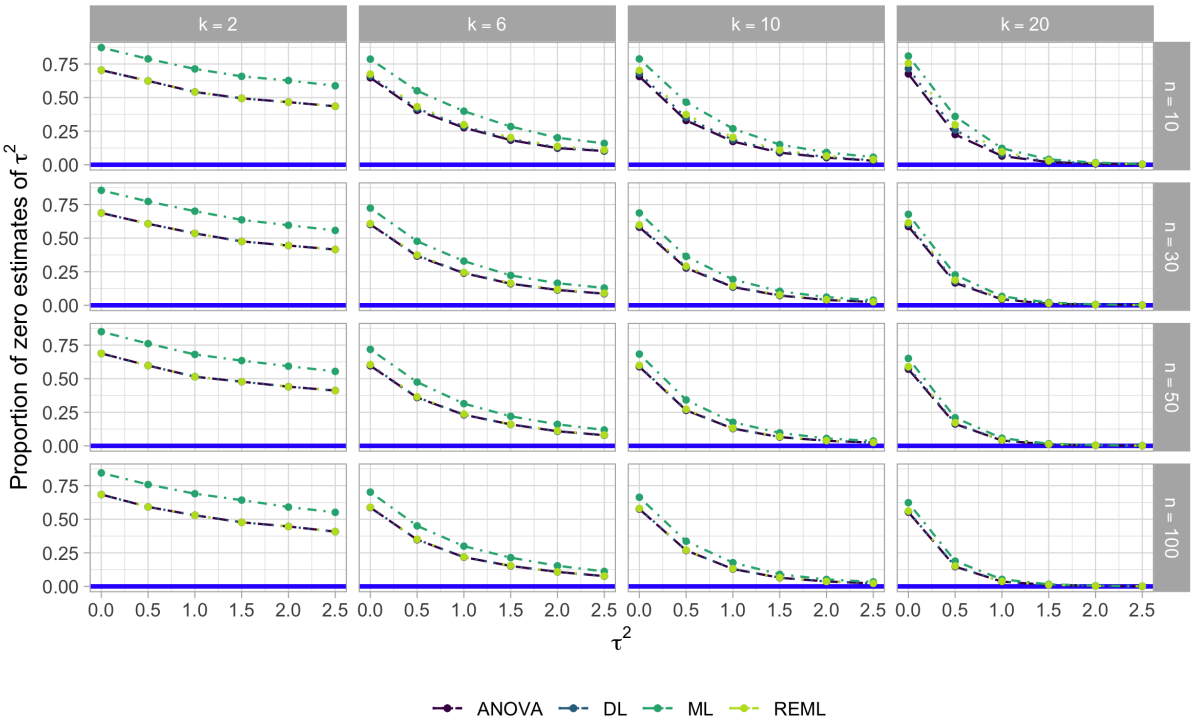

Figure 4.3. Proportion of Zero Estimates of τ^2 in a Balanced Setting. This plot displays the proportion of zero estimates of τ^2 obtained from four methods: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. Within-study sample sizes, n, range from 10 to 100 and are equal across studies. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

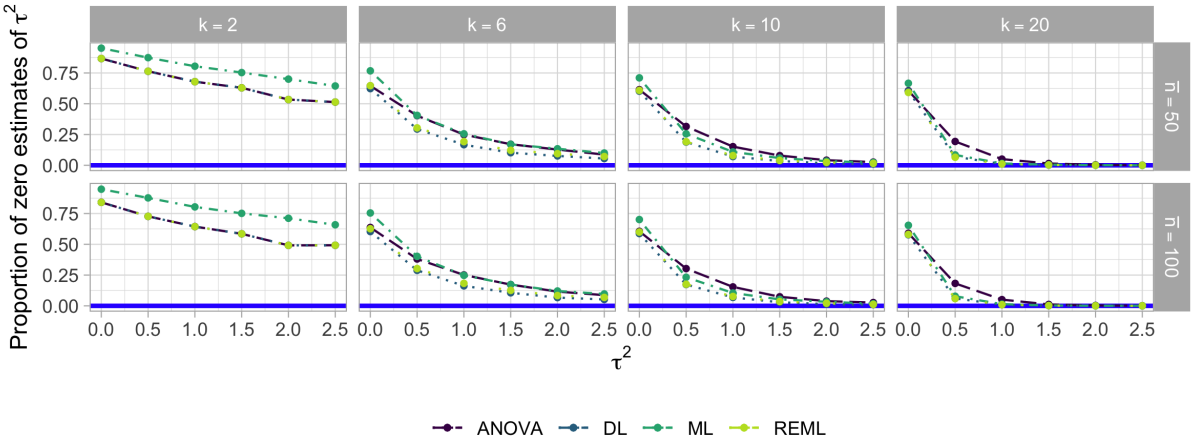

Figure 4.4. Proportion of Zero Estimates of τ^2 in an Unbalanced Setting. This plot displays the proportion of zero estimates of τ^2 obtained from four methods: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

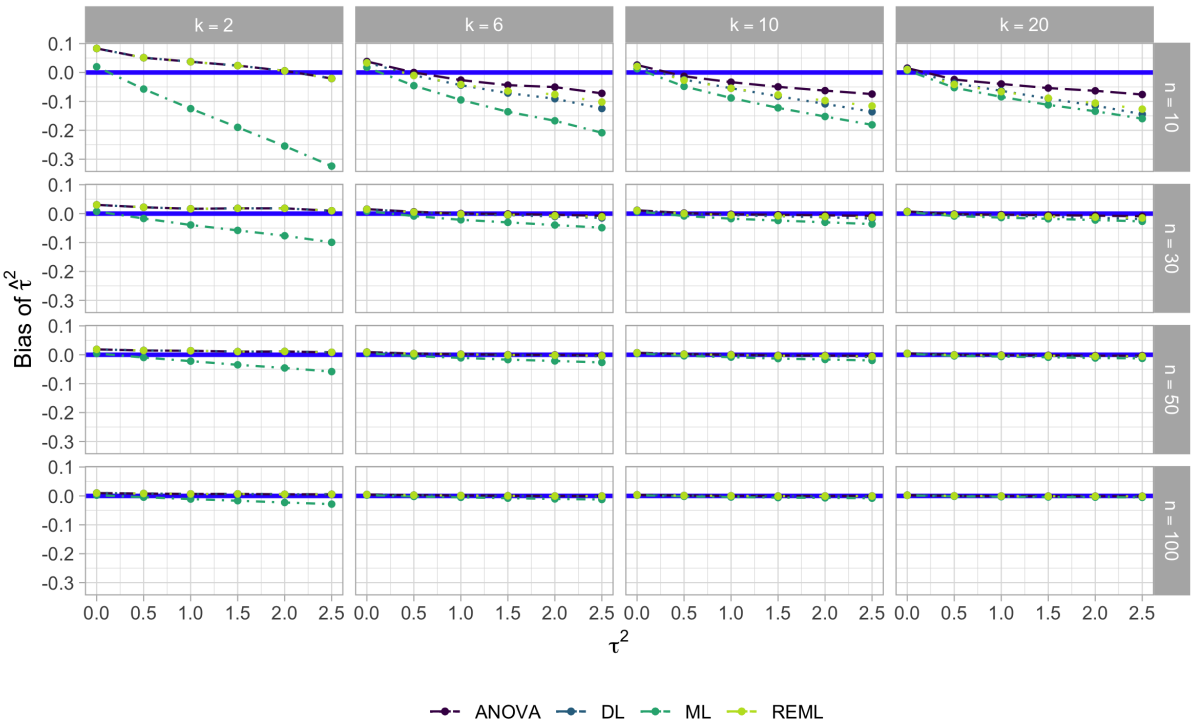

Figure 4.5. Absolute bias of $\hat{\tau}^2$ in a Balanced Setting. This plot displays the absolute bias of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. Within-study sample sizes, n, range from 10 to 100 and are equal across studies. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

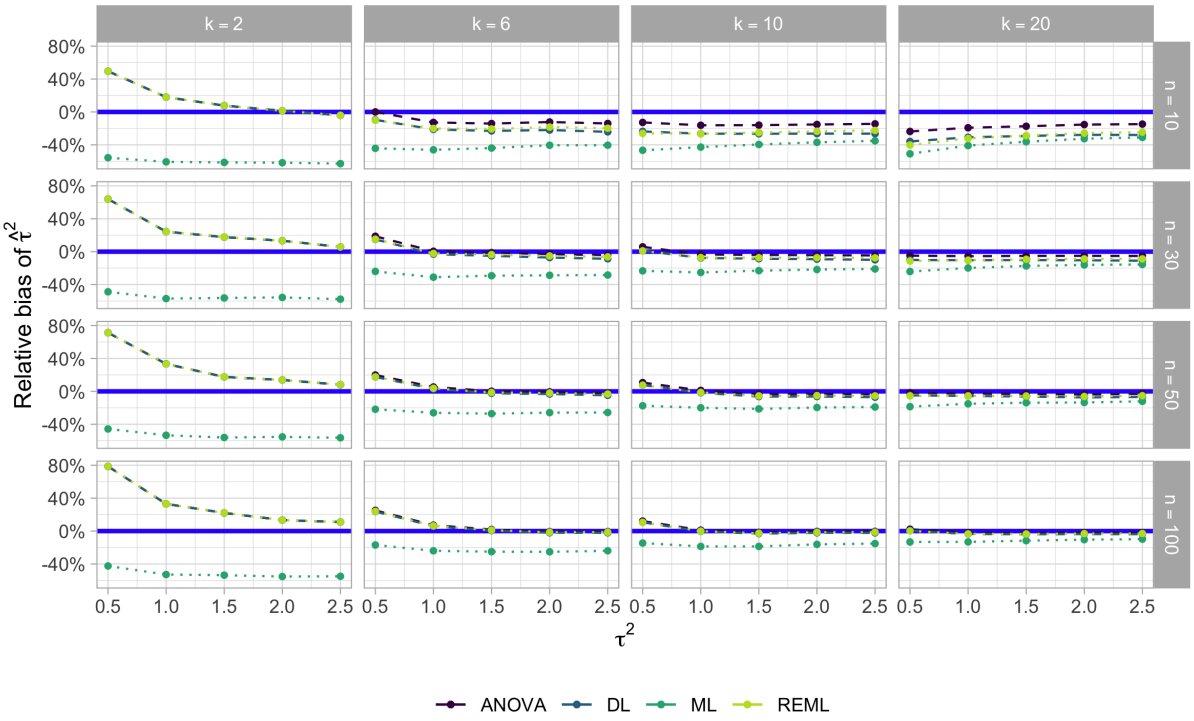

Figure 4.6. Relative bias of $\hat{\tau}^2$ in a Balanced Setting. This plot displays the relative bias of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. Within-study sample sizes, n, range from 10 to 100 and are equal across studies. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

