January 2009

# Meta-analysis for functions of heterogeneous multivariate effect sizes

Adam Hafdahl
*Washington University in St. Louis*

WASHINGTON UNIVERSITY

Department of Mathematics

# META-ANALYSIS FOR FUNCTIONS OF HETEROGENEOUS MULTIVARIATE EFFECT SIZES

by

Adam Richard Hafdahl

A thesis presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the
degree of Master of Arts in Statistics

December 2009

Saint Louis, Missouri

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

Page

# 1. INTRODUCTION

Meta-analysis is a set of techniques used to compare and combine results from similar studies or, more generally, similar samples. Most applications arise in quantitative research syntheses or systematic reviews, wherein the studies are several reported investigations on a given topic. In that context, it is useful to view a "study" as the collection of all important methodological and other features of a reported investigation apart from the specific random sample of subjects. Although statistical techniques for combining results from related studies have existed for several decades (Abrams & Jones, 1995; Chalmers, Hedges, & Cooper, 2002), such integrative reviews have become especially popular in health-, behavioral-, and social-science disciplines since the late 1970s (Glass, 1976; Schmidt & Hunter, 1977).

Most meta-analytic techniques are statistical procedures, sometimes involving graphical methods. The present thesis concerns situations where the results to be aggregated are summary statistics, such as correlations or certain simple functions of means (e.g., mean differences) or proportions (e.g., risk differences, odds ratios). I refer to these quantities generically as effect sizes (ESs), whereas many authors use "ES" somewhat more narrowly to refer to any statistical index of *bivariate* association, and some reserve it specifically for standardized mean differences (SMDs). Numerous meta-analytic techniques developed for specialized scenarios are beyond the present scope, such as vote-counting procedures, tests of combined significance, and methods for individual patient data, single-subject designs, genetic linkage studies, or survival data.

Also, I focus mainly on frequentist methods, with only occasional comments on Bayesian or empirical Bayes approaches.

It is worth noting here two major distinctions among meta-analytic techniques and commenting on their attendant meta-analytic aims; these are made more precise in Chapter 2. First, *univariate* meta-analytic models and methods are intended for a scalar ES, whereas their *multivariate* counterparts are meant for a vector-valued ES (Becker, 2000; Nam, Mengersen, & Garthwaite, 2003). For instance, a research synthesist might be interested in the Pearson product-moment correlation between two continuous variables (a univariate situation) or in two or more such correlations involving three or more variables (a multivariate situation). In the multivariate case, each study may contribute estimates of all ES components or only a strict subset. I focus herein on the former, complete-data multivariate case but comment on missing-data considerations.

Second, *fixed-effects* models and methods treat the studies at hand (and their ES parameters) as fixed and support generalization to other studies with the same features but different random subjects, whereas their *random-effects* counterparts treat the studies as random and support generalization to a larger universe of studies from which those at hand were sampled (Hedges & Vevea, 1998). Compared to fixed-effects models, random-effects models include an additional error term that essentially reflects the collective effect of all study-level features that influence a study's ES parameter. In this thesis I mainly consider random-effects models, for which meta-analysts are often interested in estimates and inferences about two (hyper)parameters: the mean and (co)variance (matrix) of ES parameters across studies. I also address fixed-effects

approaches that assume all studies share the same ES parameter (i.e., homogeneity) and whose main object of estimation and inference is this common ES. Applications of fixed-effects analyses often entail a test of homogeneity. Some fixed-effects methods permit heterogeneity among the fixed but unknown ES parameters and focus on estimation of or inferences on their mean. These heterogeneous fixed-effects models are beyond the present scope, as are meta-analytic models with study-level covariates (i.e., moderators) and various tasks of interest under heterogeneity, such as estimating or making inferences about a specific study's ES parameter or predicting a future study's ES parameter.

More specifically, this thesis concerns a problem encountered increasingly as research synthesists address more sophisticated research questions and attempt to convey them to wider audiences: How can one obtain valid estimates of and inferences about some function of an ES instead of the ES itself? Such a situation arises when meta-analytic procedures are most appropriate for available data in a particular form but the meta-analyst is interested in some transformation of that form. For example, some meta-analytic procedures perform better with a variance-stabilizing transformation of the focal ES, but for practical purposes the meta-analyst may wish to express certain estimates or other results in terms of the focal ES or a more familiar or practically informative metric. As a multivariate example, consider a research synthesis whose authors collect from each of several studies an estimate of at least one of the correlations among the three variables $Y_1$, $Y_2$, and $Y_3$ (or variables similar to these 3): They believe some of these correlations are heterogeneous and opt for a random-effects analysis, and their primary substantive

interest is in the between-studies mean and variance of the squared multiple correlation for predicting $Y_1$ from both $Y_2$ and $Y_3$. Obtaining this squared multiple correlation from every study and meta-analyzing these directly may be neither feasible nor prudent. Some examples given later involve other ES metrics (e.g., SMDs, proportions).

The major aims of the present project are to develop and evaluate meta-analytic techniques for estimation of and inferences about a wide variety of functions of heterogeneous multivariate ESs. Certain special cases of these techniques are also addressed, such as for fixed-effects models. The remainder of this thesis constitutes four chapters that cover the project's major components. Chapter 2 gives more detail on background and motivation for the problem, including notation for standard models and procedures and brief remarks on relevant literature, as well as discussion of various functions of potential interest. In Chapter 3 I propose techniques for estimation and inference, with an emphasis on the multivariate random-effects case for generic ESs; I also comment on special cases involving particular ES metrics and the univariate and fixed-effects simplifications. In Chapter 4 I report on two Monte Carlo investigations of the proposed techniques' performance under realistic conditions. Finally, in Chapter 5 I review the project's contributions and comment on limitations and directions for continued work on this and related meta-analytic problems.

## 2. BACKGROUND AND MOTIVATION

In the interest of broad applicability to various functions of a variety of ESs, the methods I propose in Chapter 3 are based on intermediate results from fairly general meta-analytic procedures used for estimation and inference in several commonly encountered situations. In this chapter I present the multivariate random-effects model underlying these procedures and describe a particular set of techniques for estimation of and inference about key elements of this model, followed by comments on alternative methods and models. Next I describe various functions of ESs that might be of interest in meta-analytic work, with particular attention to specific ES metrics. I then introduce notation for hyperparameters of such functions and comment on meta-analytic tasks to be accomplished by the proposed techniques. Emphasis is on the multivariate random-effects case, but univariate and fixed-effects specializations are also considered.

### 2.1. Multivariate Random-Effects Models and Procedures

In this section I describe a conventional random-effects model for multivariate ESs as well as a particular set of associated estimation and inference procedures for the main meta-analytic tasks. Other approaches relevant to this model are also mentioned, as are alternative random-effects models. Models and methods in this section are not original to the present work.

**2.1.1. Focal model and procedures.** Suppose that for the $i$th independent study we observe the $J$-component ES $\mathbf{t}_i = [t_{i1}, t_{i2}, ..., t_{iJ}]^T$, $i = 1, 2, ..., I$, as a realization of $\mathbf{T}_i = [T_{i1}, T_{i2}, ..., T_{iJ}]^T$, and we view $\mathbf{T}_i$ as an estimator of study $i$'s ES parameter $\boldsymbol{\theta}_i = [\theta_{i1}, \theta_{i2}, ..., \theta_{iJ}]^T$, which is in turn a realization of $\boldsymbol{\Theta}_i = [\Theta_{i1}, \Theta_{i2}, ..., \Theta_{iJ}]^T$. Both $\mathbf{T}_i$ and $\boldsymbol{\Theta}_i$ are

random column vectors, and it is understood that $\Theta_{1j}$, $\Theta_{2j}$, ..., and $\Theta_{Ij}$ are commensurable in some sense, for each $j$. (In an attempt to maintain consistency of notation, I usually use Greek letters for [hyper]parameters and Latin letters for observable or error variables; uppercase for random variables and lowercase for fixed values or realizations of random variables; and boldface and roman type for vectors and scalars, respectively. For example, $\boldsymbol{\Theta}$, $\Theta_j$, $\boldsymbol{\theta}_i$, $\theta_{ij}$, $\mathbf{T}_i$, $T_{ij}$, $\mathbf{t}_i$, and $t_{ij}$ are distinct objects in the same ES metric. Notable exceptions are my use of uppercase Greek letters for mean vectors and covariance matrices and their elements as well as matrices of known coefficients. I usually reserve italic typeface for functions, index variables, and constants.)

Many developments for multivariate random-effects meta-analysis rely on the following model for such a scenario, which treats observed ESs as independent multivariate-normal observations with known but possibly heterogeneous covariance matrices (Becker, 1992, 1995, 2000; Becker & Schram, 1994; Berkey, Hoaglin, Antczak-Bouckoms, Mosteller, & Colditz, 1998; Kalaian, 1994; Kalaian & Raudenbush, 1996):

$$\mathbf{T}_i = \mathbf{M}_{\boldsymbol{\Theta}} + \mathbf{U}_i + \mathbf{E}_i \,, \tag{1}$$

where $\mathbf{M}_{\boldsymbol{\Theta}} = [M_{\Theta 1}, M_{\Theta 2}, ..., M_{\Theta J}]^T$ is fixed and unknown; the between-studies random error (i.e., random effect) $\mathbf{U}_i = [U_{i1}, U_{i2}, ..., U_{iJ}]^T$ has expectation $E(\mathbf{U}_i) \equiv \mathbf{0}$ and fixed, unknown covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} \equiv \mathrm{Cov}(\mathbf{U}_i)\ \forall\ i$, so that $\mathbf{U}_i \sim (\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}})\ \forall\ i$ (i.e., no specified distribution); the within-study random error $\mathbf{E}_i = [E_{i1}, E_{i2}, ..., E_{iJ}]^T$ has $E(\mathbf{E}_i) \equiv \mathbf{0}$ and fixed, known covariance matrix $\boldsymbol{\Psi}_{\mathbf{T}i} \equiv \mathrm{Cov}(\mathbf{E}_i)$, and in particular $\mathbf{E}_i \sim N_J(\mathbf{0}, \boldsymbol{\Psi}_{\mathbf{T}i})$; and $\mathbf{U}_i$ and $\mathbf{E}_i$ are independent. From Model 1 (i.e., as defined by Equation 1) it follows that

study $i$'s conditional expectation and covariance matrix, given $\mathbf{U}_i = \mathbf{u}_i$, are $\boldsymbol{\theta}_i \equiv \mathrm{E}(\mathbf{T}_i \mid \mathbf{u}_i) = \mathbf{M}_{\Theta} + \mathbf{u}_i$ and $\mathrm{Cov}(\mathbf{T}_i \mid \mathbf{u}_i) = \boldsymbol{\Psi}_{\mathbf{T}i}$, and specifically $\mathbf{T}_i \mid \mathbf{u}_i \sim N_J(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_{\mathbf{T}i})$. Also, the marginal expectation and covariance matrix of $\mathbf{T}_i$ are $\mathrm{E}(\mathbf{T}_i) = \mathbf{M}_{\Theta}$ and $\mathrm{Cov}(\mathbf{T}_i) = \boldsymbol{\Sigma}_{\Theta} + \boldsymbol{\Psi}_{\mathbf{T}i}$. If we assume that $\mathbf{U}_i \sim N_J(\mathbf{0}, \boldsymbol{\Sigma}_{\Theta})\ \forall\ i$—as some estimation techniques do—then $\boldsymbol{\Theta}_i \equiv \mathbf{M}_{\Theta} + \mathbf{U}_i \sim N_J(\mathbf{M}_{\Theta}, \boldsymbol{\Sigma}_{\Theta})\ \forall\ i$, and $\mathbf{T}_i \sim N_J(\mathbf{M}_{\Theta}, \boldsymbol{\Sigma}_{\Theta} + \boldsymbol{\Psi}_{\mathbf{T}i})$. Although in practice $\boldsymbol{\Psi}_{\mathbf{T}i}$ often depends only or largely on known sample size(s), in some situations it depends non-negligibly on the unknown $\boldsymbol{\theta}_i$. I will denote element $j$, $l$ of $\boldsymbol{\Sigma}_{\Theta}$ or $\boldsymbol{\Psi}_{\mathbf{T}i}$ as $\Sigma_{\Theta jl}$ or $\Psi_{\mathbf{T}ijl}$, respectively. Because the $\boldsymbol{\Theta}_i$ are iid in Model 1, I will usually omit the subscript and write $\boldsymbol{\Theta}$.

Some common special cases and extensions of Model 1 are worth noting. When $J = 1$ the model is univariate (i.e., $\mathrm{T}_i = \mathrm{M}_{\Theta} + \mathrm{U}_i + \mathrm{E}_i$). The constraint $\boldsymbol{\Theta}_i = \boldsymbol{\theta}\ \forall\ i$ (equivalently, $\mathbf{U}_i = \mathbf{0}\ \forall\ i$, or $\boldsymbol{\Sigma}_{\Theta} = \mathbf{0}$) yields a homogeneous fixed-effects model (i.e., $\mathbf{T}_i = \boldsymbol{\theta} + \mathbf{E}_i$). A less common heterogeneous fixed-effects model treats $\mathbf{u}_i$—and hence $\boldsymbol{\theta}_i$—as fixed but unknown (i.e., $\mathbf{T}_i = \bar{\boldsymbol{\theta}} + \mathbf{u}_i + \mathbf{E}_i$). In some situations certain "hybrid" models may be appropriate, such as Model 1 with either $\boldsymbol{\Sigma}_{\Theta}$ or $\boldsymbol{\Psi}_{\mathbf{T}i}$ constrained to be diagonal (e.g., when between-studies or within-study correlation between ES components is plausibly zero or implied by the study design). Perhaps the most common extension of Model 1 is to replace $\mathbf{M}_{\Theta}$ by a linear predictor that relates each ES component to one or more study-level covariates.

Becker and Schram (1994) described a two-stage strategy for estimation of and inference about $\mathbf{M}_{\Theta}$ and $\boldsymbol{\Sigma}_{\Theta}$ in Model 1, whereby one first uses an EM algorithm to estimate $\boldsymbol{\Sigma}_{\Theta}$ by maximum likelihood (ML) then treats this estimate as known to obtain a

generalized least-squares (GLS) estimate of $\mathbf{M}_{\boldsymbol{\Theta}}$ and this estimate's covariance matrix. Although their exposition was specific to correlations as ESs, their approach extends readily to generic ESs as in Model 1. Using the assumption that $\mathbf{U}_i \sim N_J(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}) \; \forall \; i$, their EM algorithm proceeds as follows:

1. Choose starting values $\widetilde{\mathbf{M}}_{\boldsymbol{\Theta}}^{(0)}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}^{(0)}$ for estimates of $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$.

2. Estimate each study's ES parameter as $\widetilde{\boldsymbol{\theta}}_i^{(x+1)} = \widetilde{\boldsymbol{\Omega}}_i^{(x+1)} [\, \boldsymbol{\Psi}_{\mathbf{T}i}^{-1} \mathbf{t}_i + (\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}^{(x)})^{-1} \widetilde{\mathbf{M}}_{\boldsymbol{\Theta}}^{(x)} \,]$, where $\widetilde{\boldsymbol{\Omega}}_i^{(x+1)} = [\, \boldsymbol{\Psi}_{\mathbf{T}i}^{-1} + (\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}^{(x)})^{-1}]^{-1}$ has typical element $\widetilde{\Omega}_{ijl}^{(x+1)}$. This expectation step uses the empirical Bayes posterior mean and covariance matrix of $\boldsymbol{\theta}_i$, given the data and current estimates of $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$.

3. Estimate $\mathrm{M}_{\boldsymbol{\Theta}j}$ as $\widetilde{\mathrm{M}}_{\boldsymbol{\Theta}j}^{(x+1)} = I^{-1} \sum_{i=1}^{I} \widetilde{\theta}_{ij}^{(x+1)}$, and estimate $\Sigma_{\boldsymbol{\Theta}jl}$ as $\widetilde{\Sigma}_{\boldsymbol{\Theta}jl}^{(x+1)} = I^{-1} \sum_{i=1}^{I} \left( \widetilde{\Omega}_{ijl}^{(x+1)} + \widetilde{\theta}_{ij}^{(x+1)} \widetilde{\theta}_{il}^{(x+1)} \right) - \widetilde{\mathrm{M}}_{\boldsymbol{\Theta}j}^{(x+1)} \widetilde{\mathrm{M}}_{\boldsymbol{\Theta}l}^{(x+1)}$. This maximization step uses expected sufficient statistics for $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$—based on $\boldsymbol{\theta}_i$—to update estimates of these hyperparameters.

4. Repeat steps 2 and 3 until convergence (e.g., until $\widetilde{\mathbf{M}}_{\boldsymbol{\Theta}}^{(x+1)}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}^{(x+1)}$ do not differ much from $\widetilde{\mathbf{M}}_{\boldsymbol{\Theta}}^{(x)}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}^{(x)}$), and denote the resulting estimators $\widetilde{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$. One may also be interested in an empirical Bayes estimate of each study's ES parameter, $\widetilde{\boldsymbol{\theta}}_i$, and its covariance matrix, $\widetilde{\boldsymbol{\Omega}}_i$.

Now, treating $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ as known, the GLS estimator of $\mathbf{M}_{\boldsymbol{\Theta}}$ and its estimated covariance matrix are given by

$$\hat{\mathbf{M}}_{\boldsymbol{\Theta}} = [\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Xi}_{\mathbf{T}})^{-1}\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Xi}_{\mathbf{T}})^{-1}\mathbf{t} \quad (2)$$

and

$$\hat{\mathrm{Cov}}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}) = [\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Xi}_{\mathbf{T}})^{-1}\mathbf{X}]^{-1}, \quad (3)$$

where $\mathbf{X} = [\mathbf{I}_J, \mathbf{I}_J, ..., \mathbf{I}_J]^{\mathrm{T}}$, an $IJ \times J$ design matrix of stacked identity matrices (because each of $I$ studies contributes a $J$-dimensional ES estimate); $\boldsymbol{\Xi}_{\mathbf{T}}$ is block-diagonal with blocks $(\boldsymbol{\Psi}_{\mathbf{T}1} + \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}})$, $(\boldsymbol{\Psi}_{\mathbf{T}2} + \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}})$, ..., and $(\boldsymbol{\Psi}_{\mathbf{T}I} + \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}})$; and $\mathbf{t} = [\mathbf{t}_1^{\mathrm{T}}, \mathbf{t}_2^{\mathrm{T}}, ..., \mathbf{t}_I^{\mathrm{T}}]^{\mathrm{T}}$.

Taking $\hat{\mathbf{M}}_{\boldsymbol{\Theta}} \sim N_J[\mathbf{M}_{\boldsymbol{\Theta}}, \hat{\mathrm{Cov}}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})]$ to be approximately true, one may use the estimators in 2 and 3 to construct various normal-theory confidence regions involving $\mathbf{M}_{\boldsymbol{\Theta}}$: a confidence interval (CI) for $\mathrm{M}_{\boldsymbol{\Theta}j}$ or a linear combination of $\mathbf{M}_{\boldsymbol{\Theta}}$ components, simultaneous CIs for two or more such components or linear combinations, or a multivariate confidence region for two or more components or linear combinations (Harbord, Deeks, Egger, Whiting, & Sterne, 2007). For instance, when $I$ is not too small an approximate $100(1 - \alpha)\%$ CI for $\mathrm{M}_{\boldsymbol{\Theta}j}$ is $\hat{\mathrm{M}}_{\boldsymbol{\Theta}j} \pm z_{\alpha}\sqrt{\hat{\mathrm{Var}}(\hat{\mathrm{M}}_{\boldsymbol{\Theta}j})}$, where $z_{\alpha} = -\Phi^{-1}(\alpha / 2)$ is a standard-normal quantile and $\hat{\mathrm{Var}}(\hat{\mathrm{M}}_{\boldsymbol{\Theta}j})$ is element $j, j$ of $\hat{\mathrm{Cov}}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$. One may also exploit this approximate normality to test $H_0$: $\mathrm{M}_{\boldsymbol{\Theta}j} = c$ or a general linear hypothesis of the form $H_0$: $\mathbf{L}\mathbf{M}_{\boldsymbol{\Theta}} = \mathbf{c}$ for appropriate $R \times J$ contrast matrix $\mathbf{L}$ and null-hypothetical $\mathbf{c}$. The latter uses the statistic $(\mathbf{L}\hat{\mathbf{M}}_{\boldsymbol{\Theta}} - \mathbf{c})^{\mathrm{T}}[\mathbf{L}\hat{\mathrm{Cov}}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\mathbf{L}^{\mathrm{T}}]^{-1}(\mathbf{L}\hat{\mathbf{M}}_{\boldsymbol{\Theta}} - \mathbf{c})$, which under $H_0$ is distributed approximately as $\chi^2(R - 1)$ for full-rank $\mathbf{L}$.

Becker and Schram (1994) also provided a test of the ES parameters' homogeneity that does not require an assumption about the distributional form of $\mathbf{U}_i$. This uses the heterogeneity statistic

$$Q_{\Theta} = (\mathbf{t} - \mathbf{X}\hat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{\Psi}_{\mathbf{T}}^{-1} (\mathbf{t} - \mathbf{X}\hat{\boldsymbol{\theta}}) , \tag{4}$$

where

$$\hat{\boldsymbol{\theta}} = [\mathbf{X}^{\mathrm{T}} \boldsymbol{\Psi}_{\mathbf{T}}^{-1} \mathbf{X}]^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{\Psi}_{\mathbf{T}}^{-1} \mathbf{t} \tag{5}$$

is a GLS estimator of the common ES $\boldsymbol{\theta}$ in a fixed-effects version of Model 1, and $\boldsymbol{\Psi}_{\mathbf{T}}$ is block-diagonal with blocks $\boldsymbol{\Psi}_{\mathbf{T}1}$, $\boldsymbol{\Psi}_{\mathbf{T}2}$, ..., and $\boldsymbol{\Psi}_{\mathbf{T}I}$. Under $H_0$: $\boldsymbol{\Theta}_i = \boldsymbol{\theta}\ \forall\ i$ (i.e., $\boldsymbol{\Sigma}_{\Theta} = \mathbf{0}$), $Q_{\Theta}$ is distributed approximately as $\chi^2[J(I - 1)]$. One may use similar homogeneity tests for any subset of the ES components or to compare nested fixed- or random-effects models (the latter require attention to the choice of estimator for $\boldsymbol{\Sigma}_{\Theta}$).

Finally, Becker (1992) and Becker and Schram (1994) described modifications of the above estimators and tests to accommodate missing ES estimates. Details of these modified procedures are beyond the present scope. In short, they entail imputing missing ES estimates before estimating $\boldsymbol{\Sigma}_{\Theta}$, deleting relevant rows and columns of all matrices for GLS estimators and the $Q_{\Theta}$ statistic, and adjusting the homogeneity test's degrees of freedom accordingly.

**2.1.2. Specific ES metrics.** Model 1 and the associated procedures described above are applicable to several meta-analytic situations where studies contribute estimates of two or more related ESs. Perhaps the most difficult requirements to satisfy in practice are that the within-study model—$\mathbf{T}_i \mid \mathbf{u}_i \sim N_J(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_{\mathbf{T}i})$ with $\boldsymbol{\Psi}_{\mathbf{T}i}$ known—is

plausible and that adequate data are available to compute $\boldsymbol{\Psi}_{\mathbf{T}i}$, especially when $\boldsymbol{\Psi}_{\mathbf{T}i}$ depends on $\boldsymbol{\theta}_i$. Several authors have described large-sample approximations for a variety of commonly used ES metrics that conform to the within-study model reasonably well under some realistic conditions. By way of illustration, I provide details for correlation matrices in two different metrics. I then comment on some other ES metrics. It is worth bearing in mind that for some of these situations other models or techniques tailored to the particular form of data, justifiable assumptions, or analytic aims at hand may be preferable to the generic procedures described here (e.g., Hamza, van Houwelingen, & Stijnen, 2008).

Suppose the ES parameter $\boldsymbol{\Theta}$ comprises the distinct Pearson product-moment (i.e., Pearson-*r*) correlations between the component variables in $\mathbf{Y} = [Y_1, Y_2, ..., Y_P]$, which I denote $\mathbf{P} = [P_{12}, P_{13}, P_{23}, ... , P_{(P-2)P}, P_{(P-1)P}]^{\mathrm{T}}$—taking the correlation matrix's lower triangle in row-major order. Correlations are popular in observational and quasi-experimental research and are used widely in meta-analysis, especially in studies of validity generalization (Hunter & Schmidt, 2004; Schmidt & Hunter, 1977) and reliability generalization (Mason, Allam, & Brannick, 2007; Vacha-Haase, 1998). Matrices of correlations arise in numerous circumstances where several bivariate associations are of interest, including some special situations such as cross-lagged panel designs and multitrait-multimethod studies of construct validity. If $\mathbf{Y}$ is $P$-variate normal with arbitrary mean and covariance matrix, Olkin and Siotani (1976) showed that $\mathbf{R}_i = [R_{i12}, R_{i13}, R_{i23}, ... , R_{i(P-2)P}, R_{i(P-1)P}]^{\mathrm{T}}$, $\mathbf{P}$'s sample counterpart from a sample of size $n_i$, is approximately $J$-variate normal (where $J = {}_P\mathrm{C}_2$) with the following covariance $\psi_{\mathbf{R}ijl} \equiv$

$\mathrm{Cov}(R_{i\mathrm{VW}}, R_{i\mathrm{XY}})$ between components $j$ ($R_{ij} = R_{i\mathrm{VW}}$, between variables V and W) and $l$ ($R_{il} = R_{i\mathrm{XY}}$, between variables X and Y):

$$\psi_{\mathbf{R}ijl} = \frac{\left[\begin{array}{l}\rho_{i\mathrm{VW}}\rho_{i\mathrm{XY}}\left(\rho_{i\mathrm{VX}}^2 + \rho_{i\mathrm{VY}}^2 + \rho_{i\mathrm{WX}}^2 + \rho_{i\mathrm{WY}}^2\right)/2 + \rho_{i\mathrm{VX}}\rho_{i\mathrm{WY}} + \rho_{i\mathrm{VY}}\rho_{i\mathrm{WX}} - \\ \left(\rho_{i\mathrm{VW}}\rho_{i\mathrm{VX}}\rho_{i\mathrm{VY}} + \rho_{i\mathrm{WV}}\rho_{i\mathrm{WX}}\rho_{i\mathrm{WY}} + \rho_{i\mathrm{XV}}\rho_{i\mathrm{XW}}\rho_{i\mathrm{XY}} + \rho_{i\mathrm{YV}}\rho_{i\mathrm{YW}}\rho_{i\mathrm{YX}}\right)\end{array}\right]}{n_i}, \qquad (6)$$

where $\boldsymbol{\rho}_i = [\rho_{i12}, \rho_{i13}, \rho_{i23}, \ldots, \rho_{i(P-2)P}, \rho_{i(P-1)P}]^{\mathrm{T}}$ is the $i$th study's realization of $\mathbf{P}$. Note that when {V W} = {X Y}, $\psi_{\mathbf{R}ijj} \equiv \mathrm{Var}(R_{i\mathrm{VW}}) = (1-\rho_{i\mathrm{VW}}^2)^2 / n_i$. Becker and Schram's (1994) EM-GLS procedures for $\boldsymbol{\Theta}$ and $\mathbf{T}_i$ may be applied to $\mathbf{P}$ and $\mathbf{R}_i$, respectively, to estimate and make inferences about $\mathbf{M}_{\mathbf{P}}$ and $\boldsymbol{\Sigma}_{\mathbf{P}}$ (as well as $\boldsymbol{\rho}_i$).