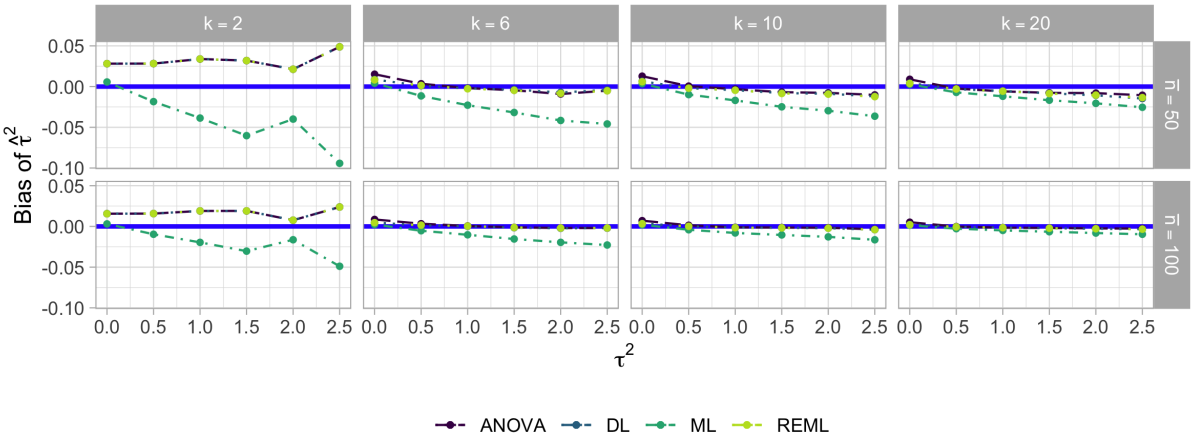

Figure 4.7. Absolute bias of $\hat{\tau}^2$ in an Unbalanced Setting. This plot displays the absolute bias of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

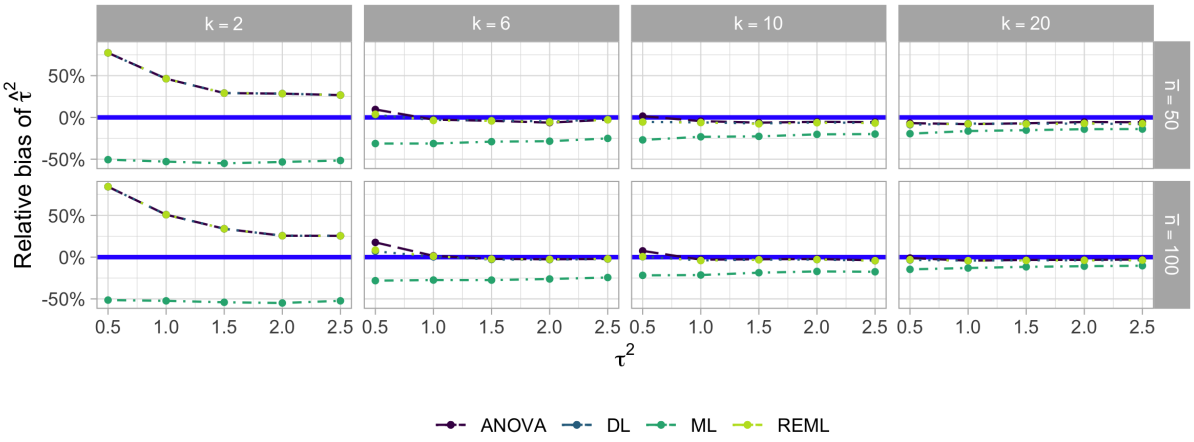

Figure 4.8. Relative bias of $\hat{\tau}^2$ in an Unbalanced Setting. This plot displays the relative bias of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

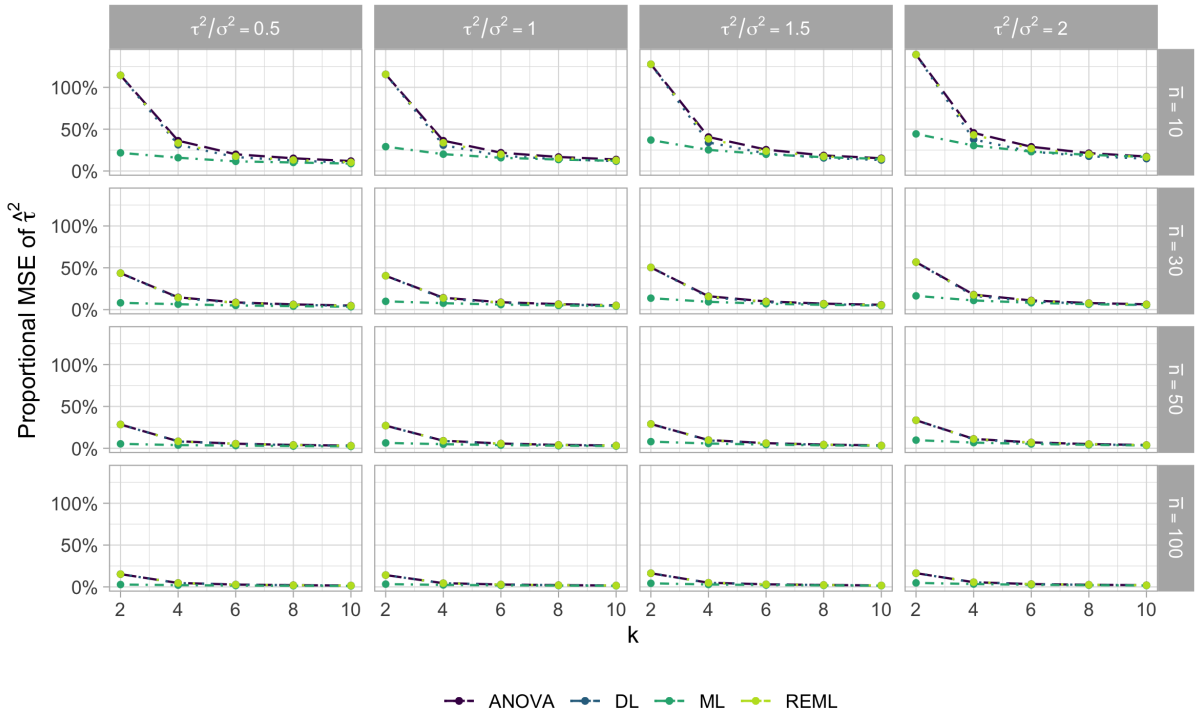

Figure 4.9. Proportional mean square error (MSE) of $\hat{\tau}^2$ in a Balanced Setting. This plot displays the proportional mean square error of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. Within-study sample sizes, n, range from 10 to 100 and are equal across studies. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

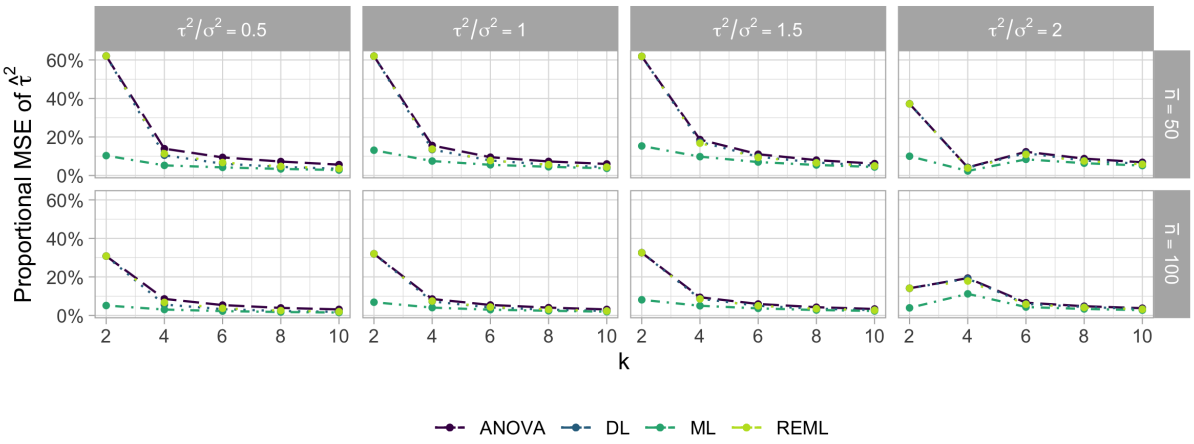

Figure 4.10. Proportional mean square error (MSE) of $\hat{\tau}^2$ in an Unbalanced Setting. This plot displays the relative bias of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

References

Biggerstaff, B. J. and Tweedie, R. L. (1997) Incorporating Variability in Estimates of Heterogeneity in the Random Effects Model in Meta-Analysis. *Statistics in Medicine*, **16**, 753–768. DOI: 10.1002/(SICI)1097-0258(19970415)16:7<753::AID-SIM494>3.0.CO;2-G.

Box, G. E. P. (1954) Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification. *Annals of Mathematical Statistics*, **25**, 290–302. Institute of Mathematical Statistics. DOI: 10.1214/aoms/1177728786.

Brockwell, S. E. and Gordon, I. R. (2001) A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, **20**, 825–840. DOI: 10.1002/sim.650.

Card, N. A. (2011) *Applied Meta-Analysis for Social Science Research.* Methodology in The Social Sciences. New York: Guilford Publications.

Chung, Y., Rabe-Hesketh, S. and Choi, I.-H. (2013) Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, **32**, 4071–4089. DOI: 10.1002/sim.5821.

Cochran, W. G. (1937) Problems Arising in the Analysis of a Series of Similar Experiments. *Supplement to the Journal of the Royal Statistical Society*, **4**, 102–118. [Wiley, Royal Statistical Society]. DOI: 10.2307/2984123.

Cochran, W. G. (1943) The Comparison of Different Scales of Measurement for Experimental Results. *The Annals of Mathematical Statistics*, **14**, 205–216. Institute of Mathematical Statistics.

Cochran, W. G. (1954) The Combination of Estimates from Different Experiments. *Biometrics*, **10**, 101–129. [Wiley, International Biometric Society]. DOI: 10.2307/3001666.

Cooper, H., Hedges, L. V. and Valentine, J. C. (2009) *The Handbook of Research Synthesis and Meta-Analysis.* Russell Sage Foundation.

Davey, J., Turner, R. M., Clarke, M. J., et al. (2011) Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC medical research methodology*, **11**, 160. DOI: 10.1186/1471-2288-11-160.

Davidson, R. and MacKinnon, J. G. (1993) *Estimation and Inference in Econometrics. OUP Catalogue.* Oxford University Press.

DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177–188. Elsevier. DOI: 10.1016/0197-2456(86)90046-2.

DerSimonian, R. and Laird, N. (2015) Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, **45**, 139–145. 10th Anniversary Special issue. DOI: 10.1016/j.cct.2015.09.002.

Eicker, F. (1963) Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions. *The Annals of Mathematical Statistics*, **34**, 447–456. Institute of Mathematical Statistics. DOI: 10.1214/aoms/1177704156.

Eicker, F. (1967) Limit theorems for regressions with unequal and dependent errors. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 59–82. University of California Press.

Emerson, J. D., Hoaglin, D. C. and Mosteller, F. (n.d.) A MODIFIED RANDOM-EFFECT PROCEDURE FOR COMBINING RISK DIFFERENCE IN SETS OF 2 x 2 TABLES FROM CLINICAL TRIALS., 22.

Friede, T., Röver, C., Wandel, S., et al. (2017) Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods*, **8**, 79–91. DOI: 10.1002/jrsm.1217.

Friedman, L. (2000) Estimators of Random Effects Variance Components in Meta-Analysis. *Journal of Educational and Behavioral Statistics*, **25**, 1–12. DOI: 10.3102/10769986025001001.

Gart, J. J. and Zweifel, J. R. (1967) On the Bias of Various Estimators of the Logit and Its Variance with Application to Quantal Bioassay. *Biometrika*, **54**, 181–187. [Oxford University Press, Biometrika Trust]. DOI: 10.2307/2333861.