Some meta-analysts prefer to analyze Fisher-$z$ correlations: $\mathbf{Z} = [Z_{12}, Z_{13}, Z_{23}, \ldots, Z_{(P-2)P}, Z_{(P-1)P}]^{\mathrm{T}}$ and $\mathbf{Z}_i = [Z_{i12}, Z_{i13}, Z_{i23}, \ldots, Z_{i(P-2)P}, Z_{i(P-1)P}]^{\mathrm{T}}$, where $Z_j = \tanh^{-1} P_j$ and $Z_{ij} = \tanh^{-1} R_{ij}$, instead of $\mathbf{P}$ and $\mathbf{R}_i$. Steiger (1980) showed that under the same conditions as for Equation 6 $\mathbf{Z}_i$ is approximately $J$-variate normal with the following covariance between components $j$ and $l$:

$$\psi_{\mathbf{Z}ijl} \equiv \mathrm{Cov}(Z_{i\mathrm{VW}}, Z_{i\mathrm{XY}}) = \frac{n_i \mathrm{Cov}(R_{i\mathrm{VW}}, R_{i\mathrm{XY}})}{(n_i - 3)(1-\rho_{i\mathrm{VW}}^2)(1-\rho_{i\mathrm{XY}}^2)}. \qquad (7)$$

When {V W} = {X Y}, $\psi_{\mathbf{Z}ijj} = \mathrm{Var}(Z_{i\mathrm{VW}}) = 1 / (n_i - 3)$. Now $\mathbf{Z}$ and $\mathbf{Z}_i$ take the place of $\boldsymbol{\Theta}$ and $\mathbf{T}_i$ in Becker and Schram's (1994) EM-GLS procedures for $\mathbf{M}_{\mathbf{Z}}$ and $\boldsymbol{\Sigma}_{\mathbf{Z}}$ (as well as $\boldsymbol{\zeta}_i$).

In practice, computing $\boldsymbol{\Psi}_{\mathbf{R}i}$ or $\boldsymbol{\Psi}_{\mathbf{Z}i}$ requires substituting a known value for $\boldsymbol{\rho}_i$. A conventional strategy is to substitute $\mathbf{r}_i$ for $\boldsymbol{\rho}_i$ (Becker & Schram, 1994). Hafdahl (2004) showed that this strategy performs poorly in several respects, especially when the studies' average $n_i$ is not large, and recommended replacing $\boldsymbol{\rho}_i$ by either a simple estimator of $\mathbf{M}_{\mathbf{P}}$,

such as $\bar{\mathbf{r}} = I^{-1} \sum_{i=1}^{I} \mathbf{r}_i$, or a simple shrinkage estimator of $\boldsymbol{\rho}_i$ that is more efficient than $\mathbf{R}_i$ (e.g., $\tilde{\boldsymbol{\rho}}_i$ constructed from $J$ applications of a univariate version of Becker & Schram's EM algorithm for $\tilde{\rho}_{ij}$, or a non-iterative variant thereof).

Multivariate ES parameters also arise in meta-analysis for quantities other than correlations. In research domains where binary outcomes are popular, probabilities and functions thereof for multiple groups or outcomes may be of interest. For instance, a research synthesis might focus on the probability of success in each of two conditions, such as a treatment or intervention and some type of control experience (e.g., wait list, placebo medication, sham surgery). Each study's meta-analytic data might consist of one success proportion from Treatment subjects and another from separate Control subjects, $\mathbf{p}_i = [p_{i\mathrm{T}}, p_{i\mathrm{C}}]^{\mathrm{T}}$. Here the two samples' independence implies diagonal $\boldsymbol{\Psi}_{\mathbf{P}i}$, but in $\boldsymbol{\Sigma}_{\boldsymbol{\Pi}}$ the between-studies covariance between $\Pi_{\mathrm{T}}$ and $\Pi_{\mathrm{C}}$ may be non-zero. If the same or related subjects experienced both conditions, such as in a case-control or case-crossover design, $\boldsymbol{\Psi}_{\mathbf{P}i}$ would include non-zero covariance between $P_{i\mathrm{T}}$ and $P_{i\mathrm{C}}$. Instead of meta-analyzing such proportions directly, one might first apply a logit, arcsine, or other transformation.

If comparing $\Pi_{\mathrm{T}}$ and $\Pi_{\mathrm{C}}$ is of central interest, one could recast the above example with proportions as a univariate problem, by transforming each study's data into a difference or ratio of proportions (i.e., risk difference, relative risk) or transformed proportions, an odds ratio, or some transformation thereof (Fleiss, 1994; Haddock, Rindskopf, & Shadish, 1998). In some meta-analyses with categorical data, however, this approach still yields multivariate ES parameters, such as with more than one

treatment group (e.g., mild exercise, vigorous exercise), outcome level (e.g., worse, unchanged, better), or outcome variable (e.g., measured at 6 months and 1 year, assessed under stress or not). Such scenarios are especially likely when synthesizing diverse studies whose different conditions or outcomes cannot justifiably be considered the same. For example, Gleser and Olkin (2000) described fixed-effects meta-analysis procedures for the following scenario: Each study contributes the proportion of "success" (occurrence of heart disease) for a control sample and for one or two independent samples who received one of three anti-hypertension therapies (different samples received different therapies), but no study includes all three therapies. These authors' procedures entail expressing each study's (incomplete) data as one or two treatment-versus-control differences in either proportions or their arcsine or logit transforms, so the ES parameter comprises three components (i.e., one difference for each therapy).

Multivariate categorical data may also arise in systematic reviews of diagnostic tests. In particular, many such tests yield a binary decision (e.g., diseased vs. non-diseased), and when administered to several subjects with and without the target disease—as assessed by a "gold-standard" reference test—they yield a $2 \times 2$ table with the frequency of positive and negative diagnosis for each independent sample of subjects. Some methods for meta-analyzing such studies entail modeling sensitivity and specificity or certain transformations thereof (e.g., logits) as a bivariate ES that is independent within studies but possibly dependent between studies (Harbord et al., 2007; Reitsma, Glas, Rutjes, Scholten, Bossuyt, & Zwinderman, 2005). Other authors have described

multivariate situations where the basic data are proportions or rates (e.g., Arends, Voko, & Stijnen, 2003).

Another common form for multivariate ES data arises when the basic univariate ES is a difference between two samples' means on a quantitative variable. Typically the samples are independent and differ on a fixed experimental or measured independent variable (e.g., intervention, demographic characteristic), the quantitative variable is dependent and random, and the difference in means between samples is standardized by a (possibly pooled) standard deviation from one or both samples. Meta-analytic procedures for mean differences between related samples (Becker, 1988; Morris & DeShon, 2002) or without standardization (Bond, Wiitala, & Richard, 2003) have also been developed but will not be addressed in detail here.

Gleser and Olkin (1994) distinguished between two types of multivariate standardized mean difference (SMD) between independent samples: *Multiple-endpoint* SMDs arise when two samples are compared on each of two or more quantitative dependent variables (i.e., endpoints), such as measurements on two or more constructs or on the same construct but either in two or more conditions (e.g., split-plot design) or on two or more occasions (e.g., pre-post design or longitudinal study). In contrast, *multiple-treatment* SMDs arise when each of two or more independent samples is compared to the same reference sample on one endpoint, such as in studies where two or more competing treatments or interventions are compared to a single control condition. Gleser and Olkin gave expressions for the asymptotic covariance matrix of multiple-endpoint and -treatment SMDs under various assumptions about homogeneity of variances,

covariances, or correlations between samples. They also described fixed-effects techniques for combining and comparing these multivariate SMDs using GLS. In practice, multiple-endpoint SMDs pose more trouble because their covariance matrix relies on within-study correlations between endpoints, which authors of primary studies rarely report. The covariance matrix for multiple-treatment SMDs relies mainly on sample sizes and possibly ratios of (heterogeneous) treatment and control variances.

One can easily imagine more complex scenarios that involve multivariate SMDs. For example, multiple-treatment SMDs may be viewed as a special case of two or more standardized mean contrasts among three or more independent samples. Instead of all pairwise comparisons with a common control group, one might instead be interested in other sets of contrasts. Such a circumstance could arise, for instance, if one were interested in comparing Treatments A and B and some studies provided these data but others provided comparisons between A or B and a third condition, Treatment C (e.g., an SMD version of Ballesteros, 2005, or Lu & Ades, 2004). Another more complex scenario might involve multiple-endpoint SMDs for which the endpoints differ in two or more ways, such as variables $M_1$, $M_2$, $R_1$, and $R_2$ for mathematics and reading measured in two experimental conditions (e.g., teaching methods, study strategies, test types) or on two occasions (e.g., immediate vs. follow-up). Furthermore, multiple-endpoint and -treatment SMDs may co-occur in a single meta-analysis. For example, in their quantitative review of exercise interventions for adults with arthritis, Conn, Hafdahl, Minor, and Nielsen (2008) were interested in SMDs on physical activity, pain, and objective and subjective functional ability: Some primary studies included more than one

treatment sample compared to a common control sample on two or more of these endpoints, in many cases at both pre- and post-intervention occasions (e.g., Lorig, Feigenbaum, Regan, Ung, Chastain, & Holman, 1986).

Numerous types of multivariate ES besides those described above have been or could be addressed in meta-analysis. Examples include variance ratios to compare groups' dispersions, coefficients in regression models (Becker & Wu, 2007), and certain parameters in models for survival data (Arends, Hunink, & Stijnen, 2008).

**2.1.3. Other approaches for this model.** Besides Becker and Schram's (1994) multivariate random-effects techniques for estimation and inferences based on Model 1, other approaches have been proposed. Becker's (1992, 1995) earlier work—perhaps the first published explanation of multivariate random-effects meta-analysis—used a method of moments estimator for $\boldsymbol{\Sigma}_{\Theta}$. In a similar vein, Berkey et al. (1998) described two iterative schemes for estimating $\mathbf{M}_{\Theta}$ and $\boldsymbol{\Sigma}_{\Theta}$ (GLS and marginal ML) in a more general model that can include study-level covariates. Although they considered only multiple-endpoint raw mean differences as ESs (surgical vs. non-surgical mean on post-treatment probing depth and attachment level, from split-mouth studies of treatments for periodontal disease), their model and procedures extend readily to other multivariate ESs. Other authors who have presented estimation strategies for Model 1 or nearly identical models include van Houwelingen, Zwinderman, and Stijnen (1993); Kalaian and Raudenbush (1996); and Whitehead (2002, Appendices A.7 and A.8).

Simulation evidence about the performance of Becker and Schram's (1994) or alternative multivariate random-effects methods is scarce. Two exceptions are Hafdahl's

(2004) aforementioned Monte Carlo study focused on correlation matrices and Riley, Abrams, Sutton, Lambert, and Thompson's (2007) simulations of competing models for bivariate random-effects data. Although the assessment of such techniques is not the primary aim of this thesis, some aspects of the Monte Carlo studies in Chapter 4 pertain to the performance of Becker and Schram's EM-GLS approach.

**2.1.4. Related random-effects models and approaches.** Other authors have described Bayesian approaches for models similar to Model 1. For example, Prevost, Mason, Griffin, Kinmonth, Sutton, and Spiegelhalter (2007) proposed a Bayesian hierarchical model for synthesizing heterogeneous correlation matrices. Their technique involves MCMC for estimation and inferences and can be implemented using widely available software (e.g., WinBUGS). One purported advantage of such a Bayesian approach over the others described above is its explicit incorporation of uncertainty about $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ into inferences about $\mathbf{M}_{\boldsymbol{\Theta}}$. One could presumably extend Prevost et al.'s approach to other multivariate ESs, and Nam et al. (2003) described additional Bayesian approaches for the multivariate random-effects case. Although those and other Bayesian approaches to the present problem are beyond the scope of this thesis, they would be well worth considering for future developments, as would more general mixed models (e.g., Arends, Voko, & Stijnen, 2003) and models for more complex meta-analytic data (e.g., Ades, 2003; Lu & Ades, 2004).

## 2.2. Functions of Effect Sizes

In many realistic situations a meta-analyst will be interested in a different ES metric that is some function $g$ of $\boldsymbol{\Theta}$, say $\boldsymbol{\Gamma} \equiv g(\boldsymbol{\Theta})$, where $g$ may be vector-valued with $K$

components, so that $\boldsymbol{\Gamma} = [g_1(\boldsymbol{\Theta}), g_2(\boldsymbol{\Theta}), ..., g_K(\boldsymbol{\Theta})]^T = [\Gamma_1, \Gamma_2, ..., \Gamma_K]^T$. For various reasons, however, it may be more desirable to conduct initial stages of the meta-analysis in terms of $\boldsymbol{\Theta}$ rather than $\boldsymbol{\Gamma}$, which introduces the complication of transforming certain meta-analytic results between these two metrics. Although an exhaustive survey of candidates for $g$ is not feasible, I describe in this section several examples of such functions and circumstances in which they might arise.

**2.2.1. Variance-stabilizing transformations.** A simple variant of the above problem arises when $g$ is (essentially) the inverse of some transformation that applies to $\boldsymbol{\Theta}$ and $\mathbf{T}_i$ componentwise (i.e., $K = J$ and $g_k = g_1 \; \forall \; k$) and yields ESs that are more amenable to Model 1 and its attendant meta-analytic procedures. Suppose each study contributes $\mathbf{g}_i = g(\mathbf{t}_i)$ but the transform $\mathbf{T}_i = g^{-1}(\mathbf{G}_i)$—or a very similar function of $\mathbf{G}_i$ (perhaps depending on sample size[s])—is nearly multivariate normal with an essentially known covariance matrix, such as an asymptotic approximation whose variances are independent of $\boldsymbol{\gamma}_i = g(\boldsymbol{\theta}_i)$. Such variance-stabilizing transformations are available for proportions (Gleser & Olkin, 2000), SMDs, correlations, and other statistics (Games & Hedges, 1987). One might apply Becker and Schram's (1994) EM-GLS procedures to $\mathbf{T}_i$ to estimate or make inferences about $\mathbf{M}_{\boldsymbol{\Theta}}$ or $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ but instead be interested in $\mathbf{M}_{\boldsymbol{\Gamma}} \equiv \mathrm{E}(\boldsymbol{\Gamma})$ or $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \equiv \mathrm{Cov}(\boldsymbol{\Gamma})$. For example, suppose $\mathbf{R}_i$ and $\mathbf{Z}_i = \tanh^{-1} \mathbf{R}_i$ (i.e., $R_{ij} = \tanh^{-1} Z_{ik} \; \forall \; j = k$) are a sample Pearson-$r$ correlation matrix and its corresponding matrix of Fisher $z$-transforms, as in Section 2.1.2: Despite statistical reasons to meta-analyze $\mathbf{Z}_i$, one might

wish to estimate or make inferences about $\mathbf{M}_P$ and $\mathbf{\Sigma}_P$ for interpretational or other pragmatic reasons (Hafdahl, 2004, 2009a, 2009b).

**2.2.2. Transformations for interpretability.** Other componentwise transformations to facilitate interpretation arise for both multivariate and univariate ESs. For instance, with data from binary diagnostic tests one might meta-analyze sensitivity and (1 minus) specificity as logits but wish to express certain results in terms of the original true- and false-positive rates (e.g., as in a traditional ROC curve; see Section 2.1.2). In a similar vein, some authors favor ES indices that are interpreted more readily by practitioners (e.g., clinicians, physicians), policymakers, or laypersons who are less familiar with statistics (Grissom & Kim, 2005; Kline, 2004, pp. 122-131; Kraemer & Kupfer, 2006; Lipsey & Wilson, 2001, pp. 146-156; Sinclair & Bracken, 1994). For example, one may express a correlation as its square (i.e., proportion of variance), a coefficient of alienation, a common language effect size (CLES; Dunlap, 1994), or a quantity in a binomial effect-size display (BESD; Rosenthal & Rubin, 1982). One may express a SMD as a (biserial or point-biserial) correlation or its square, a BESD or CLES (McGraw & Wong, 1992), any of Cohen's (1988) measures of (non)overlap between two distributions (i.e., $U_1$, $U_2$, $U_3$), a tail ratio (i.e., ratio of two groups' probabilities of scoring below [or above] some low [or high] value), or improvement over chance classification (Huberty & Lowman, 2000). Certain assumptions and criticisms of such indices should be born in mind (e.g., Thompson & Schumacker, 1997; Vargha & Delaney, 2000).

**2.2.3. Other functions.** For many functions of interest each component of $g$ depends one two or more components of $\boldsymbol{\Theta}$. Some examples for bivariate ESs such as SMDs, risk differences, relative risks, and odds ratios (or suitable transforms thereof) include a difference between or ratio of two such ESs, or the minimum or maximum of two or more. Similarly, with (possibly transformed) probabilities from each of two or more groups as ESs, $g$ might be a risk difference, relative risk, odds ratio, likelihood ratio, Youden's index, or number needed to treat (but see Altman & Deeks, 2002, and Cates, 2002). In the case of binary diagnostic tests, one might be interested in functions of the true- and false-positive rates, such as the likelihood ratio for a positive (or negative) test and the diagnostic odds ratio (Reitsma et al., 2005).

**2.2.4. Functions of correlations.** Many functions of substantive value arise with a Pearson-$r$ correlation matrix, $\mathbf{P}$, as $\boldsymbol{\Theta}$ (as in Section 2.1.2). As for scalar $g$ (i.e., $K = 1$), meta-analysts may be interested in a first- or higher-order (semi-)partial correlation between two of the $P$ variables in $\mathbf{Y}$ (i.e., after linearly partialling one or more of the $P - 2$ remaining variables from one or both focal variables), a standardized partial regression coefficient from regressing one of the $P$ variables on two or more of the remaining $P - 1$, or the (squared) multiple correlation from such a regression (or the square's complement, the coefficient of multiple alienation, sometimes expressed as its square root; Cohen, Cohen, West, & Aiken, 2003). Other common examples from multivariate analysis applicable to $\mathbf{P}$ include eigenvalues; determinants; canonical correlation coefficients; various measures of multivariate association, such as those used in set correlation (Cohen, 1982); and coefficients in path, factor, and more general structural equation

models (SEMs; Bollen, 1989). Beyond the present scope are cautions one should attend to when computing or interpreting certain of these quantities, such as critiques of standardized regression coefficients (e.g., Greenland, Schlesselman, & Criqui, 1986; Richards, 1982) and SEM using correlations instead of covariances (e.g., Cudeck, 1989).

Numerous other substantively interesting scalar functions of **P** include differences or more general contrasts or other linear combinations of either zero-order correlations or certain of the aforementioned quantities. Some examples are the difference (a) in a variable's correlation with one versus another variable (e.g., $P_{12} - P_{13}$), (b) in two variables' correlation between two within-subjects conditions or repeated-measures occasions (e.g., $P_{12,A} - P_{12,B}$ for Conditions A and B), (c) between two variables' zero-order and partial correlation of some order (e.g., $P_{12} - P_{12.3}$ as a test of mediation; Olkin & Finn, 1995) or between the corresponding standardized regression coefficients (e.g., $B_{12} - B_{12.3}$), (d) in two variables' partial correlation between different (sets of) partialled variables (e.g., $P_{12.3} - P_{12.4}$), (e) in squared multiple correlation between two (possibly nested) sets of predictors for one variable (e.g., $P^2_{1.23} - P^2_{1.2}$), and (f) between two regression coefficients from the same or different regression equations (e.g., one predictor's coefficient with vs. without another [set of] predictor[s]). One might also be interested in more complicated combinations of correlation or regression coefficients, such as a second- or higher-order difference (e.g., a treatment difference in the change in correlation: $[P_{12,A2} - P_{12,A1}] - [P_{12,B2} - P_{12,B1}]$ for Treatments A and B at Occasions 1 and 2) or a tetrad difference (e.g., $P_{13}P_{24} - P_{23}P_{14}$). Moreover, vector-valued functions of **P**

arise naturally as *K*-tuples of parameters in multiple regression and other linear models (e.g., path coefficients, factor loadings) or other collections of scalar quantities. Finally, if $\boldsymbol{\Theta}$ were $\mathbf{Z}$ then one could apply any aforementioned function to $\mathbf{Z}$ after transforming it to $\mathbf{P}$—that is, $g$ would be a composite function $h_2 \circ h_1$, where $h_1(\mathbf{Z}) \equiv \tanh \mathbf{Z} = \mathbf{P}$.

### 2.3. Meta-Analysis for Functions of Effect Sizes

Major aims of the present thesis are to develop and evaluate techniques to address certain meta-analytic tasks for a wide variety of functions of ESs. In particular, considering any function in the previous section as $\boldsymbol{\Gamma} \equiv g(\boldsymbol{\Theta})$, I am most interested in obtaining point estimators of $\mathbf{M}_{\boldsymbol{\Gamma}} \equiv \mathrm{E}(\boldsymbol{\Gamma})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \equiv \mathrm{Cov}(\boldsymbol{\Gamma})$ as well as confidence sets and tests for $\mathbf{M}_{\boldsymbol{\Gamma}}$. Of particular interest are circumstances where meta-analytic procedures such as Becker and Schram's (1994) EM-GLS techniques are more appropriate for $\mathbf{T}_i$ than $\mathbf{G}_i = g(\mathbf{T}_i)$ but $\boldsymbol{\Gamma}$ is of more substantive relevance than $\boldsymbol{\Theta}$. In this section I comment on these circumstances and on the scant extant work on this problem.

**2.3.1. Typical circumstances.** There are two primary reasons that one might wish to apply meta-analysis in the $\boldsymbol{\Theta}$ metric despite being more interested substantively in one of many options for $g$ and $\boldsymbol{\Gamma}$. First, as already alluded to, methods such as Becker and Schram's (1994) estimators and inference techniques might perform better with $\mathbf{T}_i$ than $\mathbf{G}_i$, such as if $\mathbf{T}_i$ conforms better to assumptions of normality and known covariance matrix or $\boldsymbol{\Theta}$'s distributional properties are more amenable to certain procedures (e.g., when components of $\boldsymbol{\Gamma}$ are bounded but those of $\boldsymbol{\Theta}$ are not). Second, some studies may be missing one or more components of $\mathbf{T}_i$ so that $\mathbf{G}_i$ cannot be computed for each study,

but under some missingness mechanisms (e.g., missing completely at random) it may be feasible to meta-analyze the incomplete $\mathbf{T}_i$ and use results for $\boldsymbol{\Theta}$ to obtain results for $\boldsymbol{\Gamma}$. For instance, suppose each of several studies contributes one or two (but not all 3) of $r_{i12}$, $r_{i13}$, and $r_{i23}$: One could estimate $\boldsymbol{\rho} = [\rho_{12}, \rho_{13}, \rho_{23}]^T$ in a fixed-effects meta-analysis model or $\mathbf{M}_\mathbf{P}$ and $\boldsymbol{\Sigma}_\mathbf{P}$ for $\mathbf{P} = [P_{12}, P_{13}, P_{23}]^T$ in a random-effects model. These results could be used to estimate and make inferences about, say, a common $\rho_{12.3}$ or hyperparameters of a random $P_{12.3}$ (the partial correlation between $Y_1$ and $Y_2$, partialling $Y_3$), even though $r_{i12.3}$ could not be computed for any single study.

**2.3.2. Previous relevant work.** Despite the potential value of obtaining meta-analytic results for functions of the directly analyzed ESs, little attention appears to have been paid to this problem, especially in the case I am most interested in: random-effects methods for multivariate ESs. Nevertheless, it is worth commenting on some methods developed—or at least used without explicit justification—for that case as well as some popular special cases, beginning with the latter.

First consider the *univariate fixed-effects* version of Model 1, where $J = 1$ and $U_i = 0 \; \forall \; i$, and suppose $g$ is scalar (i.e., $K = 1$). If we assume further that $u_i = 0 \; \forall \; i$ (i.e., between-studies homogeneity), then the main task is to obtain $\hat{\theta}$ as a fixed-effects estimate of the common ES, $\theta$, as in Equation 5. It is usually reasonable to use $\hat{\gamma} = g(\hat{\theta})$ as a point estimator of $\gamma \equiv g(\theta)$, and in practice this is often done without remark (e.g., Shadish & Haddock, 1994). Provided $g$ is invertible, as it is for inverse variance-stabilizing transformations, many interpretive indices, and other functions, a CI for $\gamma$ is

usually constructed by simply applying $g$ to both endpoints of a CI for $\theta$. Also, one may test $H_0$: $\gamma = c$ by testing $H_0$: $\theta = g^{-1}(c)$. For non-monotonic $g$, which is less common but plausible (e.g., $\rho^2$), different inference approaches are needed. I am not aware of methods proposed for this in a meta-analytic context, but perhaps some have been described or at least mentioned in passing among the more than 6,000 articles, chapters, dissertations, conference papers, and other works on research-synthesis methodology. One strategy is to use a delta-method approximation of the variance of $\hat{\theta}$, as Olkin and Finn (1990, 1995) suggested for a primary study, and another is to adapt a resampling technique such as bootstrapping for meta-analysis (Van den Noortgate & Onghena, 2005). Both of these are special cases of methods described in Chapter 3. I do not consider the heterogeneous fixed-effects case here (but see Bonett, 2008, 2009).

Now consider the *univariate random-effects* version of Model 1, with $J = 1$, and suppose again that $g$ is scalar. In this case, Becker and Schram's (1994) procedures—or various univariate alternatives (e.g., DerSimonian & Laird, 1986; DerSimonian & Kacker, 2007; Hedges, 1983; Raudenbush, 2009; Shadish & Haddock, 2009)—yield point estimators of $M_\Theta$ and $\Sigma_\Theta$ and a CI or test for $M_\Theta$. A meta-analyst interested in $\Gamma \equiv g(\Theta)$, however, would instead prefer estimators of and inferences about $M_\Gamma$ and $\Sigma_\Gamma$. Some authors have described or used in passing the point estimator $g(\hat{M}_\Theta)$ and similarly applied $g$ directly to the endpoints of a CI for $M_\Theta$. These may be meant as point and interval estimators of $M_\Gamma$, but this is rarely explicated: Most authors simply state that these transformations express the estimators in the desired metric (i.e., $\Gamma$). As Schulze

(2004, pp. 75-79) and Hafdahl (2009a) discussed in the situation I refer to herein as the "univariate *z*-to-*r* case"—namely, $J = K = 1$ with Z as $\Theta$ and P = tanh Z as $\Gamma$—the intended estimand of these direction transformations is ambiguous, and they may yield poor estimators of $M_\Gamma$ (e.g., with large $\Sigma_\Theta$ or $M_\Theta$). As a simple example, consider $\Theta^2$ as $\Gamma$: For many realistic distributions $\mathrm{Var}(\Theta) = E(\Theta^2) - [E(\Theta)]^2$, so the direct transformation $g(M_\Theta) = [E(\Theta)]^2$ is smaller than $M_\Gamma = E[g(\Theta)] = E(\Theta^2)$ by $\Sigma_\Theta = \mathrm{Var}(\Theta)$.