GLASS, G. V. (1976) Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, **5**, 3–8. American Educational Research Association. DOI: 10.3102/0013189X005010003.

Haldane, B. J. B. S. (1956) The Estimation and Significance of the Logarithm of a Ratio of Frequencies. *Annals of Human Genetics*, **20**, 309–311. DOI: 10.1111/j.1469-1809.1955.tb01285.x.

Hardy, R. J. and Thompson, S. G. (1996) A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, **15**, 619–629. DOI: 10.1002/(SICI)1097-0258(19960330)15:6<619::AID-SIM188>3.0.CO;2-A.

Hardy, Rebecca J. and Thompson, S. G. (1996) A Likelihood Approach to Meta-Analysis with Random Effects. *Statistics in Medicine*, **15**, 619–629. DOI: 10.1002/(SICI)1097-0258(19960330)15:6<619::AID-SIM188>3.0.CO;2-A.

Hartung, J. (1999b) An Alternative Method for Meta-Analysis. *Biometrical Journal*, **41**, 901–916. DOI: 10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W.

Hartung, J. (1999a) An Alternative Method for Meta-Analysis. *Biometrical Journal*, **41**, 901–916. DOI: 10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W.

Hartung, J. (2008) *Statistical Meta-Analysis with Applications*. Wiley series in probability and statistics. Hoboken, N.J: Wiley.

Hartung, J. and Knapp, G. (2001a) A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, **20**, 3875–3889. DOI: 10.1002/sim.1009.

Hartung, J. and Knapp, G. (2001b) On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, **20**, 1771–1782. DOI: 10.1002/sim.791.

Harville, D. A. (1977) Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320–338. Taylor & Francis. DOI: 10.1080/01621459.1977.10480998.

Hedges, L. V. (1983) A random effects model for effect sizes. *Psychological Bulletin*, **93**, 388–395. American Psychological Association. DOI: 10.1037/0033-2909.93.2.388.

Hedges, L. V. (1985) *Statistical Methods for Meta-Analysis*. Orlando: Academic Press.

Hedges, L. V. and Vevea, J. L. (1998) Fixed- and random-effects models in meta-analysis. *Psychological Methods*, **3**, 486–504. American Psychological Association. DOI: 10.1037/1082-989X.3.4.486.

Horn, S. D., Horn, R. A. and Duncan, D. B. (1975) Estimating Heteroscedastic Variances in Linear Models. *Journal of the American Statistical Association*, **70**, 380–385. DOI: 10.1080/01621459.1975.10479877.

Huber, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 221–233. University of California Press.

Hunter, J. E. and Schmidt, F. L. (2000) Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge. *International Journal of Selection and Assessment*, **8**, 275–292. DOI: 10.1111/1468-2389.00156.

IntHout, J., Ioannidis, J. P. and Borm, G. F. (2014) The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, **14**, 25. DOI: 10.1186/1471-2288-14-25.

Kacker, R. N. (2004) Combining information from interlaboratory evaluations using a random effects model. *Metrologia*, **41**, 132–136. IOP Publishing. DOI: 10.1088/0026-1394/41/3/004.

Langan, D., Higgins, J. P. T., Jackson, D., et al. (2019) A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, **10**, 83–98. DOI: 10.1002/jrsm.1316.

Li, A. H., Lo, M., Crawshaw, J. E., et al. (2021) Interventions for increasing solid organ donor registration. *Cochrane Database of Systematic Reviews*. DOI: 10.1002/14651858.CD010829.pub2.

MacKinnon, J. G. and White, H. (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, **29**, 305–325. DOI: 10.1016/0304-4076(85)90158-7.

Morris, C. N. (1983) Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, **78**, 47–55. [American Statistical Association, Taylor & Francis, Ltd.]. DOI: 10.2307/2287098.

PETTIGREW, H. M., GART, J. J. and THOMAS, D. G. (1986) The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika*, **73**, 425–435. DOI: 10.1093/biomet/73.2.425.

Rao, P. S. R. S., Kaplan, J. and Cochran, W. G. (1981) Estimators for the One-Way Random Effects Model with Unequal Error Variances. *Journal of the American Statistical Association*, **76**, 89–97. [American Statistical Association, Taylor & Francis, Ltd.]. DOI: 10.2307/2287050.

Röver, C., Knapp, G. and Friede, T. (2015) Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, **15**, 99. DOI: 10.1186/s12874-015-0091-1.

Schmidt, F. L. and Hunter, J. E. (2014) *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Sage Publications.

Schmidt, F. L. and Hunter, J. E. (2015) *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications, Ltd. DOI: 10.4135/9781483398105.

Searle, S. R. (1971) *Linear Models*. A Wiley publication in mathematical statistics. New York: Wiley.

Seide, S. E., Röver, C. and Friede, T. (2019) Likelihood-based random-effects meta-analysis with few studies: Empirical and simulation studies. *BMC Medical Research Methodology*, **19**, 16. DOI: 10.1186/s12874-018-0618-3.

Sidik, K. and Jonkman, J. N. (2002a) A simple confidence interval for meta-analysis. *Statistics in Medicine*, **21**, 3153–3159. DOI: 10.1002/sim.1262.

Sidik, K. and Jonkman, J. N. (2002b) A simple confidence interval for meta-analysis. *Statistics in Medicine*, **21**, 3153–3159. DOI: 10.1002/sim.1262.

Sidik, K. and Jonkman, J. N. (2005) Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 367–384. DOI: 10.1111/j.1467-9876.2005.00489.x.

Sidik, K. and Jonkman, J. N. (2006) Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis*, **50**, 3681–3701. DOI: 10.1016/j.csda.2005.07.019.

Sidik, K. and Jonkman, J. N. (2007) A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, **26**, 1964–1981. DOI: 10.1002/sim.2688.

Stangl, D. K., Berry, D. A. and NetLibrary, I. (2000) *Meta-Analysis in Medicine and Health Policy*. Biostatistics 4. New York ; Marcel Dekker.

Sutton, A. J. (2000) *Methods for Meta-Analysis in Medical Research*. Wiley series in probability and mathematical statistics. Chichester ; John Wiley.

Thompson, S. G. and Sharp, S. J. (1999) Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, **18**, 2693–2708. DOI: 10.1002/(SICI)1097-0258(19991030)18:20<2693::AID-SIM235>3.0.CO;2-V.

Tipton, E. (2013) Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods*, **4**, 169–187. DOI: 10.1002/jrsm.1070.

Veroniki, A. A., Jackson, D., Viechtbauer, W., et al. (2016) Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, **7**, 55–79. DOI: 10.1002/jrsm.1164.

Viechtbauer, W. (2005) Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, **30**, 261–293. DOI: 10.3102/10769986030003261.

White, H. (1980) A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48**, 817–838. [Wiley, Econometric Society]. DOI: 10.2307/1912934.

White, H. (1982) Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, **50**, 1–25. [Wiley, Econometric Society]. DOI: 10.2307/1912526.

Yates, F. and Cochran, W. G. (1938) The analysis of groups of experiments. *The Journal of Agricultural Science*, **28**, 556–580. Cambridge University Press. DOI: 10.1017/S0021859600050978.

APPENDIX A

Chapter 2

A.1. Proofs

In this section, we provide proofs of several remarks made in Chapter 2. Common to the following proofs is the distributional assumption of effect estimates, i.e. $d_i \sim AN(\delta, \sigma_i^2 + \tau^2)$. The weights and normalized weights are denoted by $w_i = 1/(\sigma_i^2 + \tau^2)$ and $\beta_i = w_i / \sum_{i=1}^k w_i$, respectively, and are treated as known throughout.

Remark 1. The modified variance estimator of the Hartung-Knapp method, $\widehat{Var}(\bar{d}_+) = \sum_{i=1}^k \beta_i (d_i - \bar{d}_+)^2/(k-1)$, is unbiased for the true variance, $1/W$.

Proof:

We can rewrite $S(\beta) = \sum_{i=1}^k \beta_i (d_i - \bar{d}_+)^2$ as $S(\beta) = \sum_{i=1}^k \beta_i d_i^2 - (\sum_{i=1}^k \beta_i d_i)^2$. Take the expected value of the first term in $S(\beta)$,

$$\begin{aligned}
E\left[\sum_{i=1}^k \beta_i d_i^2\right] &= \sum_{i=1}^k \beta_i E(d_i^2) \\
&= \sum_{i=1}^k \beta_i E[(d_i - \delta + \delta)^2] \\
&= \sum_{i=1}^k \beta_i E[(d_i - \delta)^2 + 2\delta(d_i - \delta) + \delta^2] \\
&= \sum_{i=1}^k \beta_i [E(d_i - \delta)^2 + 2\delta E(d_i - \delta) + \delta^2] \\
&= \sum_{i=1}^k \beta_i [E(d_i - \delta)^2 + \delta^2] \\
&= \sum_{i=1}^k \beta_i [Var(d_i) + \delta^2]
\end{aligned}$$

$$= \sum_{i=1}^{k} \beta_i(\sigma_i^2 + \tau^2) + \delta^2 \sum_{i=1}^{k} w_i.$$

Now let $Y = \sum_{i=1}^{k} \beta_i d_i$ and take the expected value of Y^2,

$$\begin{aligned}
E[Y^2] &= E[(Y - \mu_Y + \mu_y)^2] \\
&= E[(Y - \mu_Y)^2 + 2\mu_Y(Y - \mu_Y) + \mu_Y^2] \\
&= E(Y - \mu_Y)^2 + 2\mu_Y E(Y - \mu_Y) + \mu_Y^2 \\
&= Var(Y) + \mu_Y^2 \\
&= Var\left(\sum_{i=1}^{k} \beta_i d_i\right) + \mu_Y^2 \\
&= \sum_{i=1}^{k} \beta_i^2 Var(d_i) + \mu_Y^2 \\
&= \sum_{i=1}^{k} \beta_i^2(\sigma_i^2 + \tau^2) + \delta^2.
\end{aligned}$$

Note that $\mu_Y^2 = \left[E(\sum_{i=1}^{k} \beta_i d_i)\right]^2 = \delta^2$. Putting everything together, we get

$$\begin{aligned}
E[\widehat{Var}(\bar{d}_+)] &= \frac{1}{k-1} \frac{k-1}{\sum_{i=1}^{k} w_i} \\
&= \frac{1}{W}.
\end{aligned}$$

Remark 2. Variance estimator $\widehat{Var}(\bar{d}_+)_{HC0} = \sum_{i=1}^{k} \beta_i^2(d_i - \bar{d}_+)^2$ is biased for $Var(\bar{d}_+) = \sum_{i=1}^{k} \beta_i^2(\sigma_i^2 + \tau^2)$.