In the univariate *z*-to-*r* case, Hafdahl's (2009a) Table 1 gives the discrepancy between tanh $M_Z$ and $M_P$ in several realistic conditions (and some unrealistic ones—Hafdahl, 2009a, and Hafdahl & Williams, 2009, highlighted deficiencies in prominent Monte Carlo simulations that some authors have cited as evidence against analyzing Fisher-*z* correlations). Positing that this discrepancy is largely responsible for the sometimes poor performance of tanh $\hat{M}_Z$ and an associated CI as estimators of $M_P$, Hafdahl (2009a) proposed alternative estimators that use a more defensible integral *z*-to-*r* transformation (IZRT), which formalizes Law's (1995) "point approximation" approach. This IZRT and a version for variances to estimate $\Sigma_P$ are special cases of the more general integral transformations described in Chapter 3. Such an integral transformation, however, complicates testing $M_\Gamma$ or constructing a CI for it. Hafdahl's (2009c) simulation studies showed that Hafdahl's (2009a) proposed approaches—application of the IZRT to a CI for $M_Z$, and a delta-method variance for $\hat{M}_P$—work reasonably well in many conditions. In Chapter 3 I describe generalizations of those procedures for generic ESs as well as bootstrap approaches.

As for the *multivariate homogeneous fixed-effects* version of Model 1, with $\mathbf{U}_i = \mathbf{0}$ $\forall\ i$, the main task is to obtain $\hat{\boldsymbol{\theta}}$ as an estimate of $\boldsymbol{\theta}$, as in Equation 5. Like in the univariate case, a meta-analyst interested in $\boldsymbol{\gamma} \equiv g(\boldsymbol{\theta})$ could reasonably use $\hat{\boldsymbol{\gamma}} = g(\hat{\boldsymbol{\theta}})$ as a point estimator, where $g$ may be scalar or vector-valued. This was, for instance, Becker's (1992) approach for estimating a vector of standardized regression coefficients (say, $\boldsymbol{\beta}$) from a pooled correlation matrix ($\hat{\boldsymbol{\rho}}$). Other authors have used this approach to estimate exploratory factor models (Hafdahl, 2001), path models (S. F. Cheung, 2000), or more general SEMs (M. W.-L. Cheung & Chan, 2005; Furlow & Beretvas, 2005) from a pooled correlation matrix. Except in special cases (e.g., componentwise functions), inference about $\boldsymbol{\gamma}$ or any of its components cannot be based on inference about $\boldsymbol{\theta}$ (e.g., $g$ applied to CIs or more general confidence sets for $\boldsymbol{\theta}$). Becker gave a formula for the Jacobian matrix needed to obtain an approximate covariance matrix for $\hat{\boldsymbol{\beta}}$ from that for $\hat{\boldsymbol{\rho}}$ by the multivariate delta method. This strategy may be generalized readily to generic multivariate ESs and more general $g$, which is a special case of one approach described in Chapter 3. A drawback of this strategy in practice is the required derivatives, which will be difficult to obtain for many meta-analysts. Solutions to this barrier, including numerical differentiation and bootstrap alternatives, are also addressed in Chapter 3.

Finally, with no restrictions on Model 1, suppose a meta-analyst obtains $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$, $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$, and $\hat{\mathrm{C}}\mathrm{ov}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ but would like estimates of and inferences about $\mathbf{M}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$. As in the univariate random-effects case, authors who have described methods for these latter

estimates and inferences seem to have used $g(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ or an analogous estimator without clarifying the intended estimand. For example, Becker (1992) and Becker and Schram (1994) considered a vector of standardized regression coefficients as a function $g$ of a Pearson correlation matrix (i.e., with $\mathbf{P}$ as $\boldsymbol{\Theta}$ and $\mathbf{B}$ as $\boldsymbol{\Gamma}$): They applied the same $g$ to $\hat{\mathbf{M}}_{\mathbf{P}}$ as one would apply to $\rho$ or $\rho_i$ and obtained $\hat{\mathrm{C}}\mathrm{ov}[g(\hat{\mathbf{M}}_{\mathbf{P}})]$ by the same multivariate delta method as in the fixed-effects case. As an example involving binary outcomes, Reitsma et al. (2005) described a bivariate random-effects analysis of, essentially, logit true- and false-positive rates—say, $\boldsymbol{\Lambda} = [\Lambda_T, \Lambda_F]^T$—that compares favorably with competitor techniques for meta-analyzing diagnostic tests (Harbord et al., 2007). (They in fact analyzed logit sensitivity and specificity, $\Lambda_T$ and $-\Lambda_F$.) They reported point and interval estimates for the mean of sensitivity, specificity, and the diagnostic odds ratio in their Table 2 (i.e., $[1 + \exp(-\Lambda_T)]^{-1}$, $[1 + \exp(\Lambda_F)]^{-1}$, and $\exp[\Lambda_T - \Lambda_F]$, respectively) and suggested other useful functions of $\boldsymbol{\Lambda}$ in their Appendix 1. All of their estimates appear to entail applying the relevant function directly to $\hat{\mathbf{M}}_{\boldsymbol{\Lambda}}$ or to a CI for $\mathrm{M}_{\boldsymbol{\Lambda}\mathrm{T}}$ or $\mathrm{M}_{\boldsymbol{\Lambda}\mathrm{F}}$.

Despite the appeal of the above direct approach based on $g(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$, except in special circumstances $g(\mathbf{M}_{\boldsymbol{\Theta}}) \neq \mathbf{M}_{\boldsymbol{\Gamma}}$, and this discrepancy's form and magnitude depend on not only $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ but also the distribution of $\boldsymbol{\Theta}$. Hence, in some situations $g(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ may estimate $\mathbf{M}_{\boldsymbol{\Gamma}}$ poorly. One might argue that $g(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ is intended to estimate $g(\mathbf{M}_{\boldsymbol{\Theta}})$, but to my knowledge no author has explicated that as the estimand. Indeed, the value of $g$ at $\mathbf{M}_{\boldsymbol{\Theta}}$ seems unlikely to be of interest to an applied meta-analyst, especially when $\boldsymbol{\Gamma}$ is of

substantive interest and $\boldsymbol{\Theta}$ is merely computationally expedient: It seems more likely that he or she would wish to estimate $\mathbf{M}_{\boldsymbol{\Gamma}}$ rather than $g(\mathbf{M}_{\boldsymbol{\Theta}})$. Although $g(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ might be of interest as a prediction of $\boldsymbol{\gamma}$ in a future study for which $\boldsymbol{\theta}_{J+1} = \mathbf{M}_{\boldsymbol{\Theta}}$, it is again unclear why $\mathbf{M}_{\boldsymbol{\Theta}}$ is a point of interest for evaluating $g(\boldsymbol{\theta})$, and because $\hat{\text{Cov}}[g(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})]$ in Becker's (1992) approach does not capture error due to $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ it does not seem to be intended for prediction of a future study (cf. Harbord et al., 2007). Furthermore, an estimator for $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ may be of interest to some meta-analysts but does not seem to have been proposed for this case of a generic multivariate ES. A primary contribution of this thesis is my proposal in Chapter 3 of techniques to estimate and makes inferences about $\mathbf{M}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$.

## 3. PROPOSED ESTIMATION AND INFERENCE TECHNIQUES

In Chapter 2 I claimed that in many realistic situations a meta-analyst may wish to estimate and make inferences about the between-studies mean or covariance matrix of some function ($g$) of heterogeneous ESs ($\boldsymbol{\Theta}$, from which $\boldsymbol{\theta}_i$ arises) whose sample estimates ($\mathbf{t}_i$ from $\mathbf{T}_i$) are analyzed by a conventional approach such as Becker and Schram's (1994) EM-GLS procedures. Moreover, I contended that procedures for these tasks do not seem to have been addressed comprehensively. The major aims of the present thesis are to develop principled techniques for these meta-analytic tasks and evaluate them by Monte Carlo simulation. This chapter focuses on proposed techniques; the next, on evaluations thereof. After describing one straightforward but often infeasible approach, I describe point estimators of the target function's mean and covariance matrix followed by strategies for inference on the function's mean. I also comment briefly on two special cases of interest.

### 3.1. Direct Meta-Analysis of Function

If every study contributes a sample estimate $\mathbf{g}_i$ of the target function's parameter for that study ($\boldsymbol{\gamma}_i$) and both $\mathbf{G}_i$ and $\boldsymbol{\Gamma}$ conform reasonably well to Model 1, then one may simply use Becker and Schram's (1994) procedures to estimate $\mathbf{M}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$, make inferences about $\mathbf{M}_{\boldsymbol{\Gamma}}$ or its components, test $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ or its components, and so on. For example, in the case of correlations $\mathbf{t}_i$ might be $\mathbf{r}_i$, a sample correlation matrix among $P$ variables, and $\mathbf{g}_i$ might be $\mathbf{b}_i$, a vector of $K \leq P - 1$ standardized regression coefficients (i.e., $\mathbf{P}$ as $\boldsymbol{\Theta}$ and $\mathbf{B}$ as $\boldsymbol{\Gamma}$): Provided that one could obtain from each study $\mathbf{b}_i$ itself (with the same outcome and regressor variables) or a complete $\mathbf{r}_i$ among the same $P$ variables

as well as an approximate $\boldsymbol{\Psi}_{\mathbf{B}i} \equiv \text{Cov}(\mathbf{B}_i)$ (e.g., using Becker's, 1992, delta-method approach or a reported covariance matrix), one could apply Becker and Schram's EM algorithm to $\mathbf{b} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_K]^T$ to obtain $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{B}}$, apply GLS to obtain $\hat{\mathbf{M}}_{\mathbf{B}}$ and $\hat{\text{Cov}}(\hat{\mathbf{M}}_{\mathbf{B}})$ for CIs or other confidence sets (Equations 2 and 3), and test $H_0$: $\boldsymbol{\Sigma}_{\mathbf{B}} = \mathbf{0}$ (Equation 4). Becker and Wu (2007) described a fixed-effects variant of this direct strategy applied to unstandardized (i.e., raw) regression coefficients, and applying Becker and Schram's procedures instead would accommodate between-studies heterogeneity.

Some aspects of this direct meta-analysis of $\mathbf{g}_i$ are attractive: Not only does it yield standard meta-analytic results in the $\boldsymbol{\Gamma}$ metric immediately, but it extends readily to other multivariate random-effects models and procedures, such as models with covariates (Berkey et al., 1998; Kalaian & Raudenbush, 1996). Despite these advantages, this approach suffers from a major practical drawback: As discussed in Section 2.3.1, in many situations a complete $\mathbf{g}_i$ will not be available from every study, and even when it is the standard meta-analytic procedures may perform poorly due to properties of $\mathbf{G}_i$ or $\boldsymbol{\Gamma}$. For instance, Becker and Wu (2007) discussed several limitations of their direct approach for regression coefficients, such as the usual incomparability of regression models across studies and lack of available data for $\boldsymbol{\Psi}_{\mathbf{G}i}$. Their real-data example relied on a large national survey (NELS:88) that provided subject-level data from each of several schools, yielding all results for direct meta-analysis. Given access to such subject-level data, however, meta-analysis may be less appropriate than, say, a standard mixed linear model.

### 3.2. Point Estimation for Function's Mean

The point estimators of $\mathbf{M}_\Gamma$ I propose herein rely on a simple key idea: Transformations between moments of $\Theta$ and of $\Gamma$ should incorporate the integration that defines these moments. In particular, it is evident that $\mathbf{M}_\Gamma = g(\mathbf{M}_\Theta)$ will not hold for many functions $g$. Instead, because $\mathbf{M}_\Gamma = \mathrm{E}[g(\Theta)] = \int g(\mathbf{a}) f_\Theta(\mathbf{a}) d\mathbf{a}$, where $f_\Theta(\mathbf{a})$ is the density function for $\Theta$ at $\Theta = \mathbf{a} \in \mathbb{R}^J$, $\mathbf{M}_\Gamma$ is not in general determined by $\mathbf{M}_\Theta$ alone, $\mathbf{M}_\Theta$ and $\mathbf{\Sigma}_\Theta$, or any other set of moments of $\Theta$, unless those moments characterize $f_\Theta$. This complicates any attempt to estimate $\mathbf{M}_\Gamma$ from estimates of, say, $\mathbf{M}_\Theta$ and $\mathbf{\Sigma}_\Theta$. Nevertheless, some straightforward estimators may be obtained under certain reasonable assumptions about $f_\Theta$—such as multivariate normality or multivariate-normal lower-order moments. Two such approaches are described below.

**3.2.1. Integral transformation.** If $\Theta \sim N_J(\mathbf{M}_\Theta, \mathbf{\Sigma}_\Theta)$, as is assumed for Becker and Schram's (1994) ML estimator of $\mathbf{\Sigma}_\Theta$, then we can express $\mathbf{M}_\Gamma$ using the following function of a $J \times 1$ real vector $\mathbf{M}$ and a $J \times J$ symmetric, positive definite real matrix $\mathbf{\Sigma}$, which for convenience I call the mean integral transformation (MIT) of $\mathbf{M}$ and $\mathbf{\Sigma}$:

$$S_1(\mathbf{M}, \mathbf{\Sigma}) \equiv \int g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \ , \tag{8}$$

where $\mathbf{x} \in \mathbb{R}^J$ and $f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-J/2} |\mathbf{\Sigma}|^{-1/2} \exp[-(\mathbf{x} - \mathbf{M})^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{M}) / 2]$. If $\Theta \sim N_J(\mathbf{M}_\Theta, \mathbf{\Sigma}_\Theta)$, then $\mathbf{M}_\Gamma = S_1(\mathbf{M}_\Theta, \mathbf{\Sigma}_\Theta)$, so a sensible estimator of $\mathbf{M}_\Gamma$ based on estimators of $\mathbf{M}_\Theta$ and $\mathbf{\Sigma}_\Theta$ is

$$\hat{\mathbf{M}}_\Gamma = S_1(\hat{\mathbf{M}}_\Theta, \tilde{\mathbf{\Sigma}}_\Theta) \ . \tag{9}$$

If $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ were both ML estimators, then $\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}$ would be an ML estimator of $\mathbf{M}_{\boldsymbol{\Gamma}}$ by the invariance property of ML. Because Becker and Schram's (1994) $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ is not ML, however, $\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}$ is not in general ML. It is, nevertheless, a reasonable estimator whose performance under realistic conditions merits consideration. In practice, the integral $S_1$ will rarely exist in closed form, except when $g$ is affine or takes other special forms. It will therefore often be necessary to evaluate $S_1$ by numerical or simulation methods, such as Law (1995) and Hafdahl (2009a, 2009c) described for the univariate $z$-to-$r$ case. The following simple Monte Carlo approximation will often suffice:

1. Draw $\boldsymbol{\theta}^*_m$ from $N_J(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}})$ for $m = 1, 2, ..., M$, where $M$ is large.

2. Compute $\boldsymbol{\gamma}^*_m = g(\boldsymbol{\theta}^*_m)$.

3. Compute $\hat{\mathbf{M}}_{\boldsymbol{\Gamma}} = M^{-1} \sum_{m=1}^{M} \boldsymbol{\gamma}^*_m$.

**3.2.2. Approximation of function.** An alternative strategy is to approximate $g$ by a simpler function whose expected value is a fairly tractable function of $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$. For instance, for the univariate $z$-to-$r$ case Hafdahl (2009a) described an estimator of $M_P \equiv E(P)$ that relies on the following expectation of a second-order Taylor-series approximation of $P = \tanh Z$ expanded at $M_Z$:

$$M_P \approx E[\varphi + (1 - \varphi^2)(Z - M_Z) - \varphi(1 - \varphi^2)(Z - M_Z)^2] = \varphi[1 - \Sigma_Z(1 - \varphi^2)] , \qquad (10)$$

where $\varphi = \tanh M_Z$. This suggests the estimator $\breve{M}_P = \hat{\varphi}[1 - \widetilde{\Sigma}_Z(1 - \hat{\varphi}^2)]$, where $\hat{\varphi} = \tanh \hat{M}_Z$, which differs from Law's (1995) TS2 point estimator only in the specific estimators of $M_Z$ and $\Sigma_Z$.

For appropriate (e.g., sufficiently differentiable) functions, such a second-order Taylor-series mean (MTS2) estimator is extended readily to the more general case with $J > 1$ for $\boldsymbol{\Theta}$ or $K > 1$ for $\boldsymbol{\Gamma}$. Letting $\boldsymbol{\Delta} \equiv \boldsymbol{\Theta} - \mathbf{M}_{\boldsymbol{\Theta}}$, consider the Taylor polynomial of degree two, expanded at $\mathbf{M}_{\boldsymbol{\Theta}}$, as a quadratic approximation to $g_k(\boldsymbol{\Theta})$:

$$g_k(\boldsymbol{\Theta}) \approx Qg_k(\boldsymbol{\Theta}; \mathbf{M}_{\boldsymbol{\Theta}})$$

$$\equiv g_k(\mathbf{M}_{\boldsymbol{\Theta}}) + \nabla g_k(\mathbf{M}_{\boldsymbol{\Theta}})^{\mathrm{T}}\boldsymbol{\Delta} + \boldsymbol{\Delta}^{\mathrm{T}}\mathbf{H}g_k(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Delta} / 2 , \tag{11}$$

where $\nabla g_k(\mathbf{M}_{\boldsymbol{\Theta}})$ and $\mathbf{H}g_k(\mathbf{M}_{\boldsymbol{\Theta}})$ are, respectively, the gradient vector and Hessian matrix of $g_k(\boldsymbol{\Theta})$ with respect to $\boldsymbol{\Theta}$ evaluated at $\boldsymbol{\Theta} = \mathbf{M}_{\boldsymbol{\Theta}}$ (i.e., $\boldsymbol{\Delta} = \mathbf{0}$). Using a result on expectations of quadratic forms in a random vector that has an arbitrary distribution with finite fourth moments (e.g., Schott, 1997), we obtain $\mathrm{E}[\boldsymbol{\Delta}^{\mathrm{T}}\mathbf{H}g_k(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Delta}] = \mathrm{tr}[\mathbf{H}g_k(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}]$. Hence, for the $k$th component of $g(\boldsymbol{\Theta})$ we have the second-order Taylor-series approximation

$$\mathrm{M}_{\boldsymbol{\Gamma}k} \equiv \mathrm{E}[g_k(\boldsymbol{\Theta})] \approx \mathrm{E}[Qg_k(\boldsymbol{\Theta}; \mathbf{M}_{\boldsymbol{\Theta}})] = g_k(\mathbf{M}_{\boldsymbol{\Theta}}) + \mathrm{tr}[\mathbf{H}g_k(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}] / 2 . \tag{12}$$

This suggests the following MTS2 estimator of $\mathbf{M}_{\boldsymbol{\Gamma}}$ based on $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$:

$$\breve{\mathbf{M}}_{\boldsymbol{\Gamma}} = \begin{bmatrix} g_1(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}) + \mathrm{tr}[\mathbf{H}g_1(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}] / 2 \\ g_2(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}) + \mathrm{tr}[\mathbf{H}g_2(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}] / 2 \\ \vdots \\ g_K(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}) + \mathrm{tr}[\mathbf{H}g_K(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}] / 2 \end{bmatrix} = g(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}) + \frac{1}{2}\begin{bmatrix} \mathrm{tr}[\mathbf{H}g_1(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}] \\ \mathrm{tr}[\mathbf{H}g_2(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}] \\ \vdots \\ \mathrm{tr}[\mathbf{H}g_K(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}] \end{bmatrix} . \tag{13}$$

To be feasible for practicing meta-analysts, the second-order (mixed) partial derivatives in $\mathbf{H}g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ may be evaluated numerically. For example, one could approximate element $j$, $l$ of $\mathbf{H}g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ using the following central difference quotient:

$$\left.\frac{\partial^2 g_k(\boldsymbol{\Theta})}{\partial\Theta_j\Theta_l}\right|_{\boldsymbol{\Theta}=\hat{\mathbf{M}}_{\boldsymbol{\Theta}}} \approx \frac{\left[\begin{array}{l} g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}+\boldsymbol{\varepsilon}_j+\boldsymbol{\varepsilon}_l)-g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}+\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_l)- \\ g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}-\boldsymbol{\varepsilon}_j+\boldsymbol{\varepsilon}_l)+g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}-\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_l) \end{array}\right]}{4\varepsilon^2}, \tag{14}$$

where ε is a suitably small increment that balances truncation and round-off error, and each of $\boldsymbol{\varepsilon}_j$ and $\boldsymbol{\varepsilon}_l$ is a $J$-element column vector with ε in position $j$ or $l$, respectively, and 0 elsewhere. Compared to the MIT estimator, this MTS2 estimator will often be computationally cheaper and may be more robust to non-normality of $\boldsymbol{\Theta}$. It may perform poorly, however, for functions approximated poorly by a second-order Taylor polynomial near $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$, which may in turn depend on $\mathbf{M}_{\boldsymbol{\Theta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$, or other features of $\boldsymbol{\Theta}$'s distribution.

### 3.3. Point Estimation for Function's Covariance Matrix

Relying on the same basic ideas as for the MIT and MTS2 estimators of $\mathbf{M}_{\boldsymbol{\Gamma}}$, point estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ may be obtained by approximating either the relevant integral or $g$.

**3.3.1. Integral transformation.** Assuming $\boldsymbol{\Theta} \sim N_J(\mathbf{M}_{\boldsymbol{\Theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}})$, as for $\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}$ in Section 3.2.1, and recalling that $\mathbf{M}_{\boldsymbol{\Gamma}} = S_1(\mathbf{M}_{\boldsymbol{\Theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}})$, we can express $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ using the following covariance integral transformation (CIT) of $\mathbf{M}$ and $\boldsymbol{\Sigma}$:

$$S_2(\mathbf{M}, \boldsymbol{\Sigma}) = \int [g(\mathbf{x}) - S_1(\mathbf{M}, \boldsymbol{\Sigma})][g(\mathbf{x}) - S_1(\mathbf{M}, \boldsymbol{\Sigma})]^{\mathrm{T}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \tag{15}$$

where $\mathbf{M}$, $\boldsymbol{\Sigma}$, and $f_{\mathbf{X}}(\mathbf{x})$ are defined as for Equation 8. Namely, if $\boldsymbol{\Theta} \sim N_J(\mathbf{M}_{\boldsymbol{\Theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}})$, then $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} = S_2(\mathbf{M}_{\boldsymbol{\Theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}})$. Hence, a sensible estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ based on $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}} = S_2(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}). \tag{16}$$

As with $\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}$ would be an ML estimator if both $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ were, but even if they are not it is reasonable and warrants investigation. Except in special cases, it will be necessary to evaluate $S_2$ by numerical or simulation methods, such as Law (1995) and Hafdahl (2009a, 2009c) described for the univariate *z*-to-*r* case. In many situations the following step added to the Monte Carlo approximation of $\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}$ (see Section 3.2.1) will suffice: Compute the usual (unbiased) covariance-matrix estimator $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}} = (M - 1)^{-1}\mathbf{D}\mathbf{D}^{\mathrm{T}}$, where $\mathbf{D} = [\, \boldsymbol{\gamma}_1^* - \hat{\mathbf{M}}_{\boldsymbol{\Gamma}}, \boldsymbol{\gamma}_2^* - \hat{\mathbf{M}}_{\boldsymbol{\Gamma}}, \ldots, \boldsymbol{\gamma}_M^* - \hat{\mathbf{M}}_{\boldsymbol{\Gamma}} \,]$ is a $K \times M$ matrix of deviations.

**3.3.2. Approximation of function.** By analogy with the MTS2 estimator for $\mathbf{M}_{\boldsymbol{\Gamma}}$, one could approximate $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ using the covariance matrix of a second-order Taylor-series approximation. For the univariate *z*-to-*r* case Hafdahl (2009a, 2009c) described an estimator of $\Sigma_P \equiv \mathrm{Var}(P)$ that uses the following variance of such an approximation of P:

$$\Sigma_P \approx \mathrm{Var}[\varphi + (1 - \varphi^2)(Z - M_Z) - \varphi(1 - \varphi^2)(Z - M_Z)^2] = \Sigma_Z(1 - \varphi^2)^2(1 + 2\varphi^2\Sigma_Z) \,, \quad (17)$$

where the third and fourth moments of Z are assumed to be the same as normal—that is, $\mathrm{E}[(Z - M_Z)^3] = 0$ and $\mathrm{E}[(Z - M_Z)^4] = 3\,\Sigma_Z^2$—so $\Sigma_P$ does not require estimating these additional moments. This suggests the estimator $\breve{\Sigma}_P = \widetilde{\Sigma}_Z(1 - \hat{\varphi}^2)^2(1 + 2\,\hat{\varphi}^2\,\widetilde{\Sigma}_Z)$, which is essentially the same as Law's (1995) TS2 point estimator.

Again, for appropriate functions such a second-order Taylor-series covariance (CTS2) estimator extends readily to the more general case of $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ when $J > 1$ for $\boldsymbol{\Theta}$ or $K > 1$ for $\boldsymbol{\Gamma}$. Consider again $Qg_k(\boldsymbol{\Theta}; \mathbf{M}_{\boldsymbol{\Theta}})$, the quadratic approximation to $g_k(\boldsymbol{\Theta})$ given in

Equation 11. By some results on variances and covariances of quadratic forms in a multivariate-normal random vector (e.g., Schott, 1997), if $\boldsymbol{\Theta} \sim N_J(\mathbf{M}_{\boldsymbol{\Theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}})$ then

$$\mathrm{Cov}[(\boldsymbol{\Delta}^{\mathrm{T}}\mathbf{H}g_k(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Delta}, \boldsymbol{\Delta}^{\mathrm{T}}\mathbf{H}g_l(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Delta}] = 2\mathrm{tr}[\mathbf{H}g_k(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\mathbf{H}g_l(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}]$$

and

$$\mathrm{Cov}[\nabla g_k(\mathbf{M}_{\boldsymbol{\Theta}})^{\mathrm{T}}\boldsymbol{\Delta}, \boldsymbol{\Delta}^{\mathrm{T}}\mathbf{H}g_l(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Delta}] = 0 .$$

Hence, for the covariance of the $k$th and $l$th components of $g(\boldsymbol{\Theta})$ we have the second-order Taylor-series approximation

$$\begin{aligned}\Sigma_{\boldsymbol{\Gamma}kl} &\equiv \mathrm{Cov}[g_k(\boldsymbol{\Theta}), g_l(\boldsymbol{\Theta})] \approx \mathrm{Cov}[Qg_k(\boldsymbol{\Theta}; \mathbf{M}_{\boldsymbol{\Theta}}), Qg_l(\boldsymbol{\Theta}; \mathbf{M}_{\boldsymbol{\Theta}})] \\ &= \nabla g_k(\mathbf{M}_{\boldsymbol{\Theta}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\nabla g_l(\mathbf{M}_{\boldsymbol{\Theta}}) + \mathrm{tr}[\mathbf{H}g_k(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\mathbf{H}g_l(\mathbf{M}_{\boldsymbol{\Theta}})\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}] / 2 . \qquad (18)\end{aligned}$$

Substituting $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ for $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ in this expression yields the CTS2 estimator $\breve{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}$, of which element $k$, $l$ is

$$\breve{\Sigma}_{\boldsymbol{\Gamma}kl} = \nabla g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}\nabla g_l(\hat{\mathbf{M}}_{\boldsymbol{\Theta}}) + \mathrm{tr}[\mathbf{H}g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}\mathbf{H}g_l(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}]/ 2 . \qquad (19)$$

As with $\mathbf{H}g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$, the partial derivatives in $\nabla g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ may be evaluated numerically. For example, element $j$ of $\nabla g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ may be approximated using the following central difference quotient:

$$\left.\frac{\partial g_k(\boldsymbol{\Theta})}{\partial \Theta_j}\right|_{\boldsymbol{\Theta}=\hat{\mathbf{M}}_{\boldsymbol{\Theta}}} \approx \frac{g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}} + \boldsymbol{\varepsilon}_j) - g_k(\hat{\mathbf{M}}_{\boldsymbol{\Theta}} - \boldsymbol{\varepsilon}_j)}{2\varepsilon} . \qquad (20)$$

### 3.4. Inference on Function's Mean

It is often desirable to accompany a point estimate of $\mathbf{M}_{\boldsymbol{\Gamma}}$ with a confidence set or test. In this section I describe two general strategies for accomplishing these tasks. Each

of these is applicable to both of the MIT and MTS2 estimation approaches described above, whose mean estimators I denote generically as $\overline{\mathbf{M}}_{\boldsymbol{\Gamma}} \in \{\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}, \breve{\mathbf{M}}_{\boldsymbol{\Gamma}}\}$.