Proof:

Rewrite $\sigma_i^2 + \tau^2 = \phi_i^2$. We proceed by rewriting the sample squared residual as follows

$$\hat{e}_i^2 = (d_i - \bar{d}_+)^2 = d_i^2 - 2d_i \sum_{i=1}^{k} \beta_i d_i + \left(\sum_{i=1}^{k} \beta_i d_i\right)^2.$$

Now take the expected value of the corresponding terms,

$$E(d_i^2) = E(d_i - \delta)^2 + \delta^2 = Var(d_i) + \delta^2 = \phi_i^2 + \delta^2,$$

and

$$\begin{aligned} E\left(d_i \sum_{i=1}^{k} \beta_i d_i\right) &= E(d_i) E\left(\sum_{i=1}^{k} \beta_i d_i\right) + Cov\left(d_i, \sum_{i=1}^{k} \beta_i d_i\right) \\ &= \delta^2 + \beta_i \phi_i^2, \end{aligned}$$

and finally,

$$E\left[\left(\sum_{i=1}^{k} \beta_i d_i\right)^2\right] = \left[\sum_{i=1}^{k} \beta_i^2 \phi_i^2\right] + \delta^2.$$

Putting it all together, we obtain

$$\begin{aligned} E(d_i - \bar{d}_+)^2 &= \phi_i^2 - 2\beta_i \phi_i^2 + \sum_{i=1}^{k} \beta_i^2 \phi_i^2 \\ &= \phi_i^2 \left[1 - 2\beta_i + \frac{\sum_{i=1}^{k} \beta_i^2 \phi_i^2}{\phi_i^2}\right] \\ &= \phi_i^2 (1 - h_i), \end{aligned}$$

where $h_i = 2\beta_i - \frac{\sum_{i=1}^{k} \beta_i^2 \phi_i^2}{\phi_i^2}$. Note that h_i further reduces to β_i under $\phi_i^2 = \sigma_i^2 + \tau^2$.

Remark 3. An unbiased estimate of the variance of $\bar{d}_+$ can be obtained by adjusting the squared sample residuals by $(1 - \beta_i)^{-1}$. When the weights are equal, i.e. $\beta_i = \beta$ for all $i = 1, ..., k$, the adjustment factor is $(k - 1)/k$.

Proof:

This follows immediately from Remark 2.

APPENDIX B

Chapter 3

B.1. Supplementary simulation results

This sections includes extra simulation results that examine the bias of the variance estimators used in Chapter 3 under different levels of heterogeneity, number of studies, and configuration of sample sizes of primary studies.

Table B.1. Ratio of Model Variance to Empirical Variance for the Risk Difference: Equal Size Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Equal size	0.00	2	1.47	1.00	1.47	1.00	1.99	1.51	1.02	1.51	1.02	2.04	1.45	0.97	1.45	0.97	1.94
		4	1.31	1.00	1.31	1.00	1.33	1.29	0.99	1.29	0.99	1.32	1.31	0.99	1.31	0.99	1.33
		6	1.25	1.01	1.25	1.01	1.21	1.22	0.98	1.22	0.98	1.17	1.23	1.00	1.23	1.00	1.20
		8	1.23	1.02	1.23	1.02	1.16	1.18	0.98	1.18	0.98	1.12	1.20	1.00	1.20	1.00	1.14
		10	1.19	1.00	1.19	1.00	1.12	1.18	1.00	1.18	1.00	1.11	1.22	1.02	1.21	1.02	1.14
	0.33	2	1.24	0.97	1.24	0.97	1.95	1.25	0.97	1.25	0.97	1.95	1.27	0.99	1.27	0.99	1.98
		4	1.13	1.00	1.13	1.00	1.33	1.14	1.00	1.14	1.00	1.33	1.12	0.99	1.12	0.99	1.32
		6	1.07	0.99	1.07	0.99	1.19	1.08	0.99	1.08	0.99	1.19	1.08	1.00	1.08	1.00	1.20
		8	1.08	1.02	1.08	1.02	1.16	1.06	1.00	1.06	1.00	1.15	1.06	1.00	1.06	1.00	1.15
		10	1.07	1.03	1.07	1.02	1.14	1.02	0.98	1.02	0.98	1.09	1.06	1.01	1.05	1.01	1.12
	0.50	2	1.17	0.99	1.17	0.99	1.98	1.18	1.00	1.18	1.00	2.00	1.14	0.96	1.14	0.96	1.91
		4	1.11	1.03	1.11	1.03	1.38	1.09	1.01	1.08	1.01	1.35	1.06	0.99	1.06	0.99	1.32
		6	1.03	0.99	1.03	0.99	1.19	1.02	0.98	1.02	0.98	1.18	1.03	0.99	1.03	0.99	1.19
		8	1.03	1.01	1.03	1.00	1.15	1.02	0.99	1.02	0.99	1.14	1.02	0.99	1.02	0.99	1.13
		10	1.04	1.02	1.04	1.02	1.13	1.01	1.00	1.01	1.00	1.11	0.99	0.97	0.98	0.97	1.08

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.2. Ratio of Model Variance to Empirical Variance for the Risk Difference: One Small Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One small	0.00	2	1.67	1.38	1.67	1.38	4.60	1.67	1.39	1.67	1.39	4.52	1.66	1.36	1.66	1.36	4.48
		4	1.36	1.16	1.48	1.03	1.46	1.33	1.14	1.45	1.00	1.43	1.34	1.15	1.47	1.01	1.43
		6	1.26	1.05	1.30	0.99	1.22	1.25	1.05	1.29	0.99	1.22	1.27	1.07	1.31	1.00	1.24
		8	1.25	1.06	1.26	1.02	1.19	1.22	1.03	1.24	0.99	1.15	1.22	1.03	1.24	1.00	1.15
		10	1.19	1.02	1.20	1.00	1.12	1.16	0.99	1.17	0.97	1.08	1.20	1.03	1.22	1.01	1.13
	0.33	2	1.37	1.18	1.37	1.18	3.63	1.33	1.15	1.33	1.15	3.51	1.35	1.15	1.35	1.15	3.55
		4	1.11	1.04	1.18	0.98	1.39	1.12	1.05	1.19	0.99	1.40	1.13	1.05	1.19	1.00	1.41
		6	1.09	1.03	1.12	1.01	1.24	1.11	1.05	1.14	1.02	1.26	1.12	1.06	1.15	1.03	1.26
		8	1.05	1.00	1.06	0.99	1.15	1.06	1.02	1.08	1.00	1.16	1.05	1.00	1.06	0.98	1.14
		10	1.06	1.02	1.07	1.01	1.13	1.05	1.02	1.06	1.00	1.12	1.06	1.03	1.07	1.02	1.14
	0.50	2	1.22	1.07	1.22	1.07	3.15	1.21	1.07	1.21	1.07	3.11	1.23	1.08	1.23	1.08	3.15
		4	1.04	1.01	1.09	0.97	1.37	1.05	1.02	1.10	0.98	1.38	1.06	1.02	1.10	0.99	1.39
		6	1.04	1.02	1.06	1.01	1.23	1.04	1.02	1.06	1.00	1.23	1.04	1.03	1.07	1.01	1.23
		8	1.00	0.98	1.01	0.97	1.13	1.03	1.01	1.04	1.01	1.16	1.03	1.01	1.04	1.01	1.17
		10	0.99	0.98	1.00	0.98	1.09	0.99	0.98	1.00	0.98	1.10	1.01	1.00	1.02	1.00	1.12

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.3. Ratio of Model Variance to Empirical Variance for the Risk Difference: Half-half Large and Small Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Half-half	0.00	2	1.63	1.34	1.63	1.34	4.49	1.64	1.34	1.64	1.34	4.50	1.63	1.34	1.63	1.34	4.42
		4	1.36	1.25	1.58	1.04	1.63	1.38	1.29	1.63	1.09	1.69	1.39	1.32	1.66	1.09	1.69
		6	1.29	1.18	1.46	1.01	1.33	1.30	1.21	1.47	1.03	1.37	1.32	1.22	1.49	1.03	1.36
		8	1.30	1.17	1.42	1.02	1.26	1.32	1.20	1.44	1.04	1.28	1.27	1.16	1.39	1.00	1.23
		10	1.27	1.15	1.37	1.02	1.20	1.27	1.17	1.38	1.02	1.20	1.26	1.15	1.36	1.02	1.20
	0.33	2	1.38	1.18	1.38	1.18	3.67	1.35	1.16	1.35	1.16	3.56	1.36	1.16	1.36	1.16	3.59
		4	1.07	1.04	1.18	0.96	1.46	1.04	1.03	1.18	0.94	1.43	1.05	1.04	1.19	0.95	1.44
		6	1.06	1.04	1.14	0.99	1.29	1.03	1.03	1.13	0.97	1.26	1.07	1.06	1.16	1.00	1.30
		8	1.00	0.99	1.06	0.95	1.15	1.01	1.00	1.07	0.96	1.17	1.02	1.02	1.09	0.97	1.18
		10	0.99	0.98	1.04	0.95	1.11	1.00	1.00	1.06	0.96	1.13	1.00	1.00	1.06	0.96	1.12
	0.50	2	1.24	1.09	1.24	1.09	3.23	1.21	1.07	1.21	1.07	3.11	1.20	1.07	1.20	1.07	3.12
		4	0.99	0.97	1.06	0.92	1.37	1.00	0.99	1.07	0.94	1.40	0.99	0.98	1.07	0.93	1.39
		6	0.98	0.97	1.02	0.94	1.21	0.96	0.96	1.01	0.93	1.20	0.97	0.96	1.01	0.93	1.20
		8	0.97	0.97	1.01	0.96	1.16	0.97	0.97	1.00	0.96	1.15	0.98	0.98	1.02	0.97	1.16
		10	0.96	0.96	0.98	0.95	1.10	0.94	0.95	0.98	0.94	1.08	0.97	0.98	1.00	0.97	1.12

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.4. Ratio of Model Variance to Empirical Variance for the Risk Difference: One Large Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One large	0.00	2	1.63	1.34	1.63	1.34	4.49	1.63	1.34	1.63	1.34	4.37	1.65	1.36	1.65	1.36	4.53
		4	1.38	1.22	1.47	1.13	2.22	1.37	1.23	1.46	1.14	2.19	1.38	1.23	1.47	1.14	2.23
		6	1.26	1.11	1.32	1.04	1.67	1.30	1.17	1.37	1.09	1.71	1.25	1.11	1.32	1.03	1.63
		8	1.26	1.13	1.32	1.07	1.52	1.21	1.08	1.26	1.02	1.46	1.25	1.13	1.32	1.06	1.49
		10	1.21	1.08	1.25	1.03	1.35	1.19	1.08	1.24	1.02	1.34	1.19	1.07	1.23	1.01	1.33
	0.33	2	1.38	1.19	1.38	1.19	3.68	1.37	1.18	1.37	1.18	3.61	1.38	1.18	1.38	1.18	3.65
		4	1.01	0.94	1.05	0.91	1.63	1.02	0.96	1.07	0.92	1.65	1.02	0.95	1.06	0.92	1.63
		6	0.96	0.91	0.99	0.90	1.32	0.96	0.91	0.99	0.90	1.31	0.95	0.91	0.98	0.89	1.31
		8	0.90	0.87	0.92	0.86	1.14	0.96	0.92	0.98	0.92	1.20	0.92	0.89	0.95	0.88	1.17
		10	0.92	0.90	0.94	0.90	1.11	0.92	0.89	0.93	0.89	1.10	0.91	0.88	0.93	0.88	1.09
	0.50	2	1.23	1.09	1.23	1.09	3.21	1.24	1.10	1.24	1.10	3.17	1.19	1.06	1.19	1.06	3.08
		4	0.96	0.92	0.99	0.90	1.50	0.96	0.92	0.99	0.90	1.51	0.94	0.89	0.97	0.88	1.48
		6	0.89	0.87	0.91	0.85	1.18	0.91	0.88	0.93	0.88	1.22	0.90	0.87	0.91	0.87	1.20
		8	0.91	0.90	0.93	0.89	1.11	0.91	0.89	0.92	0.89	1.11	0.91	0.89	0.92	0.89	1.11
		10	0.94	0.93	0.95	0.93	1.09	0.94	0.93	0.95	0.93	1.09	0.92	0.90	0.92	0.91	1.07