**3.4.1. Delta method.** We may view $\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}$ as a function of $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ (i.e., MIT or MTS2) with $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ fixed and known. This treatment of $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ neglects sampling error in $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$, which is not uncommon in meta-analytic inference about the mean ES (e.g., Becker & Schram's, 1994, GLS). For many functions *g*, it will be appropriate to estimate an approximate covariance matrix for $\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}$ using the multivariate delta method. In addition to Becker and Schram's $\hat{\text{C}}\text{ov}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ from GLS (Equation 3), this requires $\mathbf{A}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})$, the $K \times J$ Jacobian matrix of $\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}$ with respect to $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ evaluated at $\hat{\mathbf{M}}_{\boldsymbol{\Theta}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}$, where the sample realization $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}$ (of $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$) is used to approximate $\text{E}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$. The typical element of $\mathbf{A}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})$—for the *k*th component of $\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}$ and the *j*th component of $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$—is

$$\mathbf{A}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})_{kj} = \left.\frac{\partial \overline{\text{M}}_{\boldsymbol{\Gamma}k}}{\partial \hat{\text{M}}_{\boldsymbol{\Theta}j}}\right|_{\hat{\mathbf{M}}_{\boldsymbol{\Theta}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}} . \tag{21}$$

In most practical applications these derivatives will require numerical evaluation (e.g., using difference quotients analogous to Equation 20), at least for $\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}$ due to the MIT's integral. If $\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}$ is obtained by the Monte Carlo strategy described above, then $\mathbf{A}(\hat{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})_{kj}$ may be obtained by shifting the same *M* multivariate-normal variates up or down by $\boldsymbol{\varepsilon}_j$ before computing $\boldsymbol{\gamma}^*_m$ as $g(\boldsymbol{\theta}^*_m + \boldsymbol{\varepsilon}_j)$ or $g(\boldsymbol{\theta}^*_m - \boldsymbol{\varepsilon}_j)$. Analytic derivatives of $\breve{\mathbf{M}}_{\boldsymbol{\Gamma}}$ may be more tractable: Because of the MST2's form (Equation 13), $\mathbf{A}(\breve{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})_{kj}$ depends

on only $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$, the partial derivative of $g_k(\boldsymbol{\Theta})$ with respect to $\Theta_j$, and the $J(J + 1) / 2$ distinct third-order (mixed) partial derivatives of $g_k(\boldsymbol{\Theta})$ with respect to $\Theta_j$ and at most two other components of $\boldsymbol{\Theta}$. Specifically,

$$\mathbf{A}(\breve{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})_{kj} = \left.\frac{\partial g_k(\boldsymbol{\Theta})}{\partial \Theta_j}\right|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}} + \frac{1}{2}\sum_{l=1}^{J}\sum_{m=1}^{J}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}lm}\left.\frac{\partial^3 g_k(\boldsymbol{\Theta})}{\partial\Theta_j\partial\Theta_l\partial\Theta_m}\right|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}} . \quad (22)$$

Regardless of how $\mathbf{A}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ is obtained, if $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ is approximately multivariate normal with covariance matrix $\hat{\mathrm{C}}\mathrm{ov}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$, then we may treat $\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}$ as approximately multivariate normal with covariance matrix

$$\hat{\mathrm{C}}\mathrm{ov}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}) = \mathbf{A}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})\,\hat{\mathrm{C}}\mathrm{ov}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\,\mathbf{A}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}, \hat{\mathbf{M}}_{\boldsymbol{\Theta}})^{\mathrm{T}} . \quad (23)$$

Hence, one may use $\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}$ and $\hat{\mathrm{C}}\mathrm{ov}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}})$ to construct various normal-theory CIs for components of $\mathbf{M}_{\boldsymbol{\Gamma}}$ or linear combinations of these components. One may also test $\mathbf{M}_{\boldsymbol{\Gamma}}$ components or linear combinations thereof using the same general-linear-hypothesis strategy as for tests of $\mathbf{M}_{\boldsymbol{\Theta}}$ (see Section 2.1.1). For some tests it may be feasible and preferable to incorporate the null hypothesis into an alternative covariance matrix, but procedures for doing this are beyond the present scope.

**3.4.2. Bootstrap.** Another strategy for inference about $\mathbf{M}_{\boldsymbol{\Gamma}}$ involves resampling. Specifically, van den Noortgate and Onghena (2005) described four bootstrap methods for inference about $\mathrm{M}_{\Theta}$ or covariate effects for SMDs in the univariate case. On the basis of whether a distributional form is assumed for either $\mathbf{U}_i$ or $\mathbf{E}_i$ in Model 1, each of their methods is deemed parametric or nonparametric. It appears feasible to generalize these

methods to handle inferences on $\mathbf{M}_{\mathbf{T}}$ as implied by a (possibly vector-valued) function of generic multivariate ESs, such as $g(\boldsymbol{\Theta})$. Each of the four methods entails a particular procedure for obtaining $B$ bootstrap samples by generating $\mathbf{t}_b^* = [\mathbf{t}_{b1}^{*\,\mathrm{T}}, \mathbf{t}_{b2}^{*\,\mathrm{T}}, ..., \mathbf{t}_{bI}^{*\,\mathrm{T}}]^{\mathrm{T}}$, $b = 1, 2, ..., B$. Generating $\mathbf{t}_b^*$, in turn, involves resampling $\mathbf{t}_{bi}^*$ by either sampling from $\mathbf{t}$ directly (with replacement) or sampling residuals or ESs (possibly via subject-level data) from parametric or empirical distributions (e.g., based on shrinkage estimators of $\boldsymbol{\theta}_i$). Some of these resampling procedures involve rescaling and "reflating" residuals to compensate for shrinkage. One potential drawback of bootstrap inference is that every resampling procedure requires sample size(s), $\boldsymbol{\Psi}_{\mathbf{T}i}$, or $\mathbf{t}_i$ from all studies, which may not be available in some circumstances.

Below are algorithms for the four bootstrap methods to obtain the pair $\mathbf{t}_{bi}^*$ and $\boldsymbol{\Psi}_{\mathbf{T}bi}^*$, as generalized from Van den Noortgate and Onghena's (2005) algorithms. For each algorithm it is assumed that we begin with the EM-GLS estimates $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\sigma}}_{\boldsymbol{\Theta}}$ (i.e., realizations of estimators $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ of $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$) and any necessary data from every study. In what follows it will be useful to denote the $i$th study's sample size(s) as $\mathbf{n}_i$, which may be a vector (e.g., with proportions or standardized mean differences).

First, the *effect-size* bootstrap is a parametric method that essentially entails executing Model 1 directly with an additional distributional assumption: $\mathbf{U}_i \sim N_J(0, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}})$, which was also made for the EM estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$. The algorithm is as follows:

1. Draw $\boldsymbol{\theta}_{bi}^*$ from $N_J(\hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\sigma}}_{\boldsymbol{\Theta}})$.

2. Compute $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$ from $\boldsymbol{\theta}^*_{bi}$ and $\mathbf{n}_i$.

3. Draw $\mathbf{t}^*_{bi}$ from a parametric family that depends on only $\boldsymbol{\theta}^*_{bi}$ and $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$, such as $N_J(\boldsymbol{\theta}^*_{bi}, \boldsymbol{\Psi}^*_{\mathbf{T}bi})$.

Note that if $\boldsymbol{\Psi}_{\mathbf{T}i}$ depends on only $\mathbf{n}_i$ (and not on $\boldsymbol{\theta}_i$), then $\boldsymbol{\Psi}^*_{\mathbf{T}bi} = \boldsymbol{\Psi}_{\mathbf{T}i}$ so in Step 3 we can draw $\mathbf{t}^*_{bi}$ from $N_J(\boldsymbol{\theta}^*_{bi}, \boldsymbol{\Psi}_{\mathbf{T}i})$. Also, if $\boldsymbol{\Psi}_{\mathbf{T}i}$ depends on only $\mathbf{n}_i$ and for step 3 we use $N_J(\boldsymbol{\theta}^*_{bi}, \boldsymbol{\Psi}_{\mathbf{T}i})$, then we can in one step draw $\boldsymbol{\theta}^*_{bi}$ from $N_J(\hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\sigma}}_{\boldsymbol{\Theta}} + \boldsymbol{\Psi}_{\mathbf{T}i})$.

Second, the *raw-data* bootstrap is another parametric method similar to the effect-size bootstrap, except that instead of sampling $\mathbf{t}^*_{bi}$ from a parametric distribution we compute it from a sample of subject-level data as would be done in a primary study. The algorithm is as follows:

1. Draw $\boldsymbol{\theta}^*_{bi}$ from $N_J(\hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\sigma}}_{\boldsymbol{\Theta}})$.

2. Sample raw data $\mathbf{Y}_i$ from a distribution that depends on only $\boldsymbol{\theta}^*_{bi}$ and $\mathbf{n}_i$.

3. Compute $\mathbf{t}^*_{bi}$ from $\mathbf{Y}_i$ and $\mathbf{n}_i$.

How raw data are sampled depends on the metric of $\boldsymbol{\Theta}$. For example, if $\boldsymbol{\Theta}$ were a matrix of Fisher *z*-transforms then $\mathbf{Y}_i$ might be a simple random sample of size $\mathbf{n}_i$ from a multivariate normal distribution whose correlation matrix corresponds to $\tanh \boldsymbol{\theta}^*_{bi}$, or we could instead draw one sample from the appropriate Wishart distribution. Van den Noortgate and Onghena (2005) did not explicitly describe how to obtain $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$ for this method, but one may infer from their explanation of the effect-size and raw-data methods

that $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$ could be computed from $\boldsymbol{\theta}^*_{bi}$ and $\mathbf{n}_i$. It may also be reasonable to compute $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$ from all studies' $\mathbf{t}^*_{bi}$ and $\mathbf{n}_i$ to mimic the computation of $\boldsymbol{\Psi}_{\mathbf{T}i}$.

Third, the *error* bootstrap is a nonparametric method whose several steps may obscure its relatively simple underlying idea: Instead of assuming parametric distributions for the errors $\mathbf{U}_i$ and $\mathbf{E}_i$, we compute residuals (i.e., estimated errors) for each study; we then obtain resampled errors by sampling with replacement (SWR) from these residuals and use these errors to compute a resampled $\mathbf{t}^*_{bi}$. The algorithm below involves using covariance matrices for residuals to "reflate" them—to compensate for shrinkage (Carpenter, Goldstein, & Rasbash, 2003)—and standardizing within-study residuals to have a similar mean and covariance matrix for all studies (then later unstandardizing them). Standardization and reflation are accomplished using the Cholesky decomposition: For covariance matrix **S**, let **C**(**S**) denote the lower-triangle matrix such that $\mathbf{C}(\mathbf{S})\mathbf{C}(\mathbf{S})^{\mathrm{T}} = \mathbf{S}$. In the following algorithm, Step 1 estimates ES parameters, Steps 2–5 resample between-studies errors, and Steps 7–13 resample within-studies errors.

1. Compute $\hat{\boldsymbol{\theta}}_i = (\boldsymbol{\Psi}^{-1}_{\mathbf{T}i} + \tilde{\boldsymbol{\sigma}}^{-1}_{\boldsymbol{\Theta}})^{-1}(\boldsymbol{\Psi}^{-1}_{\mathbf{T}i}\mathbf{t}_i + \tilde{\boldsymbol{\sigma}}^{-1}_{\boldsymbol{\Theta}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}})$.

2. Compute $\hat{\mathbf{u}}_i = \hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}}$, and collect these in the $J \times I$ matrix $\hat{\mathbf{u}} = [\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_I]$.

3. Estimate the covariance matrix of $\hat{\mathbf{u}}_i$ as $\mathbf{S}_{\hat{\mathbf{u}}} = I^{-1}\hat{\mathbf{u}}\hat{\mathbf{u}}^{\mathrm{T}}$.

4. Draw $\hat{\mathbf{u}}_{b(i)}$ by SWR from $\{\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_I\}$.

5. Compute $\mathbf{u}^*_{bi} = \mathbf{C}(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{\Theta}})\mathbf{C}(\mathbf{S}_{\hat{\mathbf{u}}})^{-1}\hat{\mathbf{u}}_{b(i)}$.

6. Compute $\boldsymbol{\theta}^*_{bi} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\Theta}} + \mathbf{u}^*_{bi}$.

7. Compute $\hat{\mathbf{e}}_i = \mathbf{t}_i - \hat{\boldsymbol{\theta}}_i$.

8. Compute $\hat{\mathbf{z}}_i = \mathbf{C}(\boldsymbol{\Psi}_{\mathbf{T}i})^{-1}\hat{\mathbf{e}}_i$, and collect these in the $J \times I$ matrix $\hat{\mathbf{z}} = [\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_I]$.

9. Estimate the covariance matrix of $\hat{\mathbf{z}}_i$ as $\mathbf{S}_{\hat{\mathbf{z}}} = I^{-1}\hat{\mathbf{z}}\hat{\mathbf{z}}^{\mathrm{T}}$.

10. Draw $\hat{\mathbf{z}}_{b(i)}$ by SWR from $\{\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_I\}$.

11. Compute $\mathbf{z}^*_{bi} = \mathbf{C}(\mathbf{S}_{\hat{\mathbf{z}}})^{-1}\hat{\mathbf{z}}_{b(i)}$.

12. Compute $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$ from $\boldsymbol{\theta}^*_{bi}$ and $\mathbf{n}_i$.

13. Compute $\mathbf{e}^*_{bi} = \mathbf{C}(\boldsymbol{\Psi}^*_{\mathbf{T}bi})\mathbf{z}^*_{bi}$.

14. Compute $\mathbf{t}^*_{bi} = \boldsymbol{\theta}^*_{bi} + \mathbf{e}^*_{bi}$.

This procedure is not fully nonparametric. For instance, Step 12 typically relies on certain distributional assumptions. Note that if $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$ depends on only $\mathbf{n}_i$ so that $\boldsymbol{\Psi}^*_{\mathbf{T}bi} = \boldsymbol{\Psi}_{\mathbf{T}i}$, then Step 12 may be omitted and the unstandardization in Step 13 is the inverse transformation of the standardization in Step 8. Also, for computational purposes one could reflate resampled residuals (e.g., Steps 3–5, Steps 9–11) for all studies at once, such as with $\mathbf{C}(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{\Theta}})\mathbf{C}(\mathbf{S}_{\hat{\mathbf{u}}})^{-1}[\hat{\mathbf{u}}_{(1)}, \hat{\mathbf{u}}_{(2)}, \ldots, \hat{\mathbf{u}}_{(I)}]$, or one might reflate before resampling, such as by SWR from the elements of $\mathbf{C}(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{\Theta}})\mathbf{C}(\mathbf{S}_{\hat{\mathbf{u}}})^{-1}\hat{\mathbf{u}}$.

Fourth, the *cases* bootstrap is a nonparametric method that treats the pairs $[\mathbf{t}_i, \boldsymbol{\Psi}_{\mathbf{T}i}]$ as random instead of treating $\boldsymbol{\Psi}_{\mathbf{T}i}$ (and $\mathbf{n}_i$) as fixed and known. The simple algorithm is as follows: Draw $[\mathbf{t}^*_{bi}, \boldsymbol{\Psi}^*_{\mathbf{T}bi}]$ by SWR from $\{[\mathbf{t}_1, \boldsymbol{\Psi}_{\mathbf{T}1}], [\mathbf{t}_2, \boldsymbol{\Psi}_{\mathbf{T}2}], \ldots, [\mathbf{t}_I, \boldsymbol{\Psi}_{\mathbf{T}I}]\}$. When $\boldsymbol{\Psi}_{\mathbf{T}i}$

depends on $\boldsymbol{\theta}_i$ in addition to $\mathbf{n}_i$, one could alternatively draw $[\mathbf{t}^*_{bi}, \mathbf{n}^*_{bi}]$ by SWR from $\{[\mathbf{t}_1, \mathbf{n}_1], [\mathbf{t}_2, \mathbf{n}_2], \ldots, [\mathbf{t}_I, \mathbf{n}_I]\}$ then compute $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$ from all studies' $\mathbf{t}^*_{bi}$ and $\mathbf{n}^*_{bi}$ to mimic the computation of $\boldsymbol{\Psi}_{\mathbf{T}i}$.

Regardless of which bootstrap resampling method is used, for the *b*th bootstrap sample one simply applies Becker and Schram's (1994) EM-GLS approach to $\mathbf{t}^*_{bi}$ and $\boldsymbol{\Psi}^*_{\mathbf{T}bi}$ (instead of $\mathbf{t}_i$ and $\boldsymbol{\Psi}_{\mathbf{T}i}$) to obtain $\hat{\boldsymbol{\mu}}^*_{\boldsymbol{\Theta}b}$ and $\tilde{\boldsymbol{\sigma}}^*_{\boldsymbol{\Theta}b}$ (i.e., realizations of $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$) and then obtains $\bar{\boldsymbol{\mu}}^*_{\boldsymbol{\Gamma}b}$ from these estimates by either of the MIT or MTS2 estimators. Repeating this for all *B* bootstrap samples yields a set of bootstrap replicates that mimic the sampling distribution of $\bar{\mathbf{M}}_{\boldsymbol{\Gamma}}$ and may be used to make inferences about $\mathbf{M}_{\boldsymbol{\Gamma}}$, such as confidence regions or tests based on percentiles or a covariance matrix. For example, a bootstrap covariance matrix for $\bar{\mathbf{M}}_{\boldsymbol{\Gamma}}$ may be estimated using $\bar{\bar{\boldsymbol{\mu}}}^*_{\boldsymbol{\Gamma}} = B^{-1}\sum_{b=1}^{B}\bar{\boldsymbol{\mu}}^*_{\boldsymbol{\Gamma}b}$:

$$\hat{\mathrm{Cov}}_B(\bar{\mathbf{M}}_{\boldsymbol{\Gamma}}) = (B-1)^{-1}\mathbf{F}\mathbf{F}^{\mathrm{T}}, \tag{24}$$

where $\mathbf{F} = [\bar{\boldsymbol{\mu}}^*_{\boldsymbol{\Gamma}1} - \bar{\bar{\boldsymbol{\mu}}}^*_{\boldsymbol{\Gamma}}, \bar{\boldsymbol{\mu}}^*_{\boldsymbol{\Gamma}2} - \bar{\bar{\boldsymbol{\mu}}}^*_{\boldsymbol{\Gamma}}, \ldots, \bar{\boldsymbol{\mu}}^*_{\boldsymbol{\Gamma}B} - \bar{\bar{\boldsymbol{\mu}}}^*_{\boldsymbol{\Gamma}}]$ is a $K \times B$ matrix of deviations.

Bootstrap inference is more computationally intensive than delta-method inference, primarily because the bootstrap entails repeating EM-GLS procedures *B* times, and some bootstrap resampling methods are susceptible to numerical problems (e.g., improper covariance matrices in the error bootstrap). Bootstrapping can, however, be applied to more functions than the delta method. Also, bootstrap inference may be less sensitive to violations of certain distributional assumptions, and it may better incorporate

uncertainty about $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$. Although obtaining bias-corrected estimators of $\mathbf{M}_{\boldsymbol{\Gamma}}$ from the bootstrap samples may also be feasible, these are beyond the present scope.

### 3.5. Special Cases

In some circumstances, such as with certain types of functions (e.g., affine, componentwise, scalar), for homogeneous fixed-effects models, or when $J = K = 1$, simplifications to the above estimators or inference procedures may arise. In some of these circumstances alternative procedures may be preferable, such as direct application of the MIT or MTS2 transformations to endpoints of CIs for components of $\mathbf{M}_{\boldsymbol{\Theta}}$ in the case of a componentwise function. Below I remark briefly on two particular circumstances of potential interest.

**3.5.1. Homogeneous fixed effects.** First, consider the homogeneous fixed-effects model mentioned in Sections 2.1.1 and 2.3.2, for whose only parameter, $\boldsymbol{\theta}$, Equation 5 gives a GLS estimator. In this case $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ is undefined because $\boldsymbol{\Theta}$ is not random (loosely speaking, $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} = \mathbf{0}$), so the EM algorithm is unnecessary. Regarding $\boldsymbol{\gamma} \equiv g(\boldsymbol{\theta})$, the MIT and MTS2 estimators both simply yield $\hat{\boldsymbol{\gamma}} = g(\hat{\boldsymbol{\theta}})$, and $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ is undefined. Application of the delta method is simplified because the Jacobian matrix of $\hat{\boldsymbol{\gamma}}$ with respect to $\hat{\boldsymbol{\theta}}$ involves only partial derivatives:

$$\hat{\mathrm{Cov}}(\hat{\boldsymbol{\gamma}}) = \mathbf{A}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}})\, \hat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}})\, \mathbf{A}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}})^{\mathrm{T}}, \tag{25}$$

where

$$\mathbf{A}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}})_{kj} = \left. \frac{\partial g_k(\boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\theta}}}. \tag{25}$$

Furthermore, bootstrap resampling does not involve between-studies error. For instance, for the effect-size and raw-data bootstrap algorithms one can omit Steps 1 and 2 and use $\boldsymbol{\theta}^*_{bi} = \hat{\boldsymbol{\theta}}$ in Step 3, and for the error bootstrap algorithm one can omit Steps 1–6 and use $\boldsymbol{\theta}^*_{bi} = \hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}$ in Steps 7 and 12.

**3.5.1. Affine transformation.** Second, suppose the function *g* is affine, so that $\boldsymbol{\Gamma} \equiv g(\boldsymbol{\Theta}) = \mathbf{W}\boldsymbol{\Theta} + \mathbf{v}$, where $\mathbf{W}$ is a $K \times J$ matrix of known coefficients, and $\mathbf{v}$ is a column vector of $J$ known values. Then the IT and TS2 estimators both simply yield $\overline{\mathbf{M}}_{\boldsymbol{\Gamma}} = \mathbf{W}\hat{\mathbf{M}}_{\boldsymbol{\Theta}} + \mathbf{v}$ and $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}} = \mathbf{W}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}\mathbf{W}^{\mathrm{T}}$, where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}} \in \{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}, \breve{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}\}$. Also, the delta method yields $\hat{\mathrm{C}}\mathrm{ov}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}) = \mathbf{W}\,\hat{\mathrm{C}}\mathrm{ov}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\,\mathbf{W}^{\mathrm{T}}$, which also follows from noting that $\mathrm{Cov}(\overline{\mathbf{M}}_{\boldsymbol{\Gamma}}) = \mathrm{Cov}(\mathbf{W}\hat{\mathbf{M}}_{\boldsymbol{\Theta}} + \mathbf{v}) = \mathbf{W}\mathrm{Cov}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})\mathbf{W}^{\mathrm{T}}$. The bootstrap resampling procedures do not simplify markedly.

# 4. MONTE CARLO STUDIES OF PROPOSED TECHNIQUES

A second major aim of the present thesis is to evaluate the techniques proposed in Chapter 3. My main interest is in point estimators' accuracy and efficiency and confidence sets' coverage performance. As illustrated in the description of the main simulation study below, Study 1, simulating even one situation requires several choices about the data and meta-analytic methods. Furthermore, some of the methods are computationally slow (e.g., MIT and CIT by simulation with bootstrap inference, especially with large $B$). Hence, it was not practically feasible to evaluate these procedures thoroughly under a wide range of plausible conditions. Study 1 offers a preliminary Monte Carlo examination of the proposed point-estimation and delta-method inference techniques for a non-trivial function $g$ in a limited number of realistic conditions. Study 2 is a follow-up Monte Carlo investigation focused mainly on bootstrap inference in a subset of conditions from Study 1, with some comparisons to delta-method inference. In the following sections I describe each study's method and report on its results, followed by a summary highlighting each study's main findings. It will often be convenient to refer collectively to the MIT and CIT estimators as IT estimators and to the MTS2 and CTS2 estimators as TS2 estimators.

## 4.1. Study 1: Method

The primary purpose of this first Monte Carlo study was to assess the proposed point estimators' and delta-method CIs' performance under fairly realistic conditions when most or all assumptions of these procedures are satisfied. To this end, data were generated to conform to Model 1, with $\boldsymbol{\Theta}$ and $\mathbf{T}_i$ both multivariate normal and $\boldsymbol{\Psi}_{\mathbf{T}i}$

known. In each of several conditions defined by the number of studies as well as features of $\mathbf{M}_{\Theta}$, $\mathbf{\Sigma}_{\Theta}$, and $\mathbf{\Psi}_{\mathbf{T}i}$, a feasible but informative number of replications (i.e., simulated meta-analyses) were run. Each such meta-analysis entailed applying proposed techniques to the simulated data to obtain estimates of and inferences about (components of) a judiciously selected vector-valued function of $\Theta$, and these results were used to estimate properties of point estimators and CIs. Details of this Monte Carlo study are given below, some of which were arrived at by considering run time, sampling error for estimates of evaluation criteria, and other information from pilot simulations.

**4.1.1. Design conditions.** Generating data that conform to Model 1 requires specifying $J$, $\mathbf{M}_{\Theta}$, $\mathbf{\Sigma}_{\Theta}$, $I$, and each simulated study's $\mathbf{\Psi}_{\mathbf{T}i}$. Enormously many options exist, even if they were limited to typical values for a specific ES metric (e.g., correlations) in an actual research domain. Furthermore, little empirical evidence about typical values is available for most of these quantities, largely because multivariate meta-analysis has been used rarely to date—even in applications where it may have been more appropriate than univariate procedures. Hence, my choices were informed by my previous work with multivariate meta-analysis, most of which has involved simulations of correlation matrices (e.g., Hafdahl, 2001, 2004, 2007, 2008) as well as several real-data applications involving correlation matrices (e.g., Hafdahl, 2001, 2009b, 2009d) or multiple-treatment SMDs (e.g., Conn, Hafdahl, Brown, & Brown, 2008; Conn, Hafdahl, Cooper, Brown, & Lusk, 2009; Conn, Hafdahl, Cooper, Ruppar, Mehr, & Russell, 2009; Conn et al., 2008; Conn, Hafdahl, Porock, McDaniel, & Nielsen, 2006).

To simplify design choices at some cost in generalizability, I fixed $J = 6$ and set $\mathbf{M}_{\Theta} = \eta \mathbf{1}_J$, a column vector of $J$ $\eta$s, with $\eta$ varied systematically as a design factor. I also varied $\mathbf{\Sigma}_{\Theta}$ and $\mathbf{\Psi}_{\mathbf{T}i}$ systematically based on the following simple structures: $\xi_1 \equiv \Sigma_{\Theta jj}$ $\forall\, j$ and $\xi_2 \equiv \text{Corr}(\Theta_j, \Theta_l) = \Sigma_{\Theta jl} / \xi_1 \;\forall\, j, l$ as variances and correlations for $\mathbf{\Sigma}_{\Theta}$, and $\Psi_{\mathbf{T}ijj} = \Psi_{\mathbf{T}ill} \;\forall\, i, j, l$ and $\phi \equiv \text{Corr}(\text{T}_{ij}, \text{T}_{il}) = \Psi_{\mathbf{T}ijl} / \Psi_{\mathbf{T}ijj} \;\forall\, i, j, l$ as variances and correlations for $\mathbf{\Psi}_{\mathbf{T}i}$. Hence, each of $\mathbf{\Sigma}_{\Theta}$ and $\mathbf{\Psi}_{\mathbf{T}i}$ was compound symmetric with homogeneous variance. Furthermore, to mimic the usual dependence of $\Psi_{\mathbf{T}ijj}$ on study $i$'s sample size, $n_i \in \mathbb{N}^*$, I used $\Psi_{\mathbf{T}ijj} = 1 / n_i \;\forall\, i, j$, so that $\Psi_{\mathbf{T}ijl} = \phi / n_i \;\forall\, i, j \neq l$. This specific relation between an ES's (approximate asymptotic) conditional variance and sample size, $\Psi_{\mathbf{T}ijj} = 1 / n_i$, indeed holds for some realistic ESs: a variance-stabilizing transformation of Hedges's $d$ (Hedges & Olkin, 1985, p. 88) and the arcsine transformation of a proportion. It is also nearly true for a Fisher-$z$ correlation, where $\Psi_{\mathbf{Z}ijj} = 1 / (n_i - 3)$, and it holds for special cases of a Pearson-$r$ correlation ($\rho_i = 0$) and a risk difference ($\pi_{i1} = \pi_{i2} = 1 / 2$ and $n_{i1} = n_{i2} = n_i / 2$). Compound symmetry for $\mathbf{\Psi}_{\mathbf{T}i}$ was not intended to represent any particular ES, and in practice $\mathbf{\Psi}_{\mathbf{T}i}$ often depends on $\boldsymbol{\theta}_i$.

The factorial simulation design consisted of $3^5 = 243$ conditions, each defined as a quintuple of five factors: three values each of the mean, $\eta \in \{0.0, 0.4, 0.8\}$; between-studies variance, $\xi_1 \in \{0.05^2, 0.10^2, 0.20^2\}$; between-studies correlation, $\xi_2 \in \{-.1, .1, .5\}$; within-study correlation, $\phi \in \{-.1, .1, .5\}$; and number of studies, $I \in \{10, 20, 40\}$. For most of the factors these levels cover fairly wide ranges of realistic values. A

possible exception is $I$, whose values seem somewhat small but are meant to reduce computational burden.

To mimic realistic sample sizes, I sampled $n_i$ from a positively skewed distribution with $\bar{n} = 100$ as its approximate expected value. To avoid idiosyncrasies of any particular sample of $I$ $n_i$ values, a new sample was drawn for each replication, which introduces an additional random component into the expectations and probabilities used for evaluation criteria. Although meta-analysis simulations often vary $\bar{n}$ as a factor, with the present ideal ESs this was not of sufficient interest to warrant the cost of a larger design: Unlike many real ESs, $\mathbf{t}_i$ was generated such that the shape of $\mathbf{T}_i$'s sampling distribution does not depend on $n_i$, so the most important consequence of varying $\bar{n}$ would be to vary the ratio of $\Psi_{\mathbf{T}ijj}$ to $\Sigma_{\mathbf{\Theta}jj}$, which already varies with $\xi_1 = \Sigma_{\mathbf{\Theta}jj}$.

**4.1.2. Data generation.** In each condition I generated 500 independent meta-analytic data sets and analyzed each of them using the several procedures specified below. Each such replication's data consisted of $I$ independent pairs $[n_i, \mathbf{t}_i]$, with study $i$'s $n_i$ and $\mathbf{t}_i$ generated independently as follows:

1. Draw $x_i$ from $\chi^2(3)$, and set $n_i = \langle [(x_i - 3) / \sqrt{2(3)}\,](\bar{n}\ / 2) + \bar{n}\,\rangle$, where $\langle a \rangle$ denotes the integer nearest $a$.

2. Draw $\boldsymbol{\theta}_i$ from $N_6(\mathbf{M}_{\mathbf{\Theta}}, \mathbf{\Sigma}_{\mathbf{\Theta}})$.

3. Draw $\mathbf{t}_i$ from $N_6(\boldsymbol{\theta}_i, \mathbf{\Psi}_{\mathbf{T}i})$.

Step 1 yields $n_i$ from a positively skewed distribution with approximate mean $\bar{n}$ and variance $(\bar{n}\ / 2)^2$ and $n_i > .38\,\bar{n}$. In Step 3 $n_i$ and $\phi$ determine $\mathbf{\Psi}_{\mathbf{T}i}$. (Steps 2 and 3 are

equivalent to drawing $\mathbf{t}_i$ from $N_6[\mathbf{M}_\Theta, \mathbf{\Sigma}_\Theta + \mathbf{\Psi}_{\mathbf{T}i}]$.) The resulting expected ratio of between-studies to "total" variance, $E[\Sigma_{\Theta jj} / (\Sigma_{\Theta jj} + \Psi_{\mathbf{T}ijj})]$, ranged from .193 with $\xi_1 = \Sigma_{\Theta jj} = 0.05^2$ to .773 with $\xi_1 = 0.20^2$, which encompass Hedges and Pigott's (2001) generic guidelines for small (ratio of .25) and large (.50) between-studies variances.

**4.1.3. Meta-analytic procedures.** Each replication's simulated meta-analytic data were analyzed by first applying Becker and Schram's (1994) EM-GLS procedures to obtain $\hat{\mathbf{M}}_\Theta$, $\widetilde{\mathbf{\Sigma}}_\Theta$, and $\hat{\text{Cov}}(\hat{\mathbf{M}}_\Theta)$. Their $Q_\Theta$ statistic (Equation 4) for testing $\mathbf{\Sigma}_\Theta$ was disregarded. Although analyses in the $\Theta$ metric were not of primary interest, these quantities were used to construct CIs for each component of $\mathbf{M}_\Theta$, in part to verify the simulation procedures and to serve as a baseline for assessing analyses in the $\Gamma$ metric.

More important, the proposed techniques for estimates and inferences in the $\Gamma \equiv g(\Theta)$ metric were applied to $\hat{\mathbf{M}}_\Theta$, $\widetilde{\mathbf{\Sigma}}_\Theta$, and $\hat{\text{Cov}}(\hat{\mathbf{M}}_\Theta)$, where $g$ was the following vector-valued function: $g(\Theta) = [\Theta_1^2, \Theta_1\Theta_2, 1 / (e^{-\Theta_3} + 1), e^{\Theta_4 - \Theta_3}, 2\sqrt{2}\sinh(\Theta_5 / \sqrt{2}), \tanh\Theta_6]^{\mathrm{T}}$. The $K = 6$ components of $g$ are basic nonlinear algebraic and transcendental functions likely to constitute realistic composite functions: $g_1(\Theta) = \Theta_1^2$ and $g_2(\Theta) = \Theta_1\Theta_2$ are quadratic, $g_3(\Theta) = 1 / (e^{-\Theta_3} + 1)$ would transform a logit to a proportion, $g_4(\Theta) = e^{\Theta_4 - \Theta_3}$ would transform two logits to an odds ratio, $g_5(\Theta) = 2\sqrt{2}\sinh(\Theta_5 / \sqrt{2})$ is the inverse variance-stabilizing transformation of a SMD in the balanced case (Hedges & Olkin, 1985, p. 88), and $g_6(\Theta) = \tanh\Theta_6$ is the inverse of Fisher's $z$-transformation.

Point estimators of $\mathbf{M}_\Gamma$ and $\mathbf{\Sigma}_\Gamma$ were obtained by each of the IT and TS2 estimation approaches in Sections 3.2 and 3.3. For IT estimators I used Monte Carlo integration with $M = 10{,}000$ samples, and for IT estimators I used numerical derivatives with the increment $\varepsilon$ chosen adaptively. Inferences about $\mathbf{M}_\Gamma$ were obtained by applying the delta method in Section 3.4.1 with numerical derivatives for both the IT and TS2 estimators. The main inference task of interest was a 95% CI for each component of $\mathbf{M}_\Gamma$, using standard-normal quantiles with no adjustment for simultaneous inference. Hence, in each condition these analyses yielded two point estimates of $\mathbf{M}_\Gamma$ and $\mathbf{\Sigma}_\Gamma$ and two sets of componentwise CIs for $\mathbf{M}_\Gamma$. (I also computed 90% and 99% CIs as well as Student-$t$ CIs at all three confidence levels using $I - 1$ as degrees of freedom, for all of which results are available upon request but not presented below.)

**4.1.4. Evaluation criteria.** Regarding the performance of point estimators of $\mathbf{M}_\Theta$, $\mathbf{M}_\Gamma$, $\mathbf{\Sigma}_\Theta$, and $\mathbf{\Sigma}_\Gamma$, my primary interest was in bias and mean squared error (MSE) for estimators of each component of $\mathbf{M}_\Theta$ and $\mathbf{M}_\Gamma$ and each diagonal element and $\mathbf{\Sigma}_\Theta$ and $\mathbf{\Sigma}_\Gamma$. As an example, for the IT estimator of $M_{\Gamma k}$, $\text{Bias}(\hat{M}_{\Gamma k}) \equiv \text{E}(\hat{M}_{\Gamma k} - M_{\Gamma k})$ and $\text{MSE}(\hat{M}_{\Gamma k}) \equiv \text{E}[(\hat{M}_{\Gamma k} - M_{\Gamma k})^2]$. Properties of $\hat{\mathbf{M}}_\Theta$, $\hat{\mathbf{M}}_\Gamma$, and $\breve{\mathbf{M}}_\Gamma$ as vectors or of $\tilde{\mathbf{\Sigma}}_\Theta$, $\hat{\mathbf{\Sigma}}_\Gamma$, and $\breve{\mathbf{\Sigma}}_\Gamma$ as matrices may be of interest for some purposes, but here I consider only scalar properties separately for each component of $\Theta$ and $\Gamma$. As for inference about $\mathbf{M}_\Gamma$, I am most interested in the coverage probability of each component's CI versus the nominal confidence level. Particular transformations of bias, MSE, and CI coverage probability to facilitate interpretation are described along with the presentation of results.

### 4.2. Study 1: Results

In this section I report selected results from the above Monte Carlo study, with ideal generic ESs as data, the delta method for inferences, and five design factors. Separate subsections are devoted to point estimators of means from $\mathbf{M}_\Theta$ and $\mathbf{M}_\Gamma$, point estimators of variances from $\mathbf{\Sigma}_\Theta$ and $\mathbf{\Sigma}_\Gamma$, and delta-method CIs for means. To keep the scope manageable and emphasize findings most relevant to anticipated applications, the presentation of results is simplified in various ways, such as by displaying summaries of results over conditions. For readers who wish to examine the simulation results independently, SAS/IML programs (Version 9.1) and data are available from the author.

**4.2.1. Point estimators of means.** In each of the 243 conditions the simulation yielded empirical estimates of bias, variance, and MSE for the EM-GLS estimator of $M_{\Theta j}$ and the IT and TS2 estimators of $M_{\Gamma k}$. Table 1 is typical of tables used to summarize simulation results in this thesis: For some outcome of interest, selected percentiles across conditions are shown separately for each component of $\Theta$ or $\Gamma$. (The 1st, 2nd, 98th, and 99th percentiles are sometimes omitted when summarizing substantially fewer than 243 conditions.) Table 1, in particular, shows percentiles across all conditions of 100 times standardized bias for each element of $\hat{\mathbf{M}}_\Theta$, $\mathrm{SBias}(\hat{M}_{\Theta j}) = \mathrm{Bias}(\hat{M}_{\Theta j}) / \mathrm{SE}(\hat{M}_{\Theta j})$, where $\mathrm{SE}(\hat{M}_{\Theta j}) = \sqrt{\mathrm{Var}(\hat{M}_{\Theta j})}$ is the estimator's empirical standard error. Standardized bias retains the sign of bias and rescales it relative to sampling error to facilitate judging its magnitude for comparisons among conditions, methods, ES metrics, function

components, and so on. Squaring standardized bias yields a relative difference between MSE and variance, or the "inflation" of MSE over variance due to bias:

$$[\text{SBias}(\bullet)]^2 = [\text{Bias}(\bullet)]^2 / [\text{SE}(\bullet)]^2 = [\text{MSE}(\bullet) - \text{Var}(\bullet)] / \text{Var}(\bullet) .$$

Furthermore, the proportion of MSE due to (squared) bias is just $\{1 + [\text{SBias}(\bullet)]^{-2}\}^{-1} = [\text{Bias}(\bullet)]^2 / \text{MSE}(\bullet)$. (Dividing bias by the estimand would yield another re-expression of bias, but this fails in one third of the present conditions where $M_{\Theta j} = \eta = 0$.)

Although my primary interest is in the proposed estimators of $\mathbf{M}_\Gamma$, comparing them to their counterparts for $\mathbf{M}_\Theta$ addresses the impact of the proposed transformation

Table 1
*Percentiles of Standardized Bias for Estimators of* $M_{\Theta j}$

| Percentile | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ |
|---|---|---|---|---|---|---|
| 100 | 13 | 12 | 11 | 13 | 13 | 13 |
| 99 | 10 | 10 | 11 | 8 | 10 | 11 |
| 98 | 9 | 9 | 9 | 8 | 9 | 10 |
| 95 | 7 | 8 | 8 | 7 | 7 | 8 |
| 90 | 6 | 6 | 6 | 6 | 6 | 6 |
| 80 | 4 | 4 | 4 | 4 | 4 | 4 |
| 75 | 4 | 4 | 3 | 3 | 3 | 3 |
| 50 | 1 | 0 | 0 | 0 | 0 | 0 |
| 25 | -2 | -2 | -3 | -3 | -3 | -3 |
| 20 | -4 | -3 | -5 | -3 | -4 | -3 |
| 10 | -6 | -6 | -6 | -6 | -6 | -5 |
| 5 | -7 | -8 | -8 | -8 | -8 | -8 |
| 2 | -9 | -10 | -10 | -9 | -9 | -9 |
| 1 | -10 | -10 | -11 | -9 | -11 | -10 |
| 0 | -11 | -13 | -14 | -13 | -13 | -11 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $100\text{Bias}(\hat{M}_{\Theta j}) / \sqrt{\text{Var}(\hat{M}_{\Theta j})}$ .

techniques. That is, the performance of the EM-GLS estimators $\hat{\mathbf{M}}_{\Theta}$ and $\tilde{\mathbf{\Sigma}}_{\Theta}$ serves as a useful baseline against which to evaluate the performance of transformations of these estimators used to estimate $\mathbf{M}_{\Gamma}$ and $\mathbf{\Sigma}_{\Gamma}$. As Table 1 shows, standardized bias for $\hat{\mathrm{M}}_{\Theta j}$ was often less than 0.05 in absolute value and always less than 0.15, so that MSE would be at most $100(0.15^2) = 2.25\%$ larger than variance. This negligibly small standardized bias showed no apparent association with any of the five simulation design factors. (Any variation among components of $\Theta$ in standardized bias or other properties of $\hat{\mathbf{M}}_{\Theta}$ is due entirely to Monte Carlo sampling error.)

Table 2 parallels Table 1 but for components of $\mathbf{M}_{\Gamma}$—separately for IT and TS2—instead of $\mathbf{M}_{\Theta}$. It will occasionally be convenient to refer simultaneously to the IT and TS2 estimators of $\mathbf{M}_{\Gamma}$, $\hat{\mathbf{M}}_{\Gamma}$ and $\breve{\mathbf{M}}_{\Gamma}$, respectively, as $\overline{\mathbf{M}}_{\Gamma} \in \{\hat{\mathbf{M}}_{\Gamma}, \breve{\mathbf{M}}_{\Gamma}\}$. With exceptions for $\Gamma_1$ and $\Gamma_2$ in certain isolated conditions, which I address below, standardized bias was at worst only slightly larger for estimators of $\mathbf{M}_{\Gamma}$ than for estimators of $\mathbf{M}_{\Theta}$: typically less than 0.10 in absolute value and rarely more than 0.20. That standardized bias was not markedly larger for $\overline{\mathbf{M}}_{\Gamma}$ than $\hat{\mathbf{M}}_{\Theta}$ is encouraging, especially given that $\overline{\mathbf{M}}_{\Gamma}$ depends on $\tilde{\mathbf{\Sigma}}_{\Theta}$, which is shown later to exhibit considerable bias in some conditions.

As for associations with design conditions, standardized bias for $\overline{\mathbf{M}}_{\Gamma}$ tended to be more positive than negative for $\overline{\mathrm{M}}_{\Gamma 4}$ and vice versa for $\overline{\mathrm{M}}_{\Gamma 6}$, and for both of these components it was most pronounced with small $\xi_1 = \Sigma_{\Theta jj}$ (i.e., small between-studies

variance component). Most prominent, however, was the substantial positive standardized bias for $\overline{M}_{\Gamma 1}$ and $\overline{M}_{\Gamma 2}$, which was sometimes over 0.50 and reached 0.83 for $\overline{M}_{\Gamma 1}$ and 0.74 for $\overline{M}_{\Gamma 2}$. These large values occurred almost exclusively when $\eta = M_{\Theta j} = 0$ and $\xi_1 \leq 0.10^2$. To highlight this association with $\eta$ and $\xi_1$, Table 3 shows percentiles of standardized bias for only $\overline{M}_{\Gamma 1}$ and $\overline{M}_{\Gamma 2}$, separately for conditions with $\eta = 0$ and $\xi_1 \leq 0.10^2$ versus all other combinations of $\eta$ and $\xi_1$.

Table 2
*Percentiles of Standardized Bias for Estimators of* $M_{\Gamma k}$

| | Integral transformation (IT) | | | | | | Taylor series, order 2 (TS2) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| 100 | 83 | 74 | 11 | 17 | 15 | 12 | 83 | 74 | 10 | 17 | 15 | 11 |
| 99 | 81 | 65 | 10 | 15 | 10 | 11 | 81 | 65 | 10 | 15 | 10 | 9 |
| 98 | 79 | 52 | 9 | 14 | 10 | 10 | 79 | 52 | 8 | 14 | 10 | 8 |
| 95 | 69 | 30 | 7 | 11 | 7 | 7 | 69 | 30 | 7 | 11 | 7 | 7 |
| 90 | 54 | 13 | 6 | 9 | 6 | 5 | 54 | 13 | 5 | 9 | 6 | 4 |
| 80 | 16 | 8 | 4 | 6 | 5 | 3 | 16 | 7 | 4 | 6 | 4 | 2 |
| 75 | 14 | 7 | 3 | 6 | 4 | 2 | 13 | 7 | 3 | 6 | 4 | 1 |
| 50 | 7 | 2 | 0 | 2 | 1 | -1 | 6 | 2 | 0 | 2 | 1 | -2 |
| 25 | 2 | -2 | -3 | -1 | -3 | -4 | 2 | -2 | -3 | -1 | -3 | -5 |
| 20 | 1 | -3 | -5 | -1 | -4 | -5 | 1 | -3 | -5 | -2 | -4 | -5 |
| 10 | -1 | -7 | -7 | -3 | -6 | -7 | -2 | -7 | -7 | -4 | -5 | -8 |
| 5 | -3 | -9 | -8 | -6 | -7 | -10 | -3 | -9 | -8 | -6 | -7 | -11 |
| 2 | -5 | -13 | -10 | -8 | -9 | -12 | -5 | -13 | -10 | -8 | -9 | -13 |
| 1 | -6 | -19 | -11 | -9 | -10 | -15 | -6 | -19 | -11 | -9 | -10 | -15 |
| 0 | -9 | -22 | -16 | -13 | -14 | -16 | -9 | -22 | -16 | -13 | -14 | -16 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $100\text{Bias}(\hat{M}_{\Gamma k}) / \sqrt{\text{Var}(\hat{M}_{\Gamma k})}$ for IT or $100\text{Bias}(\breve{M}_{\Gamma k}) / \sqrt{\text{Var}(\breve{M}_{\Gamma k})}$ for TS2.

The large standardized bias for $\overline{M}_{\Gamma 1}$ and $\overline{M}_{\Gamma 2}$ warrants further exploration, because the associated functions share a feature that largely accounts for their poor performance in some conditions. Namely, both $\Gamma_1 = g_1(\boldsymbol{\Theta}) = \Theta_1^2$ and $\Gamma_2 = g_2(\boldsymbol{\Theta}) = \Theta_1\Theta_2$ are quadratic functions of $\boldsymbol{\Theta}$, so they are not in general monotonic over their corresponding components of $\boldsymbol{\Theta}$. Although $g_1$ and $g_2$ imply simple transformations of $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ to obtain $\overline{M}_{\Gamma 1}$ and $\overline{M}_{\Gamma 2}$, these mean estimators' performance deteriorates when $\mathbf{M}_{\boldsymbol{\Theta}}$ is near a critical point of $g_1$ or $g_2$. More specifically, the unusually large

Table 3
*Percentiles of Standardized Bias for Estimators of $M_{\Gamma 1}$ and $M_{\Gamma 2}$, by Selected Conditions*

| | $\eta = 0$ and $\xi_1 \leq 0.10^2$ (54 conditions) | | | | $\eta \geq 0.4$ or $\xi_1 = 0.20^2$ (189 conditions) | | | |
|---|---|---|---|---|---|---|---|---|
| | IT | | TS2 | | IT | | TS2 | |
| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_1$ | $\Gamma_2$ |
| 100 | 83 | 74 | 83 | 74 | 19 | 20 | 19 | 20 |
| 95 | 81 | 64 | 81 | 64 | 14 | 10 | 14 | 10 |
| 90 | 79 | 50 | 79 | 50 | 11 | 8 | 11 | 8 |
| 80 | 70 | 34 | 70 | 34 | 9 | 7 | 9 | 7 |
| 75 | 69 | 25 | 69 | 25 | 8 | 6 | 8 | 6 |
| 50 | 47 | 7 | 47 | 7 | 4 | 2 | 4 | 2 |
| 25 | 22 | -2 | 22 | -2 | 1 | -3 | 1 | -2 |
| 20 | 20 | -5 | 20 | -5 | 0 | -3 | 0 | -3 |
| 10 | 11 | -12 | 11 | -12 | -2 | -6 | -2 | -6 |
| 5 | 8 | -18 | 8 | -18 | -4 | -8 | -4 | -8 |
| 0 | 2 | -22 | 2 | -22 | -9 | -11 | -9 | -10 |

*Note.* Estimators: IT = integral transformation; TS2 = Taylor series, order 2. Table entry is percentile (across specified conditions) of Monte Carlo estimate of $100\text{Bias}(\hat{M}_{\Gamma k}) / \sqrt{\text{Var}(\hat{M}_{\Gamma k})}$ for IT or $100\text{Bias}(\breve{M}_{\Gamma k}) / \sqrt{\text{Var}(\breve{M}_{\Gamma k})}$ for TS2.

positive standardized bias for $\overline{M}_{\Gamma 1}$ and $\overline{M}_{\Gamma 2}$ when $\eta = 0$ with small $\xi_1$ seems to be due mainly to both large positive bias and small variance compared to other conditions. This is easiest to illustrate for $\overline{M}_{\Gamma 1}$, for which $M_{\Gamma 1} = \Sigma_{\Theta 11} + M^2_{\Theta 1}$ and both the IT and TS2 estimators are essentially $\overline{M}_{\Gamma 1} = \widetilde{\Sigma}_{\Theta 11} + \hat{M}^2_{\Theta 1}$; a similar explanation for $M_{\Gamma 2}$ would involve $\Theta_1$ and $\Theta_2$ (e.g., covariances, products of means). First, when $\eta = M_{\Theta 1} = 0$, $\hat{M}^2_{\Theta 1}$ is often nearly 0, so $\overline{M}_{\Gamma 1}$ depends largely on $\widetilde{\Sigma}_{\Theta 11}$. As shown later, $\widetilde{\Sigma}_{\Theta 11}$ overestimated $\Sigma_{\Theta 11}$ substantially when $\Sigma_{\Theta 11} = \xi_1$ was small (relative to within-study variance), which in turn induced positive bias in $\overline{M}_{\Gamma 1}$.

Second, when $\eta = M_{\Theta 1} = 0$ the empirical standard error used to standardize the bias of $\overline{M}_{\Gamma 1}$, $SE(\overline{M}_{\Gamma 1})$, depends largely on the variance of $\widetilde{\Sigma}_{\Theta 11}$, which is small compared to the variance of $\hat{M}^2_{\Theta 1}$ that contributes more to $SE(\overline{M}_{\Gamma 1})$ when $\eta \geq 0.4$. More specifically, in any condition the true variance of $\overline{M}_{\Gamma 1}$ is essentially

$$Var(\overline{M}_{\Gamma 1}) = Var(\widetilde{\Sigma}_{\Theta 11}) + Var(\hat{M}^2_{\Theta 1}) + 2Cov(\hat{M}^2_{\Theta 1}, \widetilde{\Sigma}_{\Theta 11}) .$$

Treating the covariance in this expression as negligibly small and noting that if $\hat{M}_{\Theta 1}$ is normally distributed then

$$Var(\hat{M}^2_{\Theta 1}) = 2[Var(\hat{M}_{\Theta 1})]^2 + 4[E(\hat{M}_{\Theta 1})]^2 Var(\hat{M}_{\Theta 1}) ,$$

we see that if $E(\hat{M}_{\Theta 1}) \approx M_{\Theta 1} = 0$ then to a close approximation

$$Var(\overline{M}_{\Gamma 1}) \approx Var(\widetilde{\Sigma}_{\Theta 11}) + 2[Var(\hat{M}_{\Theta 1})]^2 .$$

The second term on the right-hand side of the latter expression is relatively small: In the present simulation, Monte Carlo estimates of $\mathrm{Var}(\widetilde{\Sigma}_{\Theta 11})$ / $\mathrm{Var}(\overline{M}_{\Gamma 1})$ were nearly always between 0.80 and 1.00 in the 81 conditions with $M_{\Theta 1} = 0$, whereas this quotient was between 0.10 and 0.15 when $M_{\Theta 1} = 0.4$ and $\Sigma_{\Theta 11} = 0.04$ (27 conditions) and otherwise between 0.00 and 0.06 (135 conditions). The key insight here is that $\mathrm{Var}(\hat{M}^2_{\Theta 1})$ and its relative contribution to $\mathrm{Var}(\overline{M}_{\Gamma 1})$ are much smaller when $\eta = 0$ than when $\eta \geq 0.4$.