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.5. Ratio of Model Variance to Empirical Variance for the Log Risk Ratio: Equal Size Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Equal size	0.00	2	1.46	0.98	1.46	0.98	1.96	1.46	0.99	1.46	0.99	1.99	1.50	1.01	1.50	1.01	2.03
		4	1.34	1.02	1.34	1.02	1.36	1.31	1.00	1.31	1.00	1.33	1.32	1.00	1.32	1.00	1.34
		6	1.24	1.00	1.25	0.99	1.19	1.26	1.01	1.26	1.01	1.21	1.25	1.01	1.25	1.01	1.21
		8	1.21	1.00	1.21	0.99	1.14	1.23	1.02	1.23	1.02	1.16	1.22	1.01	1.22	1.01	1.16
		10	1.20	1.01	1.20	1.00	1.12	1.17	0.99	1.17	0.99	1.10	1.20	1.01	1.20	1.01	1.12
	0.33	2	1.32	1.04	1.32	1.04	2.08	1.27	1.00	1.27	1.00	2.01	1.31	1.04	1.31	1.04	2.07
		4	1.13	1.00	1.13	0.99	1.32	1.13	1.00	1.14	1.00	1.33	1.14	1.01	1.14	1.01	1.35
		6	1.12	1.04	1.13	1.03	1.24	1.10	1.01	1.10	1.01	1.21	1.09	1.01	1.09	1.01	1.21
		8	1.08	1.02	1.08	1.01	1.16	1.04	0.98	1.05	0.98	1.12	1.06	1.00	1.06	0.99	1.14
		10	1.04	1.00	1.04	0.99	1.10	1.04	0.99	1.04	0.99	1.10	1.05	1.00	1.05	1.00	1.11
	0.50	2	1.19	1.01	1.19	1.01	2.03	1.18	1.01	1.18	1.01	2.02	1.21	1.03	1.21	1.03	2.06
		4	1.07	1.01	1.08	1.00	1.34	1.08	1.01	1.08	1.01	1.35	1.07	1.01	1.08	1.00	1.34
		6	1.03	1.00	1.03	0.99	1.19	1.04	1.00	1.04	1.00	1.20	1.06	1.02	1.06	1.02	1.23
		8	1.04	1.03	1.05	1.02	1.17	1.03	1.01	1.04	1.01	1.16	1.04	1.02	1.05	1.02	1.17
		10	1.02	1.02	1.03	1.01	1.12	1.03	1.02	1.04	1.02	1.13	1.03	1.02	1.04	1.02	1.13

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.6. Ratio of Model Variance to Empirical Variance for the Log Risk Ratio: One Small Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One small	0.00	2	1.65	1.36	1.65	1.36	4.53	1.63	1.35	1.63	1.35	4.43	1.67	1.39	1.67	1.39	4.54
		4	1.34	1.11	1.43	0.99	1.42	1.38	1.19	1.51	1.04	1.48	1.37	1.17	1.50	1.02	1.45
		6	1.27	1.06	1.31	1.00	1.24	1.26	1.06	1.30	0.99	1.22	1.25	1.05	1.29	0.98	1.21
		8	1.21	1.01	1.22	0.98	1.13	1.23	1.04	1.25	1.00	1.16	1.21	1.02	1.23	0.98	1.14
		10	1.22	1.03	1.23	1.01	1.13	1.17	1.00	1.18	0.98	1.10	1.17	1.00	1.18	0.98	1.09
	0.33	2	1.38	1.18	1.38	1.18	3.68	1.40	1.20	1.40	1.20	3.69	1.40	1.21	1.40	1.21	3.74
		4	1.11	1.02	1.17	0.97	1.37	1.13	1.06	1.20	1.00	1.41	1.14	1.07	1.21	1.01	1.42
		6	1.09	1.02	1.11	1.00	1.23	1.09	1.03	1.12	1.00	1.23	1.10	1.04	1.13	1.02	1.25
		8	1.05	1.00	1.06	0.99	1.14	1.07	1.02	1.08	1.01	1.16	1.07	1.03	1.09	1.01	1.17
		10	1.06	1.03	1.07	1.02	1.14	1.04	1.00	1.05	0.99	1.11	1.07	1.03	1.08	1.02	1.15
	0.50	2	1.26	1.12	1.26	1.12	3.30	1.20	1.06	1.20	1.06	3.12	1.23	1.09	1.23	1.09	3.18
		4	1.06	1.01	1.09	0.98	1.38	1.03	1.00	1.07	0.96	1.35	1.02	0.99	1.06	0.95	1.34
		6	1.03	1.00	1.04	0.99	1.22	1.01	0.99	1.03	0.98	1.20	1.03	1.01	1.05	0.99	1.22
		8	1.03	1.01	1.04	1.01	1.17	1.00	0.99	1.01	0.98	1.13	1.04	1.03	1.06	1.02	1.18
		10	1.02	1.01	1.03	1.01	1.13	1.02	1.02	1.03	1.01	1.13	1.02	1.01	1.03	1.01	1.13

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.7. Ratio of Model Variance to Empirical Variance for the Log Risk Ratio: Half-half Large and Small Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Half-half	0.00	2	1.67	1.38	1.67	1.38	4.63	1.68	1.39	1.68	1.39	4.61	1.69	1.39	1.69	1.39	4.63
		4	1.42	1.29	1.66	1.07	1.68	1.35	1.27	1.61	1.04	1.62	1.38	1.30	1.66	1.07	1.66
		6	1.33	1.19	1.49	1.01	1.35	1.31	1.20	1.47	1.02	1.35	1.28	1.16	1.43	0.98	1.31
		8	1.27	1.14	1.38	0.99	1.22	1.29	1.17	1.41	1.02	1.26	1.27	1.15	1.40	1.00	1.23
		10	1.28	1.15	1.37	1.02	1.21	1.28	1.16	1.38	1.03	1.21	1.26	1.15	1.36	1.01	1.20
	0.33	2	1.42	1.23	1.42	1.23	3.80	1.35	1.16	1.35	1.16	3.61	1.37	1.18	1.37	1.18	3.64
		4	1.10	1.06	1.22	0.97	1.49	1.09	1.08	1.24	0.99	1.50	1.07	1.04	1.20	0.96	1.46
		6	1.07	1.04	1.14	0.98	1.29	1.05	1.03	1.13	0.97	1.27	1.04	1.01	1.12	0.96	1.25
		8	1.02	1.00	1.08	0.97	1.18	1.03	1.02	1.10	0.98	1.19	1.02	1.02	1.09	0.97	1.18
		10	1.02	1.00	1.06	0.98	1.15	1.03	1.03	1.09	1.00	1.16	0.99	0.99	1.04	0.95	1.11
	0.50	2	1.25	1.11	1.25	1.11	3.26	1.22	1.08	1.22	1.08	3.16	1.23	1.09	1.23	1.09	3.19
		4	1.03	0.99	1.09	0.94	1.41	1.02	1.02	1.12	0.96	1.42	0.99	0.97	1.06	0.92	1.37
		6	1.00	0.99	1.05	0.96	1.24	0.98	0.98	1.03	0.94	1.21	0.96	0.95	1.00	0.93	1.19
		8	0.97	0.96	0.99	0.94	1.14	0.99	0.99	1.03	0.97	1.17	0.95	0.96	0.99	0.94	1.13
		10	0.98	0.98	1.00	0.97	1.13	0.96	0.96	0.98	0.95	1.10	0.98	0.98	1.00	0.97	1.12

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.8. Ratio of Model Variance to Empirical Variance for the Log Risk Ratio: One Large Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One large	0.00	2	1.69	1.39	1.69	1.39	4.69	1.64	1.35	1.64	1.35	4.48	1.65	1.36	1.65	1.36	4.52
		4	1.38	1.23	1.48	1.13	2.24	1.37	1.22	1.45	1.13	2.21	1.35	1.19	1.43	1.11	2.17
		6	1.26	1.13	1.34	1.04	1.67	1.27	1.14	1.35	1.06	1.68	1.31	1.17	1.38	1.08	1.71
		8	1.25	1.12	1.32	1.04	1.49	1.25	1.12	1.31	1.05	1.48	1.24	1.11	1.30	1.04	1.47
		10	1.22	1.10	1.28	1.03	1.36	1.19	1.07	1.24	1.01	1.33	1.19	1.08	1.24	1.02	1.34
	0.33	2	1.38	1.19	1.38	1.19	3.70	1.38	1.18	1.38	1.18	3.67	1.38	1.18	1.38	1.18	3.66
		4	1.03	0.96	1.08	0.92	1.66	1.06	1.00	1.12	0.96	1.72	1.03	0.96	1.08	0.93	1.67
		6	0.96	0.91	1.00	0.89	1.32	0.94	0.90	0.97	0.88	1.30	0.96	0.92	1.00	0.90	1.32
		8	0.95	0.91	0.98	0.90	1.19	0.95	0.92	0.98	0.91	1.19	0.93	0.90	0.96	0.89	1.18
		10	0.95	0.93	0.98	0.92	1.14	0.94	0.91	0.96	0.92	1.13	0.93	0.90	0.95	0.90	1.10
	0.50	2	1.23	1.08	1.23	1.08	3.19	1.26	1.12	1.26	1.12	3.24	1.24	1.10	1.24	1.10	3.20
		4	0.97	0.94	1.01	0.90	1.52	0.94	0.89	0.96	0.87	1.47	0.96	0.93	1.00	0.90	1.51
		6	0.91	0.89	0.93	0.87	1.21	0.92	0.90	0.95	0.89	1.22	0.91	0.89	0.93	0.87	1.20
		8	0.89	0.88	0.91	0.87	1.09	0.92	0.90	0.93	0.89	1.11	0.89	0.87	0.90	0.86	1.07
		10	0.93	0.92	0.94	0.91	1.07	0.94	0.93	0.95	0.93	1.09	0.93	0.92	0.94	0.91	1.07