Table 4
*Percentiles of Difference in Absolute Standardized Bias Between Estimators of* $M_{\Gamma k}$

| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
|---|---|---|---|---|---|---|
| 100 | 0.6 | 0.6 | 0.7 | 2.2 | 0.5 | 2.8 |
| 99 | 0.5 | 0.4 | 0.6 | 1.7 | 0.4 | 2.2 |
| 98 | 0.3 | 0.4 | 0.5 | 1.5 | 0.3 | 1.8 |
| 95 | 0.3 | 0.3 | 0.3 | 1.1 | 0.3 | 1.4 |
| 90 | 0.2 | 0.2 | 0.2 | 0.7 | 0.2 | 0.8 |
| 80 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 |
| 75 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 |
| 50 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 25 | -0.1 | -0.1 | -0.1 | -0.2 | -0.1 | -0.1 |
| 20 | -0.1 | -0.1 | -0.1 | -0.2 | -0.1 | -0.2 |
| 10 | -0.2 | -0.2 | -0.2 | -0.6 | -0.2 | -0.9 |
| 5 | -0.2 | -0.3 | -0.3 | -1.0 | -0.2 | -1.6 |
| 2 | -0.3 | -0.4 | -0.4 | -1.3 | -0.3 | -2.1 |
| 1 | -0.4 | -0.5 | -0.5 | -1.7 | -0.4 | -2.3 |
| 0 | -0.5 | -0.6 | -0.5 | -2.0 | -0.5 | -3.0 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $100[|\mathrm{Bias}(\breve{M}_{\Gamma k})| / \sqrt{\mathrm{Var}(\breve{M}_{\Gamma k})}\,] - 100[|\mathrm{Bias}(\hat{M}_{\Gamma k})| / \sqrt{\mathrm{Var}(\hat{M}_{\Gamma k})}\,]$, where $\breve{M}_{\Gamma k}$ and $\hat{M}_{\Gamma k}$ are TS2 and IT estimators.

Hence, in conditions with $\eta = 0$ and small $\xi_1$, large positive bias and small variance for $\overline{M}_{\Gamma 1}$ occur mainly because $\overline{M}_{\Gamma 1}$ depends largely on $\widetilde{\Sigma}_{\Theta 11}$.

As for the effect of estimation method, the IT and TS2 estimators of $\mathbf{M}_{\Gamma}$ yielded very similar standardized bias. Table 4 shows percentiles for 100 times these estimators' difference in absolute standardized bias, computed in each condition as $|\text{SBias}(\breve{M}_{\Gamma k})| - |\text{SBias}(\hat{M}_{\Gamma k})|$. These absolute values usually differed by less than 0.01 and never by more than 0.03. For another perspective on this agreement between IT and TS2, Figure 1

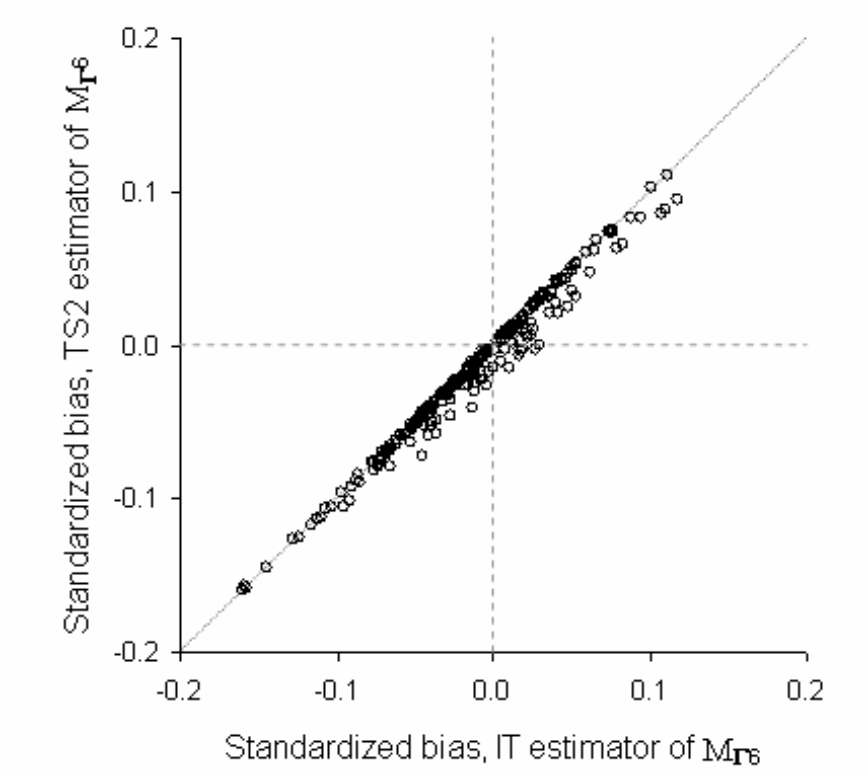


*Figure 1.* Scatterplot of standardized bias for TS2 ($\breve{M}_{\Gamma 6}$) and IT ($\hat{M}_{\Gamma 6}$) estimators of $M_{\Gamma 6}$, with identity line for reference.

shows the standardized bias of $\breve{M}_{\Gamma 6}$ plotted against that of $\hat{M}_{\Gamma 6}$, with the identity line for reference: The IT estimator's standardized bias tended to be slightly higher—more positive or less negative—than the TS2 estimator's. The linear correlation accompanying this plot, $r = .990$, was smaller than that for the other five components ($r > .995$).

Table 5
*Percentiles of Transformed Relative Efficiency for Estimators of* $M_{\Gamma k}$

| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
|---|---|---|---|---|---|---|
| 100 | 1.3 | 1.4 | 1.3 | 1.6 | 1.0 | 2.0 |
| 99 | 1.0 | 1.2 | 1.1 | 1.4 | 0.9 | 1.3 |
| 98 | 0.9 | 1.1 | 0.9 | 1.3 | 0.9 | 1.1 |
| 95 | 0.8 | 0.7 | 0.6 | 1.1 | 0.7 | 0.9 |
| 90 | 0.6 | 0.5 | 0.5 | 0.9 | 0.5 | 0.6 |
| 80 | 0.4 | 0.3 | 0.3 | 0.6 | 0.4 | 0.4 |
| 75 | 0.3 | 0.3 | 0.3 | 0.5 | 0.4 | 0.3 |
| 50 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.0 |
| 25 | -0.1 | -0.1 | -0.1 | 0.0 | -0.1 | -0.2 |
| 20 | -0.2 | -0.2 | -0.1 | 0.0 | -0.1 | -0.3 |
| 10 | -0.4 | -0.3 | -0.2 | -0.1 | -0.2 | -0.6 |
| 5 | -0.4 | -0.5 | -0.3 | -0.3 | -0.4 | -0.7 |
| 2 | -0.5 | -0.7 | -0.5 | -0.3 | -0.4 | -1.0 |
| 1 | -0.7 | -0.7 | -0.6 | -0.3 | -0.4 | -1.0 |
| 0 | -1.0 | -1.1 | -0.9 | -0.4 | -0.8 | -1.2 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $100[\mathrm{MSE}(\hat{M}_{\Gamma k}) / \mathrm{MSE}(\breve{M}_{\Gamma k})] - 100$, where $\hat{M}_{\Gamma k}$ and $\breve{M}_{\Gamma k}$ are IT and TS2 estimators.

Finally, the IT and TS2 estimators of $\mathbf{M}_{\Gamma}$ can also be compared in terms of MSE. (I do not report MSE, variance, or their square roots for either estimator, because judging or comparing their magnitudes in a practically relevant way is difficult.) Table 5 displays

a transformed version of the relative efficiency of $\breve{M}_{\Gamma k}$ with respect to $\hat{M}_{\Gamma k}$: $100\text{MSE}(\hat{M}_{\Gamma k}) / \text{MSE}(\breve{M}_{\Gamma k}) - 100$. This transformed measure is negative or positive when the TS2 estimator is less or more efficient, respectively, than its IT counterpart. For example, if $\text{MSE}(\hat{M}_{\Gamma k}) = .005$ and $\text{MSE}(\breve{M}_{\Gamma k}) = .006$, then the TS2 estimator's relative efficiency is .005 / .006 = 0.833, and the transformed relative efficiency is 100(0.833) – 100 = -16.7, so that TS2 is 16.7% less efficient than IT. (In this example IT is 100/.833 – 100 = 20% more efficient than TS2, but such asymmetry is negligible when relative efficiency is near 1.0, as in the present data.) The two estimators were nearly equally efficient, in that TS2 was rarely more than 1% more or less efficient than IT and never more than 2% more or less. Nonetheless, the slight efficiency differences tended to favor the TS2 estimator, which may be due in part to Monte Carlo error in the IT estimator. This slight advantage for TS2 was most noticeable for $\Gamma_4 = g_4(\boldsymbol{\Theta}) = \exp(\Theta_4 - \Theta_3)$, and closer inspection revealed that TS2's advantage was most pronounced when the between-studies variance component was largest ($\xi_1 = 0.20^2$).

**4.2.2. Point estimators of variances.** In each condition the simulation yielded empirical estimates of bias, variance, and MSE for the EM-GLS estimators of $\Sigma_{\boldsymbol{\Theta} jj}$ and the IT and TS2 estimators of $\Sigma_{\boldsymbol{\Gamma} kk}$. Analogous to Table 1 for $\hat{M}_{\boldsymbol{\Theta} j}$, the top panel of Table 6 shows percentiles of 100 times standardized bias for $\widetilde{\Sigma}_{\boldsymbol{\Theta} jj}$, $\text{Bias}(\widetilde{\Sigma}_{\boldsymbol{\Theta} jj}) / \text{SE}(\widetilde{\Sigma}_{\boldsymbol{\Theta} jj})$. Because $\Sigma_{\boldsymbol{\Theta} jj} = \xi_1 > 0$ for all $j$ in all conditions, it is also reasonable to express bias for $\widetilde{\Sigma}_{\boldsymbol{\Theta} jj}$ as relative bias, $\text{Bias}(\widetilde{\Sigma}_{\boldsymbol{\Theta} jj}) / \Sigma_{\boldsymbol{\Theta} jj}$, as shown in the bottom panel of Table 6 (times

Table 6

*Percentiles of Standardized and Relative Bias for Estimators of* $\Sigma_{\Theta jj}$

| Percentile | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ |
|---|---|---|---|---|---|---|
| | | | Standardized bias | | | |
| 100 | 61 | 63 | 62 | 67 | 61 | 61 |
| 99 | 59 | 60 | 60 | 60 | 60 | 60 |
| 98 | 59 | 59 | 58 | 59 | 58 | 60 |
| 95 | 57 | 57 | 56 | 57 | 57 | 56 |
| 90 | 54 | 54 | 54 | 54 | 54 | 54 |
| 80 | 48 | 49 | 50 | 50 | 50 | 49 |
| 75 | 45 | 45 | 46 | 46 | 46 | 45 |
| 50 | 4 | 5 | 5 | 5 | 5 | 5 |
| 25 | -12 | -10 | -11 | -10 | -11 | -11 |
| 20 | -14 | -13 | -13 | -13 | -14 | -13 |
| 10 | -17 | -17 | -17 | -18 | -18 | -17 |
| 5 | -20 | -20 | -20 | -20 | -19 | -20 |
| 2 | -23 | -24 | -23 | -22 | -22 | -23 |
| 1 | -26 | -25 | -24 | -23 | -23 | -24 |
| 0 | -29 | -25 | -27 | -27 | -34 | -26 |
| | | | Relative bias | | | |
| 100 | 111 | 113 | 108 | 122 | 112 | 117 |
| 99 | 104 | 105 | 104 | 110 | 106 | 106 |
| 98 | 101 | 102 | 99 | 101 | 102 | 101 |
| 95 | 95 | 94 | 92 | 99 | 97 | 95 |
| 90 | 83 | 82 | 85 | 83 | 85 | 86 |
| 80 | 58 | 55 | 59 | 59 | 57 | 59 |
| 75 | 40 | 43 | 41 | 42 | 43 | 40 |
| 50 | 2 | 3 | 3 | 3 | 3 | 3 |
| 25 | -4 | -4 | -4 | -4 | -4 | -4 |
| 20 | -5 | -5 | -5 | -5 | -5 | -5 |
| 10 | -8 | -8 | -8 | -7 | -7 | -7 |
| 5 | -9 | -10 | -9 | -9 | -9 | -9 |
| 2 | -11 | -11 | -12 | -10 | -11 | -10 |
| 1 | -12 | -13 | -12 | -10 | -11 | -12 |
| 0 | -15 | -13 | -14 | -14 | -17 | -13 |

*Note.* Table entry for standardized bias is percentile (across 243 conditions) of Monte Carlo estimate of $100\text{Bias}(\widetilde{\Sigma}_{\Theta jj}) / \sqrt{\text{Var}(\widetilde{\Sigma}_{\Theta jj})}$ ; for relative bias, $100\text{Bias}(\widetilde{\Sigma}_{\Theta jj}) / \Sigma_{\Theta jj}$.

100). I will focus on standardized bias, mainly to consider sampling variance when assessing bias and to maintain consistency with the above treatment of mean estimators. Several key patterns of bias across conditions, methods, and components of $\Theta$ and $\Gamma$ were similar between standardized and relative bias.

It is evident that $\widetilde{\Sigma}_{\Theta jj}$ exhibited greater bias (relative to sampling variance) than $\hat{M}_{\Theta j}$. Although standardized bias for $\widetilde{\Sigma}_{\Theta jj}$ fell below -0.20 in only about 5% of the conditions and then almost never below -0.30, the magnitude of positive standardized bias is more troubling: It exceeded 0.50 in nearly 20% of the conditions and occasionally exceeded 0.60—making MSE($\widetilde{\Sigma}_{\Theta jj}$) up to about 35% larger than Var($\widetilde{\Sigma}_{\Theta jj}$). Closer inspection revealed that standardized bias varied among conditions due primarily to $\xi_1 = \Sigma_{\Theta jj}$. As shown in Table 7, which presents percentiles separately for each value of $\xi_1$, standardized bias was markedly positive for $\xi_1 = .05^2$ (usually between 0.40 and 0.60), negligibly negative to slightly positive for $\xi_1 = 0.10^2$ (usually between -0.10 and 0.20), and somewhat to negligibly negative for $\xi_1 = 0.20^2$ (usually between -0.25 and -0.05). Other design factors' effects were substantially less pronounced, such as a slight tendency for standardized bias to be larger in absolute value with fewer studies. (Any variation in properties of $\widetilde{\Sigma}_{\Theta jj}$ among components of $\Theta$ is due entirely to Monte Carlo sampling error.)

Table 8 shows percentiles of standardized bias separately for IT and TS2 estimators of $\Sigma_{\Gamma kk}$. It will occasionally be convenient to refer simultaneously to the IT

Table 7

*Percentiles of Standardized Bias for Estimators of $\Sigma_{\Theta jj}$, by Selected Conditions*

| | $\xi_1 = 0.05^2$ (81 conditions) | | | | | | $\xi_1 = 0.10^2$ (81 conditions) | | | | | | $\xi_1 = 0.20^2$ (81 conditions) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentile | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ |
| 100 | 61 | 63 | 62 | 67 | 61 | 61 | 20 | 23 | 24 | 23 | 22 | 23 | 0 | -2 | -5 | -5 | -5 | -4 |
| 98 | 60 | 61 | 60 | 62 | 60 | 60 | 20 | 20 | 20 | 20 | 18 | 21 | -3 | -4 | -7 | -6 | -5 | -5 |
| 95 | 59 | 60 | 60 | 60 | 58 | 60 | 18 | 18 | 16 | 18 | 15 | 18 | -5 | -5 | -8 | -6 | -7 | -6 |
| 90 | 58 | 58 | 57 | 59 | 57 | 57 | 14 | 17 | 13 | 14 | 15 | 16 | -8 | -7 | -8 | -8 | -8 | -8 |
| 80 | 56 | 55 | 56 | 56 | 56 | 56 | 10 | 12 | 11 | 13 | 12 | 10 | -10 | -10 | -10 | -9 | -10 | -10 |
| 75 | 56 | 55 | 55 | 56 | 55 | 56 | 10 | 11 | 11 | 11 | 10 | 8 | -12 | -10 | -11 | -10 | -11 | -11 |
| 50 | 50 | 51 | 51 | 51 | 52 | 52 | 4 | 5 | 5 | 5 | 5 | 5 | -15 | -14 | -14 | -15 | -15 | -14 |
| 25 | 45 | 45 | 47 | 46 | 46 | 46 | 0 | -2 | 0 | -1 | -1 | -1 | -18 | -18 | -18 | -19 | -18 | -18 |
| 20 | 43 | 44 | 44 | 46 | 44 | 44 | -2 | -3 | -2 | -2 | -2 | -2 | -19 | -19 | -18 | -19 | -18 | -19 |
| 10 | 40 | 41 | 40 | 41 | 39 | 43 | -5 | -6 | -4 | -5 | -4 | -5 | -21 | -22 | -22 | -20 | -20 | -21 |
| 5 | 39 | 40 | 37 | 39 | 37 | 41 | -6 | -9 | -6 | -6 | -7 | -9 | -23 | -24 | -23 | -22 | -23 | -24 |
| 2 | 37 | 38 | 36 | 35 | 37 | 40 | -9 | -10 | -8 | -8 | -8 | -10 | -28 | -25 | -25 | -25 | -24 | -25 |
| 0 | 36 | 35 | 33 | 33 | 36 | 40 | -13 | -12 | -12 | -10 | -9 | -12 | -29 | -25 | -27 | -27 | -34 | -26 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $100\text{Bias}(\widetilde{\Sigma}_{\Theta jj}) / \sqrt{\text{Var}(\widetilde{\Sigma}_{\Theta jj})}$ .

and TS2 estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}$ and $\breve{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}$, respectively, as $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}} \in \{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}, \breve{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}\}$. For the most part these estimators' standardized bias mimicked that of $\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}$, in that the distributions across conditions were similar. Also, as Table 9 shows, $\xi_1$ largely accounts for variation in standardized bias among conditions.

Closer inspection highlights two notable exceptions, however, where estimators of $\Sigma_{\boldsymbol{\Gamma}kk}$ performed differently than $\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}$. First, standardized bias for $\overline{\Sigma}_{\boldsymbol{\Gamma}11}$ and $\overline{\Sigma}_{\boldsymbol{\Gamma}22}$ was substantially positive—at least compared to other components—not only when $\xi_1 = 0.05^2$

Table 8
*Percentiles of Standardized Bias for Estimators of* $\Sigma_{\boldsymbol{\Gamma}kk}$

| | Integral transformation (IT) | | | | | | Taylor series, order 2 (TS2) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| 100 | 64 | 78 | 62 | 79 | 61 | 61 | 64 | 78 | 62 | 79 | 61 | 61 |
| 99 | 60 | 76 | 60 | 75 | 60 | 60 | 61 | 76 | 60 | 75 | 60 | 60 |
| 98 | 59 | 73 | 58 | 73 | 58 | 59 | 59 | 73 | 58 | 73 | 58 | 59 |
| 95 | 56 | 68 | 56 | 70 | 57 | 56 | 56 | 68 | 56 | 71 | 57 | 56 |
| 90 | 55 | 60 | 54 | 57 | 54 | 54 | 55 | 61 | 54 | 57 | 54 | 54 |
| 80 | 50 | 51 | 50 | 44 | 50 | 49 | 50 | 52 | 50 | 43 | 50 | 49 |
| 75 | 46 | 46 | 46 | 38 | 46 | 45 | 46 | 46 | 46 | 37 | 46 | 46 |
| 50 | 20 | 19 | 5 | 3 | 5 | 5 | 20 | 19 | 5 | -2 | 4 | 8 |
| 25 | -1 | -3 | -11 | -8 | -9 | -12 | -1 | -3 | -6 | -23 | -15 | -2 |
| 20 | -4 | -5 | -13 | -9 | -12 | -15 | -4 | -5 | -8 | -27 | -18 | -5 |
| 10 | -10 | -11 | -18 | -12 | -16 | -18 | -10 | -11 | -13 | -33 | -22 | -13 |
| 5 | -12 | -16 | -21 | -15 | -17 | -20 | -12 | -16 | -17 | -37 | -24 | -18 |
| 2 | -16 | -19 | -24 | -18 | -20 | -24 | -16 | -19 | -20 | -41 | -26 | -21 |
| 1 | -18 | -19 | -26 | -21 | -22 | -26 | -18 | -19 | -21 | -41 | -29 | -23 |
| 0 | -23 | -22 | -29 | -24 | -33 | -30 | -24 | -22 | -24 | -48 | -38 | -27 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $100\text{Bias}(\hat{\Sigma}_{\boldsymbol{\Gamma}kk}) / \sqrt{\text{Var}(\hat{\Sigma}_{\boldsymbol{\Gamma}kk})}$ for IT and $100\text{Bias}(\breve{\Sigma}_{\boldsymbol{\Gamma}kk}) / \sqrt{\text{Var}(\breve{\Sigma}_{\boldsymbol{\Gamma}kk})}$ for TS2.

Table 9

*Percentiles of Standardized Bias for Estimators of $\Sigma_{\Pi kk}$, by Selected Conditions*

| | $\xi_1 = 0.05^2$ (81 conditions) | | | | | | $\xi_1 = 0.10^2$ (81 conditions) | | | | | | $\xi_1 = 0.20^2$ (81 conditions) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| | | | | | | | Integral transformation (IT) | | | | | | | | | | | |
| 100 | 64 | 78 | 62 | 79 | 61 | 61 | 47 | 49 | 24 | 38 | 22 | 24 | 27 | 27 | -6 | 5 | -3 | -5 |
| 98 | 61 | 77 | 60 | 76 | 60 | 60 | 43 | 48 | 20 | 37 | 18 | 20 | 26 | 23 | -7 | 1 | -4 | -6 |
| 95 | 59 | 73 | 60 | 73 | 58 | 60 | 40 | 45 | 15 | 33 | 16 | 18 | 25 | 20 | -8 | -1 | -5 | -8 |
| 90 | 58 | 70 | 57 | 72 | 57 | 57 | 37 | 42 | 13 | 26 | 15 | 15 | 23 | 18 | -9 | -2 | -7 | -9 |
| 80 | 56 | 66 | 55 | 65 | 56 | 56 | 33 | 35 | 10 | 16 | 12 | 10 | 18 | 14 | -11 | -5 | -8 | -10 |
| 75 | 55 | 64 | 55 | 58 | 55 | 55 | 29 | 33 | 10 | 11 | 10 | 7 | 14 | 10 | -11 | -5 | -9 | -12 |
| 50 | 53 | 54 | 51 | 49 | 51 | 51 | 11 | 12 | 5 | 1 | 5 | 5 | -5 | -6 | -15 | -8 | -13 | -16 |
| 25 | 46 | 44 | 46 | 37 | 46 | 46 | 3 | 1 | -1 | -6 | -1 | -1 | -11 | -13 | -19 | -12 | -16 | -19 |
| 20 | 44 | 42 | 44 | 30 | 45 | 45 | 2 | 0 | -2 | -8 | -2 | -2 | -11 | -15 | -19 | -13 | -17 | -19 |
| 10 | 41 | 38 | 40 | 21 | 39 | 43 | -1 | -5 | -5 | -11 | -4 | -6 | -15 | -17 | -23 | -15 | -19 | -22 |
| 5 | 40 | 36 | 37 | 16 | 37 | 41 | -5 | -6 | -6 | -14 | -7 | -9 | -17 | -19 | -25 | -16 | -21 | -24 |
| 2 | 39 | 32 | 35 | 11 | 37 | 40 | -7 | -7 | -8 | -18 | -8 | -11 | -19 | -20 | -26 | -21 | -23 | -27 |
| 0 | 37 | 26 | 33 | 9 | 36 | 39 | -10 | -10 | -12 | -22 | -9 | -13 | -23 | -22 | -29 | -24 | -33 | -30 |

*(table continues)*

| Percentile | $\xi_1 = 0.05^2$ (81 conditions) | | | | | | $\xi_1 = 0.10^2$ (81 conditions) | | | | | | $\xi_1 = 0.20^2$ (81 conditions) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| Taylor series, order 2 (TS2) | | | | | | | | | | | | | | | | | | |
| 100 | 64 | 78 | 62 | 79 | 61 | 61 | 47 | 49 | 25 | 37 | 22 | 24 | 27 | 27 | -1 | -3 | -9 | 21 |
| 98 | 62 | 77 | 60 | 76 | 60 | 60 | 43 | 48 | 21 | 36 | 17 | 22 | 26 | 23 | -2 | -9 | -11 | 17 |
| 95 | 60 | 73 | 60 | 74 | 58 | 59 | 40 | 45 | 16 | 32 | 15 | 19 | 25 | 20 | -2 | -10 | -11 | 14 |
| 90 | 58 | 70 | 57 | 73 | 57 | 57 | 37 | 42 | 14 | 24 | 14 | 17 | 22 | 18 | -3 | -16 | -13 | 10 |
| 80 | 56 | 65 | 56 | 65 | 56 | 56 | 33 | 35 | 11 | 12 | 11 | 12 | 18 | 15 | -5 | -21 | -15 | 5 |
| 75 | 55 | 64 | 55 | 59 | 55 | 55 | 29 | 33 | 11 | 8 | 9 | 11 | 14 | 11 | -6 | -23 | -15 | 3 |
| 50 | 53 | 54 | 52 | 48 | 52 | 52 | 11 | 12 | 5 | -2 | 4 | 6 | -5 | -6 | -10 | -29 | -20 | -5 |
| 25 | 46 | 44 | 47 | 37 | 46 | 46 | 3 | 1 | 0 | -12 | -2 | 2 | -11 | -13 | -14 | -34 | -22 | -14 |
| 20 | 44 | 42 | 44 | 30 | 44 | 44 | 2 | 0 | -1 | -14 | -3 | 0 | -11 | -15 | -15 | -36 | -23 | -17 |
| 10 | 41 | 38 | 41 | 20 | 39 | 43 | -1 | -5 | -4 | -17 | -5 | -4 | -15 | -17 | -19 | -37 | -25 | -20 |
| 5 | 40 | 36 | 37 | 14 | 37 | 42 | -5 | -6 | -6 | -18 | -8 | -7 | -17 | -19 | -20 | -41 | -26 | -22 |
| 2 | 39 | 32 | 36 | 10 | 37 | 40 | -6 | -7 | -7 | -24 | -9 | -9 | -19 | -20 | -21 | -42 | -30 | -25 |
| 0 | 36 | 26 | 34 | 7 | 36 | 40 | -10 | -10 | -11 | -28 | -10 | -11 | -24 | -22 | -24 | -48 | -38 | -27 |

*Note.* Table entry is percentile (across specified conditions) of Monte Carlo estimate of $100\text{Bias}(\hat{\Sigma}_{\Gamma kk}) / \sqrt{\text{Var}(\hat{\Sigma}_{\Gamma kk})}$ for IT and $100\text{Bias}(\breve{\Sigma}_{\Gamma kk}) / \sqrt{\text{Var}(\breve{\Sigma}_{\Gamma kk})}$ for TS2.

(across other design factors) but also when $\xi_1 \geq 0.10^2$ and $\eta = 0$, most notably with fewer studies. This mostly likely occurred for similar reasons as the substantial standardized bias for $\overline{M}_{\Gamma 1}$ and $\overline{M}_{\Gamma 2}$ in the same conditions: For $\Gamma_1 = \Theta_1^2$ in particular, $\Theta_1 \sim N(M_{\Theta 1}, \Sigma_{\Theta 11})$ implies that $\Sigma_{\Gamma 11} = 2\Sigma_{\Theta 11}(\Sigma_{\Theta 11} + 2M_{\Theta 1}^2)$, so when $M_{\Theta 1} = \eta = 0$ the dominant contribution to $\overline{\Sigma}_{\Gamma 11}$ should be from $\widetilde{\Sigma}_{\Theta 11}$, which was substantially positively biased with smaller $\xi_1$. Furthermore, because $E(\hat{M}_{\Theta 1}^2) = Var(\hat{M}_{\Theta 1}) + [E(\hat{M}_{\Theta 1})]^2$, when $\eta = 0$ $\hat{M}_{\Theta 1}^2$ overestimates $M_{\Theta 1}^2 = 0$ more with larger $Var(\hat{M}_{\Theta 1})$, such as with smaller *I* or larger $\xi_1$. Hence, even when $\xi_1 = 0.20^2$ so that negative bias in $\widetilde{\Sigma}_{\Theta 11}$ lowers $\overline{\Sigma}_{\Gamma 11}$, when $\eta = 0$ with larger $Var(\hat{M}_{\Theta 1})$ the positive bias in $\hat{M}_{\Theta 1}^2$ increases $\overline{\Sigma}_{\Gamma 11}$. Also, the variance of $\widetilde{\Sigma}_{\Theta 11}$ is much smaller than that of $2\hat{M}_{\Theta 1}^2$ when $\eta \geq 0.4$, so the variance of $\overline{\Sigma}_{\Gamma 11}$ used in its standardized bias is much smaller when $\eta = 0$ than when $\eta \geq 0.4$. Analogous reasoning applied to $\Gamma_2 = \Theta_1\Theta_2$ would involve covariances.

Second, in some conditions estimators of $\Sigma_{\Theta 22}$ and $\Sigma_{\Theta 44}$ tended to yield notably different standardized bias than those for variances of other components, primarily with smaller $\xi_1$ when $\xi_2$ and $\phi$ (i.e., between- and within-studies correlations) were discrepant. For example, with smaller $\xi_1$ when $\xi_2 \leq .1$ and $\phi = .5$, standardized bias for $\overline{\Sigma}_{\Gamma 22}$ and $\overline{\Sigma}_{\Gamma 44}$ was markedly higher and lower, respectively, than for other components; the opposite pattern held when $\xi_2 = .5$ and $\phi \leq .1$. This dependence on $\xi_2$ and $\phi$ is not surprising, given that $\Gamma_2$ and $\Gamma_4$ depend on more than one component of $\boldsymbol{\Theta}$.

Table 10
*Percentiles of Difference in Absolute Standardized Bias Between Estimators of* $\Sigma_{\Gamma kk}$

| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
|---|---|---|---|---|---|---|
| 100 | 0.7 | 0.6 | 1.5 | 37.4 | 8.3 | 13.3 |
| 99 | 0.5 | 0.4 | 1.4 | 34.5 | 8.2 | 5.7 |
| 98 | 0.4 | 0.4 | 1.3 | 31.9 | 8.0 | 5.6 |
| 95 | 0.3 | 0.3 | 1.1 | 27.1 | 7.8 | 5.2 |
| 90 | 0.2 | 0.2 | 0.9 | 24.2 | 7.2 | 4.0 |
| 80 | 0.2 | 0.2 | 0.7 | 16.4 | 5.8 | 2.6 |
| 75 | 0.1 | 0.1 | 0.3 | 12.7 | 5.1 | 2.3 |
| 50 | 0.0 | 0.0 | 0.0 | 0.9 | 0.0 | 0.2 |
| 25 | -0.1 | -0.1 | -4.1 | -0.8 | -0.2 | -1.2 |
| 20 | -0.1 | -0.1 | -4.5 | -1.0 | -0.9 | -3.1 |
| 10 | -0.2 | -0.2 | -5.6 | -1.9 | -1.1 | -12.7 |
| 5 | -0.3 | -0.3 | -6.7 | -3.3 | -1.2 | -14.9 |
| 2 | -0.4 | -0.5 | -7.1 | -4.1 | -1.3 | -17.6 |
| 1 | -0.5 | -0.8 | -7.3 | -4.3 | -1.4 | -18.2 |
| 0 | -0.6 | -1.6 | -7.6 | -4.6 | -1.5 | -21.8 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $100[|\text{Bias}(\breve{\Sigma}_{\Gamma kk})| / \sqrt{\text{Var}(\breve{\Sigma}_{\Gamma kk})}\,] - 100[|\text{Bias}(\hat{\Sigma}_{\Gamma kk})| / \sqrt{\text{Var}(\hat{\Sigma}_{\Gamma kk})}\,]$, where $\breve{\Sigma}_{\Gamma kk}$ and $\hat{\Sigma}_{\Gamma kk}$ are TS2 and IT estimators.

As for comparing estimators of $\Sigma_{\Gamma kk}$, Table 10 shows percentiles for the difference between the TS2 and IT estimators' absolute standardized bias, $|\text{SBias}(\breve{\Sigma}_{\Gamma kk})| - |\text{SBias}(\hat{\Sigma}_{\Gamma kk})|$. For $\Sigma_{\Gamma 11}$ and $\Sigma_{\Gamma 22}$ these differences were negligible, as we might expect given that for these components the estimators differ primarily due to Monte Carlo sampling error in the IT estimator. Other components of $\Gamma$ yielded notable differences in standardized bias, however, mainly when $\xi_1$ was large. The most pronounced discrepancy occurred for $\Sigma_{\Gamma 44}$ when $\xi_1 = 0.20^2$: Both the IT and TS2 estimators usually

exhibited negative bias, but TS2's was substantially larger (typically -0.40 to -0.20) than IT's (typically -0.10 to 0.0), especially with larger *I*. TS2 also exhibited somewhat more negative bias than IT for $\Sigma_{\Gamma 55}$, mainly when $\xi_1 = 0.20^2$. On the other hand, the IT variance estimator was sometimes more biased than its TS2 counterpart, such as for $\Sigma_{\Gamma 66}$ when $\xi_1 = 0.20^2$ with smaller $\eta$ and *I* (with larger *I* the estimators were often biased in different directions) or for $\Sigma_{\Gamma 33}$ when $\xi_1 = 0.20^2$. The scatterplot in Figure 2 shows the correspondence between standardized bias for $\breve{\Sigma}_{\Gamma 66}$ and $\hat{\Sigma}_{\Gamma 66}$ across conditions.

Regarding the efficiency of IT and TS2 estimators of $\Sigma_{\Gamma kk}$, Table 11 shows percentiles for the transformed relative efficiency of TS2 with respect to IT:

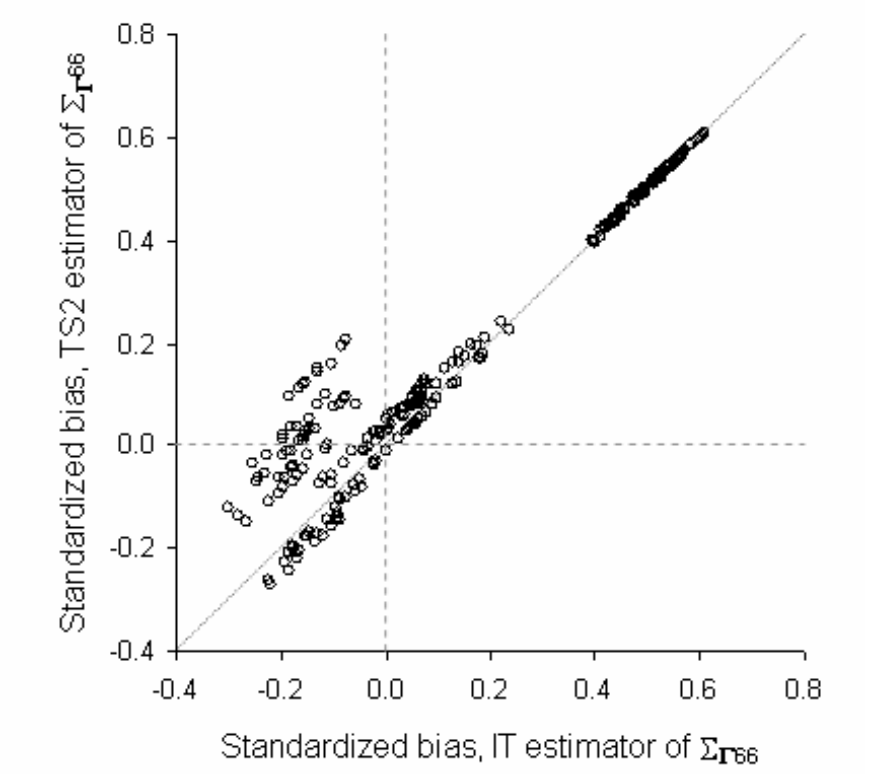


*Figure 2.* Scatterplot of standardized bias for TS2 ($\breve{\Sigma}_{\Gamma 66}$) and IT ($\hat{\Sigma}_{\Gamma 66}$) estimators of $\Sigma_{\Gamma 66}$, with identity line for reference.

100MSE($\hat{\Sigma}_{\Gamma kk}$) / MSE($\breve{\Sigma}_{\Gamma kk}$) – 100. The two estimators did not differ substantially in efficiency for $\Sigma_{\Gamma 11}$ or $\Sigma_{\Gamma 22}$ but sometimes did for other components of $\Gamma$, primarily with larger $\xi_1$. More specifically, TS2 tended to be less efficient than IT for $\Sigma_{\Gamma 33}$ (up to 8% less, mainly with smaller $\eta$ and *I*) and especially for $\Sigma_{\Gamma 66}$ (up to 29% less, mainly with smaller $\eta$) but more efficient than IT for $\Sigma_{\Gamma 55}$ (up to 10% more) and especially for $\Sigma_{\Gamma 44}$ (up to 52% more, mainly with smaller *I*).

Table 11
*Percentiles of Transformed Relative Efficiency for Estimators of* $\Sigma_{\Gamma kk}$

| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
|---|---|---|---|---|---|---|
| 100 | 3 | 3 | 1 | 52 | 10 | 5 |
| 99 | 3 | 2 | 0 | 42 | 9 | 4 |
| 98 | 2 | 2 | 0 | 41 | 9 | 3 |
| 95 | 2 | 2 | -1 | 35 | 8 | 3 |
| 90 | 1 | 1 | -1 | 29 | 8 | 2 |
| 80 | 1 | 1 | -1 | 21 | 7 | 2 |
| 75 | 1 | 0 | -1 | 17 | 6 | 1 |
| 50 | 0 | 0 | -2 | 9 | 3 | -5 |
| 25 | 0 | 0 | -4 | 5 | 2 | -12 |
| 20 | 0 | -1 | -5 | 5 | 2 | -16 |
| 10 | -1 | -1 | -6 | 3 | 1 | -22 |
| 5 | -1 | -1 | -7 | 3 | 1 | -25 |
| 2 | -2 | -2 | -7 | 2 | 1 | -27 |
| 1 | -2 | -3 | -8 | 2 | 0 | -27 |
| 0 | -3 | -5 | -8 | 2 | 0 | -29 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of 100[MSE($\hat{\Sigma}_{\Gamma kk}$) / MSE($\breve{\Sigma}_{\Gamma kk}$)] – 100, where $\hat{\Sigma}_{\Gamma kk}$ and $\breve{\Sigma}_{\Gamma kk}$ are IT and TS2 estimators.

It is important to note that unlike the IT and TS2 estimators of $M_{\Gamma k}$, which performed quite similarly in nearly all conditions for all components of $\Gamma$, the IT and TS2 estimators of $\Sigma_{\Gamma kk}$ sometimes performed rather differently for $\Gamma_3$ through $\Gamma_6$. To complicate matters further, these differences in performance between variance estimators were typically such that the estimator with larger standardized bias was more efficient. This could occur if, for instance, $\hat{\Sigma}_{\Gamma kk}$ and $\breve{\Sigma}_{\Gamma kk}$ had the same (unstandardized) absolute bias but differed in variance, in which case one might prefer the more precise estimator (i.e., with smaller variance), but the present situation was usually more ambiguous: In conditions where the IT and TS2 estimators of $\Sigma_{\Gamma 33}$ through $\Sigma_{\Gamma 66}$ tended to differ in absolute bias and variance—mainly when $\xi_1$ was large—the more precise estimator also tended to be more biased, but the proportion of MSE due to bias was small enough that the estimator with smaller variance had smaller MSE. As an illustrative example, when $[I, \eta, \xi_1, \xi_2, \phi] = [20, 0.8, 0.20^2, -.1, .1]$ the IT estimator had $\text{Bias}(\hat{\Sigma}_{\Gamma 44}) = -0.00339$ and $\text{Var}(\hat{\Sigma}_{\Gamma 44}) = 0.00215$, whereas for the TS2 estimator $\text{Bias}(\breve{\Sigma}_{\Gamma 44}) = -0.01276$ and $\text{Var}(\breve{\Sigma}_{\Gamma 44}) = 0.00146$. Hence, standardized bias was much larger for TS2 (-0.334) than IT (-0.073) but MSE was smaller for TS2 (0.00162) than IT (0.00216).

**4.2.3. Confidence intervals for means.** In each condition the simulation yielded empirical estimates of the coverage probability of standard-normal 95% CIs for $M_{\Theta j}$ (from EM-GLS estimates) and for $M_{\Gamma k}$ (by the delta method applied to IT and TS2 estimates). For CIs for $M_{\Theta j}$, the top panel of Table 12 shows percentiles of the empirical coverage percentage's difference from nominal 95%, which estimates $100\pi_j - 95$, where

$\pi_j = Pr(M_{\Theta jL} < M_{\Theta j} < M_{\Theta jU})$ is the coverage probability for the 95% CI ($M_{\Theta jL}$, $M_{\Theta jU}$). The bottom panel of Table 12 shows percentiles of this departure from nominal expressed as a logit difference, which estimates $\ln\{[\pi_j / (1 - \pi_j)] / [.95 / (1 - .95)]\}$. Compared to the percentage difference, this logit difference (i.e., log odds ratio) amplifies deviations for coverage probabilities further from .5, consequently emphasizing over- versus undercoverage. I will focus on the percentage difference. For a 95% CI that attains nominal coverage probability, the standard error of coverage probability estimated from the present 500 replications is slightly below .01. Approximate 90% and 99% acceptance

Table 12
*Percentiles of Deviation from Nominal Coverage Percentage and Logit Probability for 95% Confidence Intervals for* $M_{\Theta j}$

| Percentile | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ |
|---|---|---|---|---|---|---|
| | | | Percentage | | | |
| 100 | 2.8 | 3.0 | 2.8 | 3.6 | 2.8 | 3.2 |
| 99 | 2.2 | 2.3 | 2.4 | 2.3 | 2.4 | 2.2 |
| 98 | 2.0 | 1.8 | 2.0 | 2.0 | 1.6 | 2.0 |
| 95 | 1.6 | 1.6 | 1.6 | 1.6 | 1.4 | 1.4 |
| 90 | 1.0 | 1.2 | 1.0 | 1.0 | 1.0 | 1.0 |
| 80 | 0.4 | 0.6 | 0.5 | 0.6 | 0.6 | 0.4 |
| 75 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.0 |
| 50 | -0.8 | -0.8 | -0.8 | -1.2 | -0.8 | -1.0 |
| 25 | -2.3 | -2.2 | -1.9 | -2.2 | -2.2 | -2.0 |
| 20 | -2.6 | -2.5 | -2.4 | -2.4 | -2.4 | -2.4 |
| 10 | -3.4 | -3.6 | -3.6 | -3.4 | -3.4 | -3.2 |
| 5 | -4.4 | -4.0 | -5.0 | -4.0 | -4.2 | -4.0 |
| 2 | -5.6 | -4.4 | -5.6 | -4.6 | -4.8 | -5.1 |
| 1 | -5.9 | -4.6 | -6.7 | -4.7 | -5.3 | -5.6 |
| 0 | -7.0 | -7.4 | -7.6 | -6.6 | -6.2 | -6.8 |

*(table continues)*

| Percentile | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ |
|---|---|---|---|---|---|---|
| | | | Logit probability | | | |
| 100 | 0.85 | 0.95 | 0.85 | 1.31 | 0.85 | 1.05 |
| 99 | 0.60 | 0.65 | 0.68 | 0.65 | 0.69 | 0.60 |
| 98 | 0.54 | 0.48 | 0.54 | 0.53 | 0.41 | 0.53 |
| 95 | 0.40 | 0.40 | 0.40 | 0.40 | 0.34 | 0.34 |
| 90 | 0.23 | 0.29 | 0.23 | 0.23 | 0.23 | 0.23 |
| 80 | 0.09 | 0.13 | 0.12 | 0.13 | 0.13 | 0.09 |
| 75 | 0.04 | 0.09 | 0.04 | 0.09 | 0.04 | 0.00 |
| 50 | -0.16 | -0.16 | -0.16 | -0.23 | -0.16 | -0.19 |
| 25 | -0.40 | -0.39 | -0.34 | -0.39 | -0.39 | -0.36 |
| 20 | -0.45 | -0.43 | -0.42 | -0.42 | -0.42 | -0.42 |
| 10 | -0.56 | -0.58 | -0.58 | -0.56 | -0.56 | -0.53 |
| 5 | -0.68 | -0.63 | -0.75 | -0.63 | -0.65 | -0.63 |
| 2 | -0.81 | -0.68 | -0.82 | -0.70 | -0.73 | -0.75 |
| 1 | -0.85 | -0.70 | -0.92 | -0.72 | -0.78 | -0.82 |
| 0 | -0.95 | -0.99 | -1.01 | -0.91 | -0.87 | -0.93 |

*Note.* Table entry for percentage is percentile (across 243 conditions) of Monte Carlo estimate of $100\pi_j - 95$, where $\pi_j$ is coverage probability of confidence interval for $M_{\Theta j}$; for logit probability, $\ln[\pi_{\Theta j} / (1 - \pi_{\Theta j})] - \ln(.95 / .05)$.

intervals around .95 are, respectively, (.934, .966) and (.925 and .975), which correspond to the following intervals for percentage difference: (-1.6, 1.6) and (-2.5, 2.5). (These normal-approximation intervals depart slightly from their exact binomial counterparts, which require additional criteria to be defined uniquely.)

The coverage probability of 95% CIs for $M_{\Theta j}$ was often near nominal but departed markedly in some conditions, most notably with fewer studies. As Table 13 shows by displaying percentiles separately for each value of $\xi_1$, coverage probability tended to be slightly too high when $\xi_1 = 0.05^2$ (rarely by more than 2%), somewhat too low when $\xi_1 =$

Table 13

*Percentiles of Deviation from Nominal Coverage Percentage for 95% Confidence Intervals for $M_{\Theta j}$, by Selected Conditions*

| Per-centile | $\xi_1 = 0.05^2$ (81 conditions) | | | | | | $\xi_1 = 0.10^2$ (81 conditions) | | | | | | $\xi_1 = 0.20^2$ (81 conditions) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ | $\Theta_1$ | $\Theta_2$ | $\Theta_3$ | $\Theta_4$ | $\Theta_5$ | $\Theta_6$ |
| 100 | 2.8 | 3.0 | 2.8 | 3.6 | 2.8 | 3.2 | 1.2 | 1.6 | 1.0 | 0.8 | 1.6 | 1.8 | 0.6 | 1.8 | 2.0 | 0.6 | 0.8 | 1.2 |
| 98 | 2.4 | 2.6 | 2.5 | 2.4 | 2.6 | 2.2 | 0.9 | 0.9 | 0.8 | 0.7 | 1.0 | 1.3 | 0.3 | 0.3 | 0.7 | 0.4 | 0.3 | 0.6 |
| 95 | 2.2 | 2.0 | 2.2 | 2.0 | 1.8 | 2.0 | 0.4 | 0.6 | 0.6 | 0.4 | 0.4 | 0.4 | 0.2 | 0.0 | 0.4 | -0.2 | 0.0 | 0.0 |
| 90 | 1.8 | 1.8 | 1.6 | 1.8 | 1.4 | 1.6 | 0.2 | 0.4 | 0.2 | 0.0 | 0.2 | 0.0 | -0.2 | -0.2 | -0.4 | -0.4 | -0.6 | -0.4 |
| 80 | 1.4 | 1.6 | 1.2 | 1.4 | 1.2 | 1.4 | -0.2 | -0.4 | -0.2 | -0.6 | -0.4 | -0.4 | -1.0 | -0.8 | -1.2 | -1.2 | -1.0 | -1.0 |
| 75 | 1.2 | 1.2 | 1.0 | 1.2 | 1.2 | 1.0 | -0.4 | -0.6 | -0.4 | -0.8 | -0.6 | -0.6 | -1.4 | -1.0 | -1.2 | -1.2 | -1.2 | -1.2 |
| 50 | 0.6 | 0.6 | 0.4 | 0.6 | 0.6 | 0.4 | -1.2 | -1.2 | -1.0 | -1.4 | -1.2 | -1.4 | -2.4 | -2.2 | -2.2 | -2.2 | -2.4 | -2.0 |
| 25 | 0.0 | -0.2 | 0.0 | 0.2 | 0.0 | -0.2 | -2.2 | -2.0 | -1.8 | -2.2 | -2.0 | -2.2 | -3.4 | -3.6 | -3.8 | -3.6 | -3.4 | -3.4 |
| 20 | -0.2 | -0.2 | -0.2 | 0.2 | 0.0 | -0.2 | -2.6 | -2.2 | -2.0 | -2.2 | -2.2 | -2.4 | -3.6 | -3.8 | -4.4 | -3.8 | -3.6 | -3.8 |
| 10 | -0.6 | -0.6 | -0.8 | -0.4 | -0.6 | -0.6 | -3.0 | -2.6 | -2.8 | -2.4 | -2.4 | -2.6 | -4.8 | -4.0 | -5.2 | -4.4 | -4.2 | -4.8 |
| 5 | -1.0 | -1.0 | -0.8 | -0.8 | -0.8 | -1.0 | -3.2 | -3.4 | -3.0 | -3.0 | -2.6 | -3.2 | -5.6 | -4.4 | -5.6 | -4.6 | -5.0 | -5.4 |
| 2 | -1.3 | -1.2 | -1.1 | -1.7 | -1.5 | -1.2 | -3.8 | -4.4 | -3.8 | -3.3 | -3.6 | -3.4 | -6.2 | -4.9 | -6.7 | -4.9 | -5.6 | -5.9 |
| 0 | -1.6 | -1.6 | -1.8 | -2.2 | -2.0 | -1.6 | -4.4 | -4.6 | -6.8 | -3.6 | -4.4 | -3.8 | -7.0 | -7.4 | -7.6 | -6.6 | -6.2 | -6.8 |

*Note.* Table entry is percentile (across specified conditions) of Monte Carlo estimate of $100\pi_j - 95$, where $\pi_j$ is coverage probability of confidence interval for $M_{\Theta j}$.

$0.10^2$ (rarely by more than 3%), and notably too low when $\xi_1 = 0.20^2$ (usually by between 1% and 5%). The slight overcoverage when $\Sigma_{\boldsymbol{\Theta}jj} = \xi_1 = 0.05^2$ may be due largely to substantial overestimation of $\Sigma_{\boldsymbol{\Theta}jj}$ in those conditions, especially for smaller *I*. This positive bias in $\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}$ inflates the standard error of $\hat{\mathrm{M}}_{\boldsymbol{\Theta}j}$, but its influence on the CI's width via $\boldsymbol{\Psi}_{\mathbf{T}i} + \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ (see $\boldsymbol{\Xi}_{\mathbf{T}}$ in Equation 3) may be slight, because when $\xi_1 = 0.05^2$ even positively biased $\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}$—typically $0.003 < \mathrm{E}(\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}) < 0.005$—tends to be small relative to $\boldsymbol{\Psi}_{\mathbf{T}ijj} \approx 1 / \bar{n} = 1 / 100 = 0.01$. At the other extreme, undercoverage when $\xi_1 = 0.20^2$ may be due largely to underestimation of $\Sigma_{\boldsymbol{\Theta}jj}$, which was worse for smaller *I* and would impact the CI's width more than when $\xi_1 = 0.05^2$, because $\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}$—typically $0.036 < \mathrm{E}(\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}) < 0.039$—is larger relative to $\boldsymbol{\Psi}_{\mathbf{T}ijj}$.

Treating $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ as known by using standard-normal quantiles may also affect CI performance, primarily by narrowing CIs based on fewer studies (i.e., smaller *I*) relative to CIs that would treat $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ as an estimate. This may counteract the inflation of $\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}$ when $\xi_1 = 0.05^2$, fortuitously yielding nearly nominal coverage probability. It may also exacerbate the CI-narrowing influence of negatively biased $\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}$ when $\xi_1 = 0.20^2$, especially because when *I* was smaller the bias and variance of $\widetilde{\Sigma}_{\boldsymbol{\Theta}jj}$ were both larger, which makes treating $\hat{\mathrm{Var}}(\hat{\mathrm{M}}_{\boldsymbol{\Theta}j})$ as known for a standard-normal CI even less appropriate. Finally, treating $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ as known may partly explain why with $\xi_1 = 0.10^2$ and *I* = 10 the CI coverage probability tended be too low despite a slight overestimation of $\Sigma_{\boldsymbol{\Theta}jj}$:

The narrowing influence of using standard-normal quantiles more than compensates for the widening influence of positively biased $\widetilde{\Sigma}_{\Theta jj}$. (Performance of Student-*t* CIs, which are not reported here, was largely consistent with these speculations.)