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.9. Ratio of Model Variance to Empirical Variance for the Log Odds Ratio: Equal Size Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Equal size	0.00	2	1.49	1.00	1.49	1.00	2.00	1.47	0.99	1.47	0.99	1.97	1.45	0.97	1.45	0.97	1.94
		4	1.31	1.00	1.31	1.00	1.33	1.32	1.01	1.32	1.01	1.35	1.34	1.02	1.34	1.02	1.36
		6	1.24	1.00	1.25	0.99	1.19	1.26	1.01	1.26	1.01	1.21	1.24	0.99	1.24	0.99	1.19
		8	1.21	1.00	1.21	1.00	1.14	1.22	1.01	1.22	1.01	1.15	1.20	0.99	1.20	0.99	1.13
		10	1.20	1.02	1.20	1.01	1.12	1.19	1.00	1.19	1.00	1.11	1.19	1.00	1.19	1.00	1.11
	0.33	2	1.29	1.01	1.29	1.01	2.03	1.29	1.02	1.28	1.02	2.04	1.27	1.00	1.27	1.00	1.99
		4	1.13	1.00	1.14	0.99	1.33	1.11	0.98	1.11	0.98	1.30	1.16	1.02	1.16	1.02	1.36
		6	1.10	1.01	1.10	1.01	1.21	1.09	1.01	1.10	1.01	1.21	1.13	1.04	1.13	1.04	1.25
		8	1.05	1.00	1.06	0.99	1.13	1.07	1.01	1.07	1.01	1.15	1.08	1.02	1.08	1.01	1.16
		10	1.06	1.02	1.07	1.01	1.13	1.06	1.01	1.06	1.01	1.12	1.07	1.02	1.07	1.02	1.14
	0.50	2	1.17	0.99	1.17	0.99	1.98	1.17	0.99	1.17	0.99	1.99	1.18	1.00	1.18	1.00	1.99
		4	1.06	0.99	1.07	0.99	1.32	1.09	1.01	1.09	1.01	1.35	1.10	1.03	1.10	1.03	1.37
		6	1.05	1.02	1.06	1.01	1.22	1.06	1.02	1.06	1.02	1.23	1.02	0.99	1.02	0.99	1.18
		8	1.02	1.01	1.03	1.00	1.15	1.00	0.98	1.00	0.98	1.12	0.99	0.97	0.99	0.97	1.11
		10	1.00	0.99	1.00	0.98	1.09	1.01	1.00	1.01	1.00	1.11	1.03	1.02	1.03	1.01	1.13

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.10. Ratio of Model Variance to Empirical Variance for the Log Odds Ratio: One Small Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One small	0.00	2	1.68	1.39	1.68	1.39	4.61	1.70	1.40	1.70	1.40	4.60	1.67	1.38	1.67	1.38	4.53
		4	1.34	1.12	1.45	0.99	1.42	1.35	1.15	1.47	1.01	1.44	1.34	1.14	1.46	1.00	1.42
		6	1.28	1.07	1.32	1.01	1.24	1.24	1.04	1.28	0.97	1.20	1.28	1.07	1.32	1.01	1.24
		8	1.21	1.02	1.23	0.99	1.15	1.20	1.02	1.22	0.98	1.14	1.20	1.01	1.21	0.98	1.13
		10	1.19	1.01	1.19	0.99	1.11	1.22	1.04	1.23	1.02	1.14	1.21	1.04	1.22	1.01	1.14
	0.33	2	1.37	1.18	1.37	1.18	3.67	1.38	1.19	1.38	1.19	3.64	1.36	1.17	1.36	1.17	3.62
		4	1.13	1.05	1.19	0.99	1.41	1.15	1.08	1.22	1.01	1.43	1.08	1.03	1.16	0.96	1.35
		6	1.09	1.02	1.12	1.01	1.24	1.06	0.99	1.08	0.97	1.19	1.09	1.03	1.12	1.00	1.22
		8	1.09	1.04	1.11	1.03	1.19	1.05	1.00	1.06	0.98	1.14	1.05	1.00	1.06	0.98	1.14
		10	1.05	1.01	1.06	1.01	1.13	1.07	1.03	1.08	1.02	1.15	1.05	1.02	1.07	1.01	1.13
	0.50	2	1.26	1.12	1.26	1.12	3.25	1.24	1.09	1.24	1.09	3.17	1.24	1.09	1.24	1.09	3.17
		4	1.09	1.05	1.14	1.01	1.42	1.04	1.01	1.08	0.98	1.37	1.07	1.04	1.12	1.00	1.40
		6	1.02	1.00	1.04	0.99	1.21	1.04	1.01	1.05	1.00	1.23	1.03	1.00	1.04	0.99	1.22
		8	1.03	1.02	1.04	1.01	1.17	1.02	1.00	1.03	1.00	1.15	1.00	0.99	1.01	0.98	1.13
		10	1.02	1.01	1.02	1.01	1.13	1.02	1.01	1.03	1.01	1.13	1.02	1.01	1.03	1.01	1.13

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.11. Ratio of Model Variance to Empirical Variance for the Log Odds Ratio: Half-half Large and Small Studies

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
Half-half	0.00	2	1.64	1.35	1.64	1.35	4.50	1.62	1.33	1.62	1.33	4.43	1.67	1.38	1.67	1.38	4.54
		4	1.42	1.30	1.66	1.07	1.67	1.37	1.28	1.63	1.06	1.64	1.38	1.28	1.63	1.07	1.66
		6	1.34	1.21	1.50	1.03	1.37	1.29	1.20	1.47	1.01	1.33	1.35	1.24	1.53	1.05	1.38
		8	1.26	1.13	1.37	0.99	1.22	1.30	1.19	1.43	1.03	1.26	1.29	1.17	1.41	1.02	1.26
		10	1.27	1.14	1.36	1.01	1.20	1.21	1.11	1.31	0.97	1.14	1.28	1.18	1.40	1.03	1.21
	0.33	2	1.38	1.18	1.38	1.18	3.68	1.35	1.16	1.35	1.16	3.60	1.34	1.15	1.34	1.15	3.54
		4	1.11	1.08	1.24	0.99	1.51	1.09	1.07	1.22	0.98	1.50	1.08	1.07	1.22	0.97	1.46
		6	1.08	1.05	1.16	0.98	1.29	1.04	1.03	1.13	0.97	1.26	1.05	1.03	1.13	0.97	1.27
		8	1.03	1.01	1.09	0.97	1.19	1.04	1.04	1.11	0.99	1.20	1.01	0.99	1.06	0.95	1.16
		10	1.00	0.99	1.05	0.96	1.12	1.04	1.04	1.10	1.01	1.17	1.04	1.04	1.09	1.00	1.17
	0.50	2	1.23	1.09	1.23	1.09	3.18	1.23	1.09	1.23	1.09	3.19	1.24	1.10	1.24	1.10	3.19
		4	0.98	0.96	1.06	0.91	1.36	1.02	1.01	1.09	0.95	1.42	1.00	0.99	1.08	0.94	1.39
		6	0.98	0.97	1.02	0.94	1.21	0.97	0.96	1.01	0.94	1.20	0.98	0.98	1.04	0.95	1.22
		8	0.95	0.95	0.98	0.93	1.12	0.94	0.94	0.98	0.93	1.11	0.97	0.98	1.01	0.96	1.15
		10	0.97	0.97	0.99	0.96	1.12	0.97	0.98	1.01	0.97	1.12	0.97	0.97	0.99	0.96	1.11

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

Table B.12. Ratio of Model Variance to Empirical Variance for the Log Odds Ratio: One Large Study

			$(\pi_c, \pi_t) = (0.10, 0.10)$					$(\pi_c, \pi_t) = (0.50, 0.50)$					$(\pi_c, \pi_t) = (0.30, 0.40)$				
Pattern	I^2	k	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3	Z	KnHa	H	HC2	HC3
One large	0.00	2	1.71	1.41	1.71	1.41	4.63	1.69	1.39	1.69	1.39	4.59	1.67	1.38	1.67	1.38	4.56
		4	1.36	1.21	1.46	1.11	2.20	1.39	1.23	1.48	1.14	2.24	1.36	1.20	1.44	1.12	2.21
		6	1.29	1.16	1.37	1.06	1.71	1.31	1.17	1.38	1.09	1.73	1.26	1.12	1.33	1.04	1.66
		8	1.21	1.09	1.28	1.01	1.44	1.24	1.12	1.30	1.04	1.48	1.23	1.11	1.29	1.04	1.46
		10	1.22	1.10	1.27	1.03	1.35	1.22	1.10	1.27	1.04	1.36	1.20	1.08	1.25	1.02	1.35
	0.33	2	1.42	1.22	1.42	1.22	3.78	1.42	1.22	1.42	1.22	3.72	1.40	1.20	1.40	1.20	3.72
		4	1.05	0.99	1.11	0.94	1.69	1.07	1.01	1.13	0.97	1.73	1.03	0.96	1.08	0.92	1.65
		6	0.98	0.93	1.02	0.91	1.34	0.97	0.93	1.01	0.90	1.32	0.94	0.90	0.98	0.88	1.28
		8	0.95	0.92	0.98	0.90	1.19	0.93	0.89	0.95	0.89	1.17	0.94	0.91	0.97	0.90	1.18
		10	0.90	0.88	0.92	0.87	1.08	0.93	0.90	0.95	0.91	1.12	0.93	0.90	0.95	0.90	1.11
	0.50	2	1.22	1.08	1.22	1.08	3.19	1.19	1.05	1.19	1.05	3.09	1.22	1.08	1.22	1.08	3.17
		4	0.96	0.92	1.00	0.89	1.50	0.93	0.89	0.96	0.87	1.46	0.95	0.91	0.98	0.89	1.49
		6	0.88	0.86	0.90	0.85	1.18	0.89	0.87	0.92	0.86	1.19	0.94	0.92	0.96	0.90	1.25
		8	0.92	0.90	0.93	0.89	1.12	0.90	0.89	0.91	0.88	1.11	0.92	0.90	0.93	0.89	1.12
		10	0.93	0.92	0.95	0.91	1.08	0.90	0.89	0.91	0.88	1.04	0.93	0.91	0.93	0.91	1.07

Note:
Ratio of model variance to empirical variance for the following methods: 1. DerSimonian and Laird (Z), 2. Hartung-Knapp (KnHa), 3. Hartung (H), 4. Heteroscedasticity-consistent variance (HC2), and 5. Heteroscedasticity-consistent variance (HC3). Vertical panels correspond to pairs of true underlying probabilities, and columns within panels correspond to the five methods. Horizontal panels correspond to study patterns, and rows within panels correspond to the number of studies for three levels of heterogeneity.

APPENDIX C

Chapter 4

C.1. Supplementary details

The expectation and variance of quadratic forms has been given by Searle (1971). A general result in the context of one-way random-effects models was presented by Rao *et al.* (1981) and is restated here for convenience. The variances of $\hat{\tau}^2_{DL}$ and $\hat{\tau}^2_A$ are found by simple applications of these general results.

Expectation and variance of quadratic forms: A general result

Let the quadratic function Q be

$$Q = \sum_{i=1}^{k} c_i (y_i - \bar{y})^2,$$

where $\bar{y} = \sum_{i=1}^{k} m_i y_i$ and m_i and c_i are positive constants. Let a_i be a vector with element $1 - m_i$ in the i^{th} row and $-m_j$ in the rest, $i \neq j$. Denote the product of $a_i a_i^{'}$ by A and $\sum_{i=1}^{k} c_i A_i$ by B. Let $D = diag\{\sigma_i^2 + \tau^2\}$, i.e. a matrix with $\sigma_i^2 + \tau^2$ in the diagonal. The mean and variance of Q are

$$E(Q) = tr(BD) = \sum_{i=1}^{k} c_i (1 - 2m_i)(\sigma_i^2 + \tau^2) + \left(\sum_{i=1}^{k} c_i \right) \sum_{i=1}^{k} m_i^2 (\sigma_i^2 + \tau^2)$$

and

$$\begin{aligned} Var(Q) &= 2tr[(BD)^2] \\ &= 2 \sum_{i=1}^{k} c_i^2 (1 - 2m_i)^2 (\sigma_i^2 + \tau^2)^2 + 2 \Big(\sum_{i=1}^{k} c_i^2 \Big) \Big[\sum_{i=1}^{k} m_i^2 (\sigma_i^2 + \tau^2) \Big]^2 \\ &+ 4 \Big[\sum_{i=1}^{k} c_i^2 (1 - 2m_i)(\sigma_i^2 + \tau^2) \Big] \Big[\sum_{i=1}^{k} m_i^2 (\sigma_i^2 + \tau^2) \Big] \\ &+ 2 \Big[\Big(\sum_{i=1}^{k} c_i^2 \Big) - \sum_{i=1}^{k} c_i^2 \Big] \Big[\sum_{i=1}^{k} m_i^2 (\sigma_i^2 + \tau^2) \Big]^2 \end{aligned}$$

$$+4\left[\sum_{i=1}^{k} c_i m_i^2(\sigma_i^2+\tau^2)^2\right]\left(\sum_{i=1}^{k} c_i\right)$$
$$+4\left[\sum_{i=1}^{k} c_i m_i(\sigma_i^2+\tau^2)\right]^2 - 8\left[\sum_{i=1}^{k} c_i^2 m_i^2(\sigma_i^2+\tau^2)^2\right]$$
$$-8\left[\sum_{i=1}^{k} m_i^2(\sigma_i^2+\tau^2)\right]\left\{\left[\sum_{i=1}^{k} c_i m_i(\sigma_i^2+\tau^2)\right]\left(\sum_{i=1}^{k} c_i\right) - \left[\sum_{i=1}^{k} c_i^2 m_i(\sigma_i^2+\tau^2)\right]\right\}.$$

Variance of $\hat{\tau}_A^2$

A simple application of the general result with $c_i = 1$ and $m_i = 1/k$ gives

$$\begin{aligned} Var(\hat{\tau}_A^2) &= \frac{2(k-2)}{k(k-1)^2}\sum_{i=1}^{k}(\sigma_i^2+\tau^2)^2 + \frac{2}{k^2(k-1)^2}\left[\sum_{i=1}^{k}(\sigma_i^2+\tau^2)\right]^2 \\ &= \frac{2(k-2)}{k(k-1)^2}\sum_{i=1}^{k}\sigma_i^4 + \frac{2}{k^2(k-1)^2}\left(\sum_{i=1}^{k}\sigma_i^2\right)^2 + \frac{4}{k(k-1)}\tau^2\sum_{i=1}^{k}\sigma_i^2 + \frac{2}{(k-1)}\tau^4. \end{aligned}$$

Variance of $\hat{\tau}_{DL}^2$

The variance of the DerSimonian and Laird estimator can be obtained in a similar way, where $c_i = w_i = 1/\sigma_i^2$ and $m_i = w_i / \sum_{i=1}^{k} w_i$. It follows

$$\begin{aligned} Var(\hat{\tau}_{DL}^2) &= \frac{Var(Q)}{\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}\right)^2} \\ &= \frac{2(k-1) + 4\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}\right)\tau^2 + 2\left(\sum_{i=1}^{k} w_i^2 - 2\frac{\sum_{i=1}^{k} w_i^3}{\sum_{i=1}^{k} w_i} + \frac{(\sum_{i=1}^{k} w_i^2)^2}{(\sum_{i=1}^{k} w_i)^2}\right)\tau^4}{\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}\right)^2}. \end{aligned}$$

C.2. Supplementary simulation results

This section includes extra simulation results for the various scenarios considered. Note that we included a subset of the figures in Chapter 4 to summarize the entirety of the simulation results. The following figures cover a broader range of scenarios. The types of scenarios considered are described in the captions.

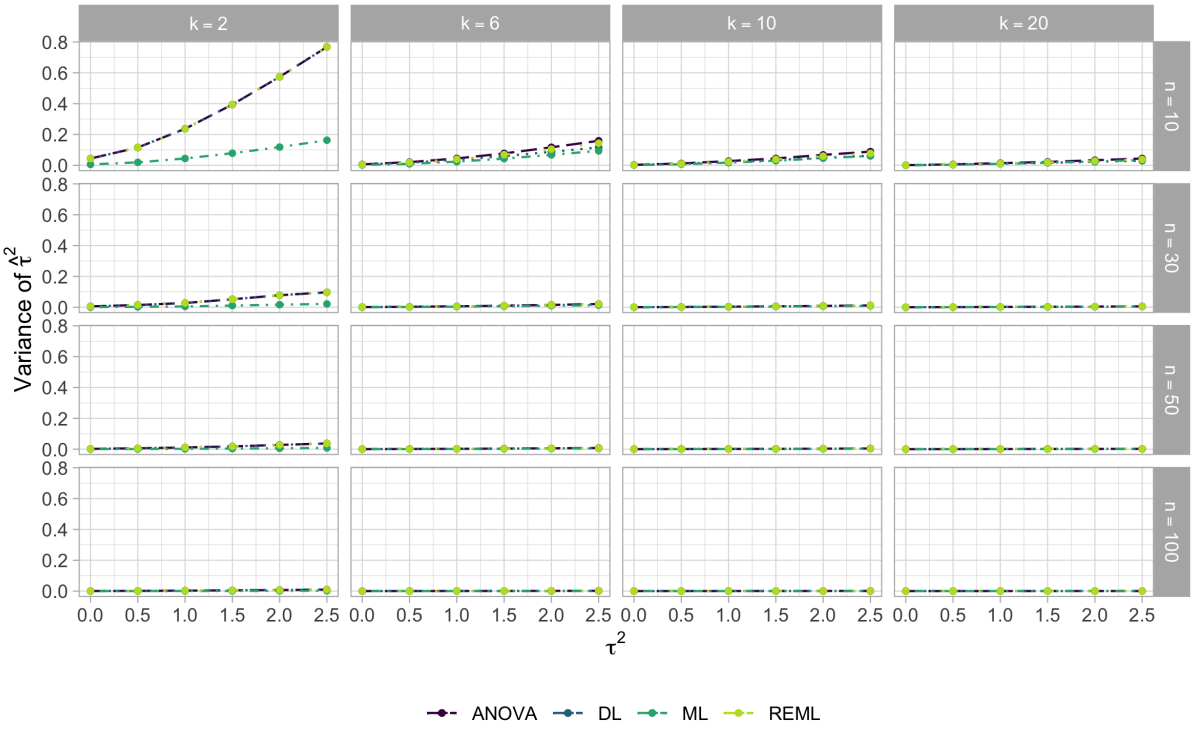

Figure C.1. Variance of $\hat{\tau}^2$ in a Balanced Setting. This plot displays the variance of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. Within-study sample sizes, n, range from 10 to 100 and are equal across studies. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

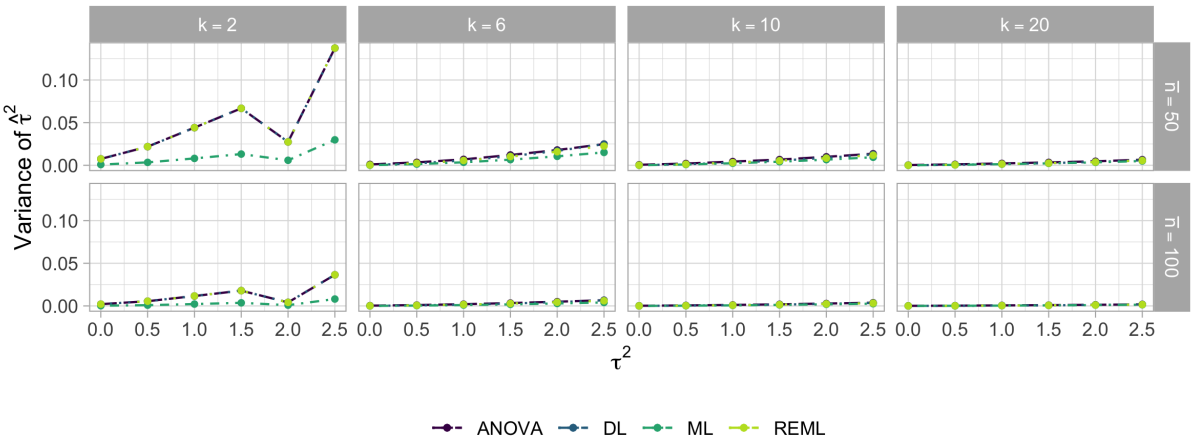

Figure C.2. Variance of $\hat{\tau}^2$ in an Unbalanced Setting: Half-half Large and Small Studies. This plot displays the variance of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

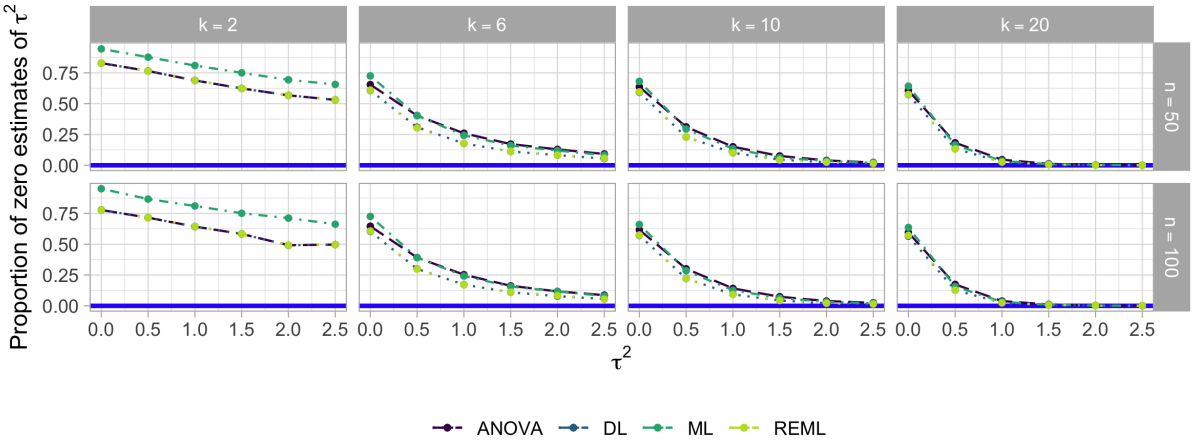

Figure C.3. Proportion of Zero Estimates of τ^2: One Small Study. This plot displays the proportion of zero estimates of τ^2 obtained from four methods: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

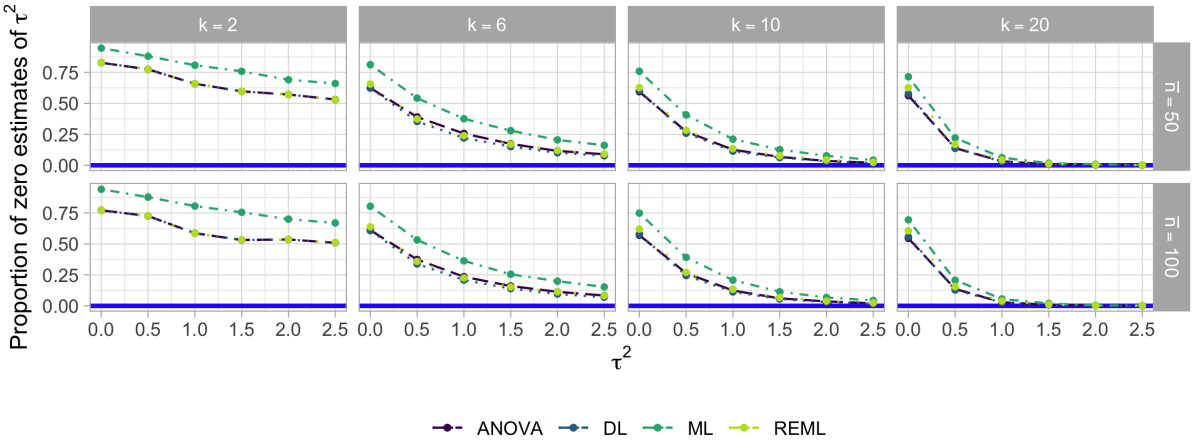

Figure C.4. Proportion of Zero Estimates of τ^2: One Large Study. This plot displays the proportion of zero estimates of τ^2 obtained from four methods: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

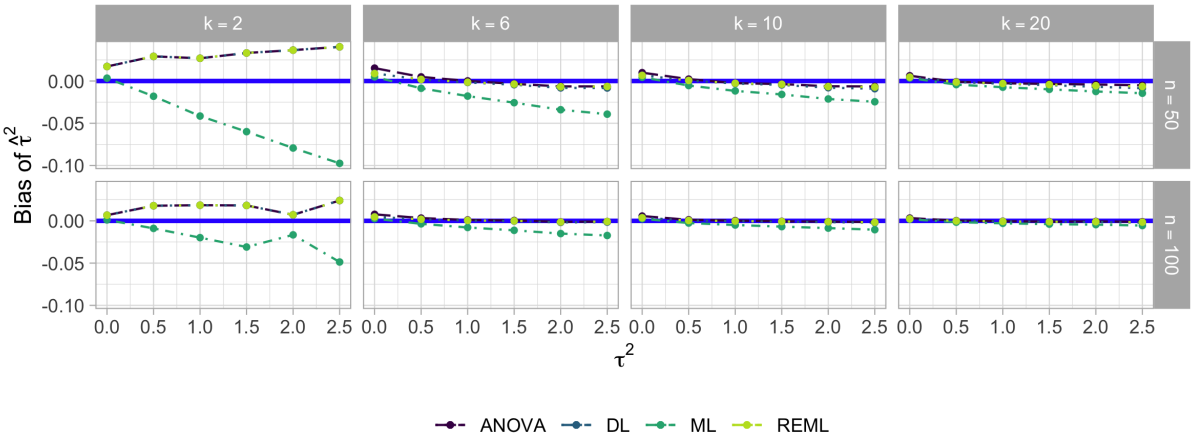

Figure C.5. Absolute bias of $\hat{\tau}^2$: One Small Study. This plot displays the absolute bias of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

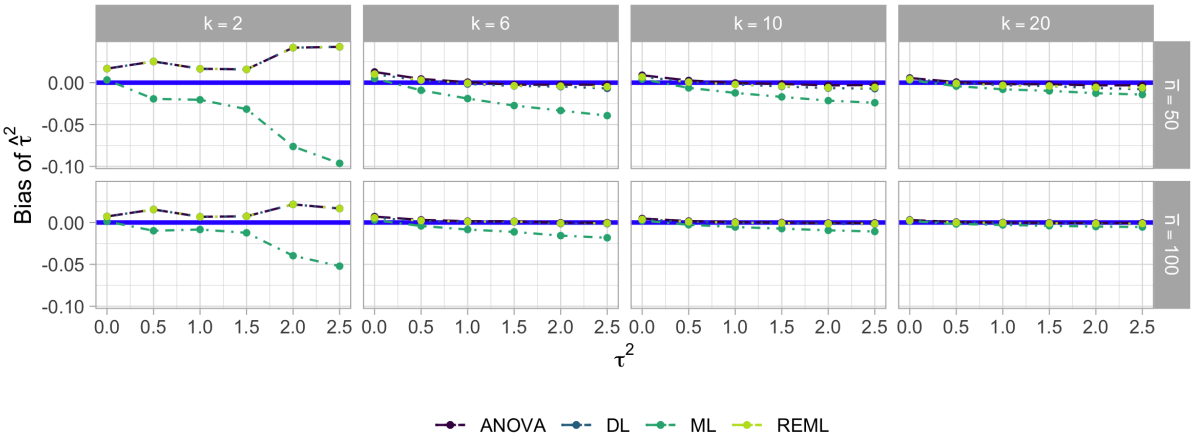

Figure C.6. Absolute bias of $\hat{\tau}^2$: One Large Study. This plot displays the absolute bias of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

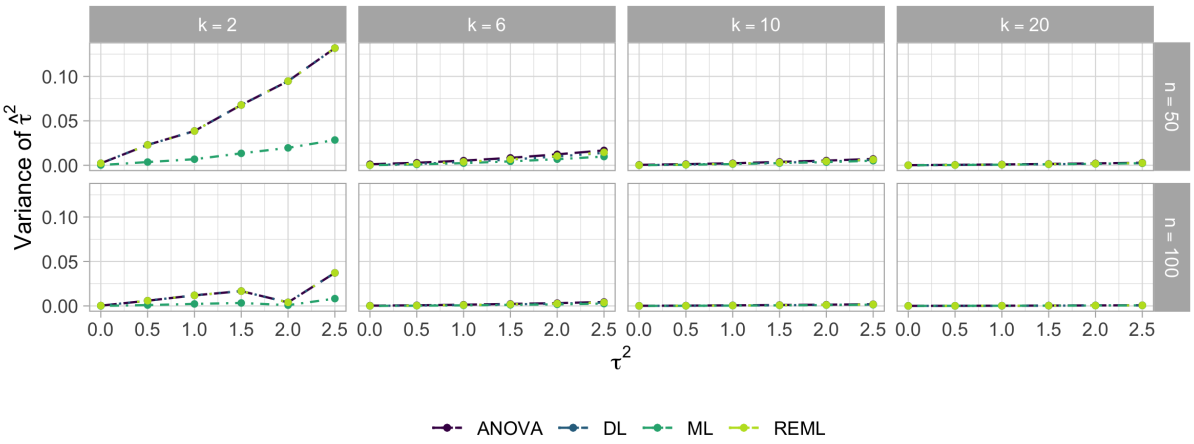

Figure C.7. Variance of $\hat{\tau}^2$: One Small Study. This plot displays the variance of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

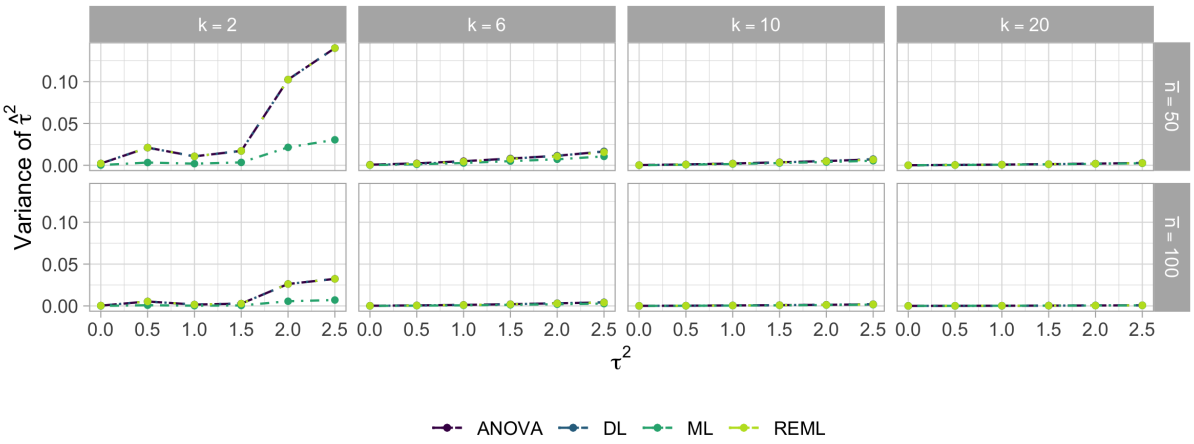

Figure C.8. Variance of $\hat{\tau}^2$: One Large Study. This plot displays the variance of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

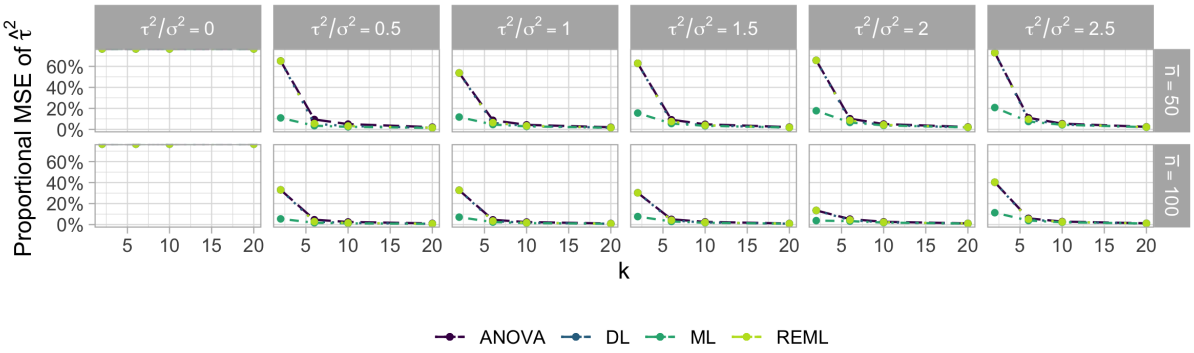

Figure C.9. Proportional mean square error (MSE) of $\hat{\tau}^2$: One Small Study. This plot displays the proportional mean square error of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).

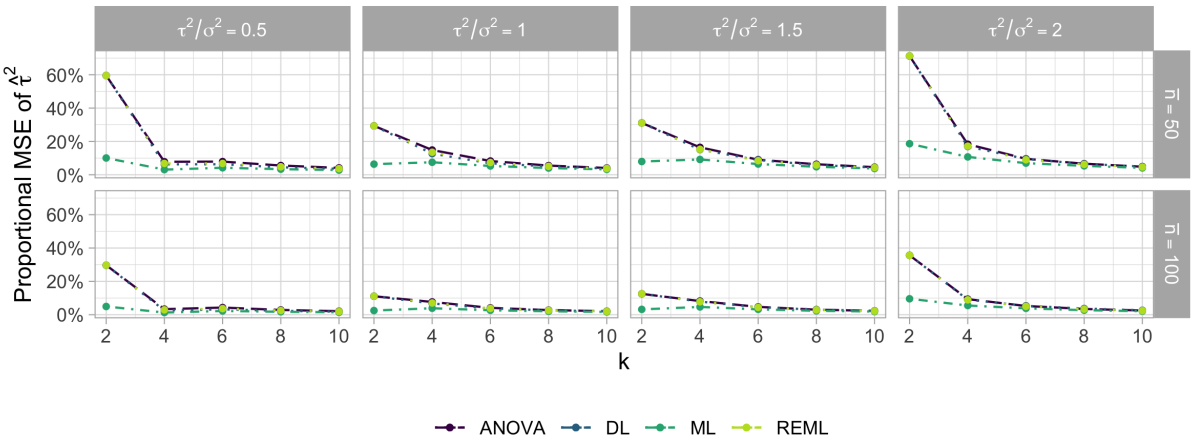

Figure C.10. Proportional mean square error (MSE) of $\hat{\tau}^2$: One Large Study. This plot displays the proportional mean square error of four heterogeneity variance estimators: ANOVA, DL, ML, and REML. The number of studies, k, ranges from 2 to 20, represented by the columns. The average within-study sample sizes, $\bar{n}$, include 50 and 100. The results are shown for different degrees of heterogeneity (τ^2/σ^2 ranges from 0 to 2.5).