Table 14
*Percentiles of Deviation from Nominal Coverage Percentage for 95% Confidence Intervals for* $M_{\Gamma k}$

| | Integral transformation (IT) | | | | | | Taylor series, order 2 (TS2) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| 100 | 2.2 | 2.6 | 2.8 | 3.0 | 2.8 | 2.8 | 2.0 | 2.6 | 2.8 | 3.0 | 2.8 | 2.8 |
| 99 | 1.8 | 1.7 | 2.4 | 2.2 | 2.4 | 2.0 | 1.8 | 1.7 | 2.4 | 2.0 | 2.4 | 2.0 |
| 98 | 1.8 | 1.2 | 1.8 | 1.8 | 1.8 | 1.6 | 1.6 | 1.2 | 1.8 | 1.8 | 1.8 | 1.8 |
| 95 | 1.0 | 1.2 | 1.4 | 1.4 | 1.4 | 1.2 | 1.0 | 1.0 | 1.4 | 1.4 | 1.4 | 1.4 |
| 90 | 0.4 | 0.8 | 1.0 | 0.8 | 1.0 | 0.8 | 0.6 | 0.6 | 1.0 | 0.8 | 1.0 | 0.8 |
| 80 | -0.4 | -0.2 | 0.4 | 0.4 | 0.6 | 0.2 | -0.2 | -0.1 | 0.4 | 0.4 | 0.6 | 0.2 |
| 75 | -0.6 | -0.8 | 0.3 | 0.0 | 0.2 | 0.0 | -0.6 | -0.6 | 0.2 | 0.2 | 0.2 | 0.0 |
| 50 | -3.6 | -3.0 | -0.8 | -1.0 | -0.8 | -1.2 | -3.6 | -2.8 | -0.8 | -1.0 | -1.0 | -1.2 |
| 25 | -45.6 | -39.1 | -2.0 | -2.6 | -2.1 | -2.4 | -45.1 | -38.9 | -2.0 | -2.6 | -2.0 | -2.4 |
| 20 | -49.3 | -44.5 | -2.4 | -3.1 | -2.5 | -2.8 | -49.5 | -44.7 | -2.4 | -3.2 | -2.4 | -2.8 |
| 10 | -61.1 | -55.9 | -3.8 | -4.6 | -3.4 | -3.6 | -61.2 | -56.3 | -3.6 | -4.4 | -3.2 | -3.8 |
| 5 | -65.8 | -64.6 | -5.0 | -5.0 | -4.0 | -4.8 | -65.9 | -64.4 | -5.0 | -5.0 | -4.2 | -4.8 |
| 2 | -67.2 | -66.2 | -5.7 | -5.6 | -5.0 | -5.6 | -67.6 | -66.7 | -5.7 | -5.8 | -4.8 | -5.8 |
| 1 | -68.1 | -67.4 | -6.2 | -6.4 | -5.4 | -6.7 | -68.5 | -67.6 | -6.5 | -6.5 | -5.4 | -6.4 |
| 0 | -70.6 | -68.2 | -8.4 | -7.6 | -6.2 | -8.6 | -69.8 | -68.4 | -8.2 | -7.4 | -6.0 | -8.4 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $100\pi_k - 95$, where $\pi_k$ is coverage probability of IT or TS2 confidence interval for $M_{\Gamma k}$.

Table 14 shows coverage results for delta-method CIs for $M_{\Gamma k}$ based on both IT and TS2 estimators. Apart from notable exceptions for $\Gamma_1$ and $\Gamma_2$ that occur in isolated conditions, CIs for $M_{\Gamma k}$ yielded very similar coverage as those for $M_{\Theta j}$, in terms of the

distribution of deviation from nominal coverage percentage as well as the deviation in particular conditions. To illustrate this close correspondence, Figure 3 shows a scatterplot of the deviation of coverage proportion from .95 for CIs for $M_{\Theta 6}$ and IT CIs for $M_{\Gamma 6}$—recall that $\Gamma_6 = \tanh \Theta_6$: Although CIs for $M_{\Gamma 6}$ tended to exhibit somewhat

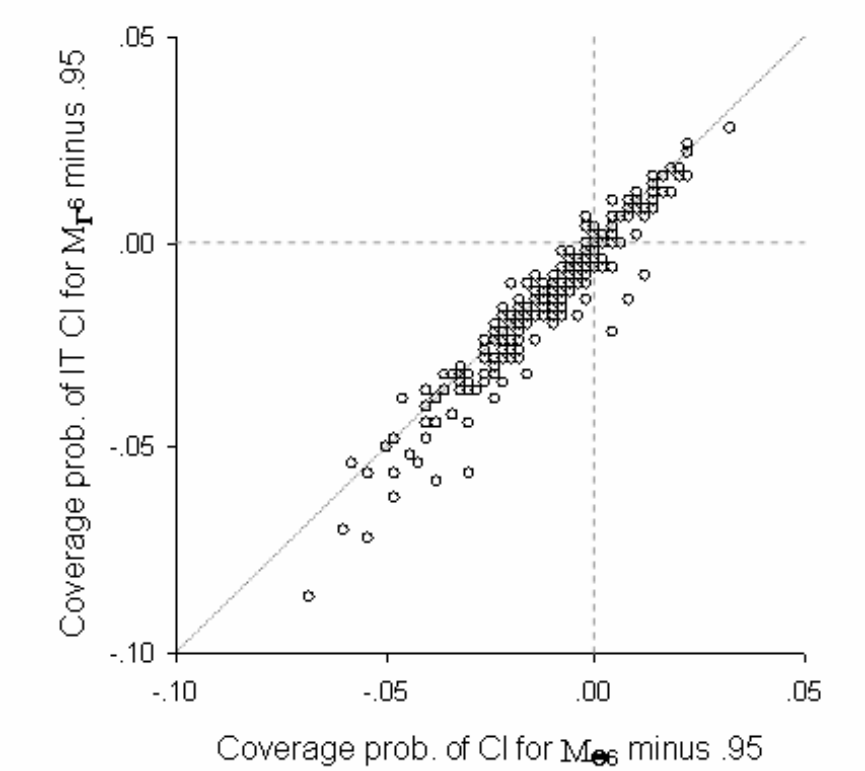


*Figure 3.* Scatterplot of deviation of coverage probability from nominal .95 for CIs for $M_{\Gamma 6}$ (by IT) and $M_{\Theta 6}$, with identity line for reference.

lower coverage probability than their counterparts for $M_{\Theta 6}$, these two coverage proportions were very highly correlated ($r$ = .958), indicating that these two CIs tended to be higher or lower in the same conditions. Hence, the IT and TS2 transformations of $\hat{\mathbf{M}}_{\Theta}$ and $\tilde{\boldsymbol{\Sigma}}_{\Theta}$ into $\overline{\mathbf{M}}_{\Gamma}$ combined with delta-method estimates of Cov($\overline{\mathbf{M}}_{\Gamma}$) do not appear

to introduce much additional error relevant to CI construction, but they also faithfully transfer problems with CIs for $\mathbf{M}_\Theta$ to CIs for $\mathbf{M}_\Gamma$ (e.g., due to bias in estimators of $\mathbf{\Sigma}_\Theta$ and quantiles that treat $\widetilde{\mathbf{\Sigma}}_\Theta$ as known).

Table 15
*Percentiles of Deviation from Nominal Coverage Percentage for 95% Confidence Intervals for $M_{\Gamma 1}$ and $M_{\Gamma 2}$, by Selected Conditions*

| | $\eta = 0$ (81 conditions) | | | | $\eta \geq 0.4$ (162 conditions) | | | |
|---|---|---|---|---|---|---|---|---|
| | IT | | TS2 | | IT | | TS2 | |
| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_1$ | $\Gamma_2$ |
| 100 | -32.0 | -21.8 | -32.6 | -21.4 | 2.2 | 2.6 | 2.0 | 2.6 |
| 98 | -34.4 | -21.9 | -34.5 | -21.9 | 1.8 | 1.4 | 1.8 | 1.4 |
| 95 | -36.4 | -22.6 | -36.6 | -22.6 | 1.4 | 1.2 | 1.4 | 1.2 |
| 90 | -38.2 | -29.6 | -38.2 | -29.4 | 0.8 | 1.0 | 0.8 | 1.0 |
| 80 | -43.8 | -35.2 | -43.8 | -35.6 | 0.2 | 0.4 | 0.2 | 0.4 |
| 75 | -45.8 | -39.2 | -45.2 | -39.0 | 0.0 | 0.2 | 0.0 | 0.2 |
| 50 | -56.8 | -50.2 | -57.2 | -50.0 | -1.2 | -1.4 | -1.3 | -1.3 |
| 25 | -63.2 | -58.4 | -64.0 | -58.6 | -3.6 | -3.0 | -3.6 | -2.8 |
| 20 | -65.2 | -62.6 | -65.2 | -62.6 | -4.0 | -3.4 | -4.0 | -3.4 |
| 10 | -66.8 | -66.0 | -67.2 | -65.8 | -5.4 | -4.4 | -5.0 | -4.8 |
| 5 | -67.4 | -66.4 | -67.6 | -67.0 | -6.4 | -6.0 | -6.4 | -5.8 |
| 2 | -69.2 | -67.9 | -68.9 | -67.8 | -7.3 | -6.6 | -7.2 | -6.4 |
| 0 | -70.6 | -68.2 | -69.8 | -68.4 | -8.4 | -8.8 | -8.2 | -8.6 |

*Note.* Estimators: IT = integral transformation; TS2 = Taylor series, order 2. Table entry is percentile (across specified conditions) of Monte Carlo estimate of $100\pi_k - 95$, where $\pi_k$ is coverage probability of IT or TS2 confidence interval for $M_{\Gamma k}$.

The abysmal performance of CIs for $M_{\Theta 1}$ and $M_{\Theta 2}$ in some conditions warrants further explanation. Table 15 presents percentiles of deviation from nominal coverage percentage for these two CIs separately for conditions with $\eta = 0$ and $\eta \geq 0.40$: CIs for

these two components of $\mathbf{M_\Gamma}$ performed as well as for other components except when $M_{\Theta j}$ = η = 0, in which case their coverage percentage was below nominal 95% by at best more than 30% and 20% for $M_{\Gamma 1}$ and $M_{\Gamma 2}$, respectively, and at worst by around 70% below nominal. This extremely poor coverage when η = 0 was worse with larger *I* or $\xi_1$ (i.e., more studies, larger between-studies variance, or both). Although this dreadful CI coverage and its association with *I* and $\xi_1$ may be influenced by bias in $\overline{M}_{\Gamma 1}$ or $\overline{M}_{\Gamma 2}$, these estimators' non-normality in some conditions, or negative bias in $\widetilde{\Sigma}_{\Theta 11}$ or $\widetilde{\Sigma}_{\Theta 22}$, the dominant culprits are mostly likely two properties of the delta method applied to these quadratic functions when η = 0.

These problematic features are most easily described for $M_{\Gamma 1}$, which involves only one component of $\mathbf{\Theta}$. (Recall that $\Gamma_1 = \Theta_1^2$ and $\Gamma_2 = \Theta_1\Theta_2$.) To a close approximation the delta-method 100(1 – α)% CI for $M_{\Gamma 1}$ by both IT and TS2 is

$$(\hat{M}_{\Theta 1}^2 + \widetilde{\Sigma}_{\Theta 11}) \pm z_\alpha 2|\hat{M}_{\Theta 1}|\sqrt{\hat{\mathrm{Var}}(\hat{M}_{\Theta 1})} \ ,$$

where $2|\hat{M}_{\Theta 1}|$ is the absolute value of the estimated derivative of $\overline{M}_{\Gamma 1}$ with respect to $\hat{M}_{\Theta 1}$. Two aspects of this CI are evidently problematic when $M_{\Theta 1}$ = η = 0, in which case the essentially unbiased $\hat{M}_{\Theta 1}$ is near zero. First, with $\hat{M}_{\Theta 1}^2$ also near zero $\widetilde{\Sigma}_{\Theta 11}$ dominates the point estimator of $M_{\Gamma 1}$, especially for larger $\Sigma_{\Theta 11} = \xi_1$. Note again that

$$\mathrm{Var}(\overline{M}_{\Gamma 1}) \approx \mathrm{Var}(\widetilde{\Sigma}_{\Theta 11} + \hat{M}_{\Theta 1}^2) = \mathrm{Var}(\widetilde{\Sigma}_{\Theta 11}) + \mathrm{Var}(\hat{M}_{\Theta 1}^2) + 2\mathrm{Cov}(\hat{M}_{\Theta 1}^2, \widetilde{\Sigma}_{\Theta 11}) \ ,$$

of which the covariance term is usually negligible. The delta method as implemented here, however, treats $\widetilde{\Sigma}_{\boldsymbol{\Theta}11}$ as fixed and known (see Section 3.4.1) and essentially yields $4\hat{M}^2_{\boldsymbol{\Theta}1}\,\hat{\text{Var}}(\hat{M}_{\boldsymbol{\Theta}1})$ as an estimated approximation of $\text{Var}(\hat{M}^2_{\boldsymbol{\Theta}1})$—if we ignore other components of $\boldsymbol{\Gamma}$—which in turn is substantially smaller than $\text{Var}(\overline{M}_{\boldsymbol{\Gamma}1})$.

Table 16
*Percentiles of Difference in Absolute Deviation from Nominal Coverage Percentage Between 95% Confidence Intervals for* $M_{\boldsymbol{\Gamma}k}$

| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
|---|---|---|---|---|---|---|
| 100 | 1.2 | 1.4 | 1.0 | 0.8 | 0.6 | 0.6 |
| 99 | 0.8 | 1.1 | 0.6 | 0.6 | 0.5 | 0.4 |
| 98 | 0.8 | 0.8 | 0.4 | 0.6 | 0.4 | 0.4 |
| 95 | 0.6 | 0.6 | 0.4 | 0.4 | 0.2 | 0.4 |
| 90 | 0.4 | 0.4 | 0.2 | 0.2 | 0.2 | 0.2 |
| 80 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 75 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 50 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 25 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 |
| 20 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 |
| 10 | -0.4 | -0.4 | -0.2 | -0.2 | -0.2 | -0.2 |
| 5 | -0.4 | -0.6 | -0.4 | -0.4 | -0.4 | -0.4 |
| 2 | -0.8 | -0.8 | -0.4 | -0.4 | -0.6 | -0.6 |
| 1 | -0.9 | -1.1 | -0.6 | -0.6 | -0.6 | -0.6 |
| 0 | -1.6 | -1.8 | -0.6 | -0.6 | -0.8 | -0.8 |

*Note.* Table entry is percentile (across 243 conditions) of Monte Carlo estimate of $|100\pi_{k,\text{TS2}} - 95| - |100\pi_{k,\text{IT}} - 95|$, where $\pi_{k,\text{TS2}}$ and $\pi_{k,\text{IT}}$ are coverage probability of TS2 and IT confidence intervals for $M_{\boldsymbol{\Gamma}k}$.

Second, $\hat{M}^2_{\boldsymbol{\Theta}1}$ and, consequently, the delta method's estimated derivative and approximate variance are more likely to be arbitrarily small with larger *I*. Even if $\widetilde{\Sigma}_{\boldsymbol{\Theta}11}$

were known so that approximating $\mathrm{Var}(\hat{M}^2_{\Theta 1})$ also approximated $\mathrm{Var}(\overline{M}_{\Gamma 1})$ well, $4\hat{M}^2_{\Theta 1}\,\hat{\mathrm{Var}}(\hat{M}_{\Theta 1})$ would approximate $\mathrm{Var}(\hat{M}^2_{\Theta 1})$ poorly when $M_{\Theta 1} = 0$. A somewhat better variance approximation—based on a second-order Taylor polynomial for $\overline{M}_{\Gamma 1}$—is $4\hat{M}^2_{\Theta 1}\,\hat{\mathrm{Var}}(\hat{M}_{\Theta 1}) + 2[\,\hat{\mathrm{Var}}(\hat{M}_{\Theta 1})\,]^2$, the last term of which does not tend to 0 along with $\hat{M}_{\Theta 1}$. In short, when $\eta = 0$ the delta method's approximation of $\mathrm{Var}(\overline{M}_{\Gamma 1})$ is much too small because it neglects sampling error in $\widetilde{\Sigma}_{\Theta 11}$ (worse with larger $\Sigma_{\Theta 11}$) and underestimates $\mathrm{Var}(\hat{M}^2_{\Theta 1})$ (worse with larger *I*).

As indicated by their difference in absolute deviation from nominal coverage percentage, shown in Table 16, the IT and TS2 CIs performed quite similarly in most conditions. For $M_{\Gamma 3}$ through $M_{\Gamma 6}$, these absolute deviations from 95% usually differed by less than 0.5% in either direction and never by more than 1.0%. For example, when $[I, \eta, \xi_1, \xi_2, \phi] = [40, 0.8, 0.20^2, -.1, .5]$ the coverage percentages for the IT and TS2 CIs for $M_{\Gamma 3}$ were, respectively, 95.6% and 94.8%, whose absolute deviations from nominal differ by $|94.8 - 95.0| - |95.6 - 95.0| = 0.2 - 0.6 = -0.4\%$. The IT and TS2 CIs' deviations from nominal 95% occasionally differed somewhat more for $M_{\Gamma 1}$ and $M_{\Gamma 2}$, but this usually occurred for very large deviations when $\eta = 0$ with large *I* or $\xi_1$ (e.g., absolute deviations more than 30%), and the difference never exceeded 2.0% in either direction. There was no apparent tendency for the absolute deviations to be larger for IT or TS2 for any component of $\mathbf{M}_{\Gamma}$. Figure 4 illustrates this symmetry for IT and TS2 CIs for $M_{\Gamma 6}$, and

analogous plots for $M_{\Gamma 1}$ through $M_{\Gamma 5}$ look similar. (Differences in signed deviation from nominal 95% were not notably larger than absolute deviation for any component of $\mathbf{M}_{\Gamma}$.)

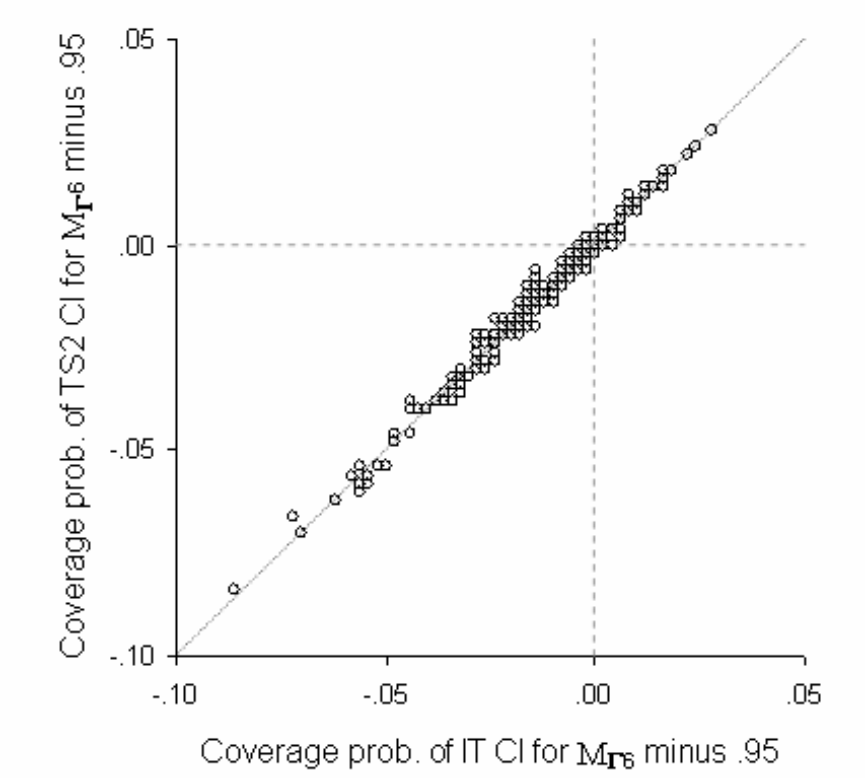


*Figure 4.* Scatterplot of deviation of coverage probability from nominal .95 for TS2 and IT CIs for $M_{\Gamma 6}$, with identity line for reference.

### 4.3. Study 2: Method

This second Monte Carlo study was conducted to evaluate the performance of bootstrap inference. Most aspects of this study were the same as for Study 1, except that bootstrap methods were used instead of the delta method to construct CIs. Bootstrap resampling takes considerably longer than delta-method computations, and three different bootstrap methods were evaluated. Due to time constraints, only a subset of Study 1's conditions were included in Study 2 (see below). It is likely that some choices for

bootstrap methods that were meant to reduce computational burden (e.g., small $B$) also decreased the performance of bootstrap inferences. Nevertheless, this preliminary evaluation of bootstrap inference for functions of ESs serves as valuable groundwork for future investigations.

**4.3.1. Design conditions and data generation.** The same design factors as in Study 1 were used, but only the two extreme values of each factor were included. Hence, the factorial simulation design consisted of $2^5 = 32$ conditions, each defined as a quintuple of five factors: $\eta \in \{0.0, 0.8\}$, $\xi_1 \in \{0.05^2, 0.20^2\}$, $\xi_2 \in \{-.1, .5\}$, $\phi \in \{-.1, .5\}$, and $I \in \{10, 40\}$. I generated 500 independent meta-analytic data sets, each consisting of $I$ independent pairs $[n_i, \mathbf{t}_i]$, using the same procedures as in Study 1.

**4.3.2. Meta-analytic procedures.** The same estimation methods as in Study 1 were applied to each replication's simulated meta-analytic data to obtain $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$, $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$, and $\hat{\mathrm{Cov}}(\hat{\mathbf{M}}_{\boldsymbol{\Theta}})$ as well as IT and TS2 point estimators of $\mathbf{M}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ for the same function $g$. To construct CIs for components of $\mathbf{M}_{\boldsymbol{\Gamma}}$, however, the effect-size, error, and cases bootstrap methods in Section 3.4.2 were used instead of the delta method to estimate $\mathrm{Cov}(\hat{\mathbf{M}}_{\boldsymbol{\Gamma}})$ and $\mathrm{Cov}(\breve{\mathbf{M}}_{\boldsymbol{\Gamma}})$. The raw-data bootstrap was not used, because generating subject-level data whose ESs are from $\mathbf{T}_i \sim N_6(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_{\mathbf{T}i})$ would be equivalent to the effect-size bootstrap. Specifically, each of these three resampling methods was applied to the same replication's data and point estimates using $B = 30$ bootstrap samples, and the covariance matrix of each method's bootstrap sample (Equation 24) was used to obtain standard errors with which to construct standard-normal 95% CIs for $\mathrm{M}_{\boldsymbol{\Gamma}k}$. Hence, the

only difference between delta-method and bootstrap CIs was the estimation of a standard error. (Again, I also computed 90% and 99% CIs as well as Student-*t* CIs at all three confidence levels using $I - 1$ as degrees of freedom, and results from these are available upon request but will not be presented.)

**4.3.3. Evaluation criteria.** The point estimators in Study 2 were identical to those in Study 1, so their performance is not discussed again. My primary interest was in the coverage probability of bootstrap CIs for each $\boldsymbol{\Gamma}$ component. Hence, in each condition 36 distinct coverage probabilities were estimated: one for each component of $\mathbf{M}_{\boldsymbol{\Gamma}}$ by each of two estimation methods crossed with each of three bootstrap methods.

### 4.4. Study 2: Results

In this section I report selected findings from the Monte Carlo study of simple bootstrap CIs. As in Study 1, the presentation is simplified by displaying numerical results as percentiles over conditions. In each of the 32 conditions, the simulation yielded estimated coverage probabilities for 95% CIs constructed for each component of $\mathbf{M}_{\boldsymbol{\Gamma}}$ using standard errors from each point estimator of $\mathbf{M}_{\boldsymbol{\Gamma}}$ (IT and TS2) combined with each bootstrap method (effect size, error, and cases). Of primary interest are comparisons among bootstrap methods as well as between bootstrap and delta-method CIs.

**4.4.1. Comparison of bootstrap methods.** Table 17 displays percentiles of deviation from nominal coverage percentage for each combination of $\mathbf{M}_{\boldsymbol{\Gamma}}$ estimator and bootstrap method. The error bootstrap performed substantially worse than the other bootstrap methods, yielding coverage percentages that were at best 3% below nominal

Table 17

*Percentiles of Deviation from Nominal Coverage Percentage for 95% Confidence Intervals for $M_{\Gamma k}$, by Bootstrap Method*

| Per-centile | Effect size bootstrap | | | | | | Error bootstrap | | | | | | Cases bootstrap | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| | | | | | | | | | Integral transformation (IT) | | | | | | | | | |
| 100 | 5.0 | 5.0 | 1.6 | 1.4 | 1.8 | 1.6 | -4.2 | -2.6 | -4.8 | -3.6 | -4.4 | -3.6 | 1.8 | 2.4 | -0.2 | 0.6 | 0.2 | 0.0 |
| 95 | 5.0 | 5.0 | 1.1 | 1.0 | 1.5 | 1.0 | -6.0 | -6.3 | -5.0 | -4.9 | -5.6 | -5.2 | 1.1 | 0.9 | -0.6 | -0.6 | -1.0 | -0.3 |
| 90 | 4.6 | 4.8 | 0.6 | 0.6 | 1.2 | 0.6 | -6.5 | -8.2 | -5.6 | -6.6 | -6.7 | -6.0 | 0.6 | 0.4 | -1.2 | -0.8 | -1.4 | -1.0 |
| 80 | 4.1 | 3.6 | 0.2 | 0.4 | 0.4 | 0.0 | -9.9 | -9.6 | -6.4 | -7.5 | -7.7 | -7.4 | -0.4 | 0.0 | -1.4 | -1.2 | -2.1 | -1.6 |
| 75 | 1.7 | 3.4 | 0.1 | 0.2 | 0.3 | -0.2 | -10.6 | -9.8 | -6.5 | -7.8 | -8.0 | -7.4 | -0.4 | -0.3 | -1.4 | -1.5 | -2.4 | -1.8 |
| 50 | -1.3 | -0.1 | -1.4 | -1.2 | -1.5 | -1.7 | -13.4 | -12.6 | -9.3 | -10.3 | -8.8 | -9.8 | -4.0 | -3.0 | -3.7 | -3.5 | -3.2 | -3.2 |
| 25 | -4.8 | -2.9 | -3.3 | -3.5 | -2.6 | -3.0 | -18.6 | -16.1 | -10.5 | -11.6 | -11.5 | -11.6 | -6.5 | -5.1 | -5.2 | -5.5 | -4.8 | -5.4 |
| 20 | -5.1 | -4.9 | -3.8 | -4.1 | -3.0 | -4.0 | -19.8 | -17.6 | -11.2 | -11.9 | -11.9 | -11.9 | -7.0 | -5.7 | -5.6 | -5.8 | -5.3 | -6.1 |
| 10 | -6.9 | -5.6 | -4.6 | -4.9 | -4.2 | -5.3 | -26.7 | -24.9 | -12.0 | -12.6 | -13.7 | -12.2 | -10.8 | -7.1 | -6.0 | -6.2 | -6.0 | -7.2 |
| 5 | -7.5 | -5.9 | -5.4 | -5.2 | -4.8 | -6.2 | -28.6 | -29.0 | -12.2 | -13.9 | -13.9 | -12.6 | -13.5 | -10.1 | -6.9 | -6.4 | -7.4 | -7.2 |
| 0 | -9.0 | -6.2 | -7.0 | -5.8 | -7.2 | -7.4 | -28.8 | -31.4 | -12.4 | -14.6 | -14.6 | -13.2 | -14.2 | -13.2 | -8.4 | -9.0 | -8.4 | -8.8 |

*(table continues)*

| Per-centile | Effect size bootstrap | | | | | | Error bootstrap | | | | | | Cases bootstrap | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| | | | | | | | Taylor series, order 2 (TS2) | | | | | | | | | | | |
| 100 | 5.0 | 5.0 | 1.8 | 1.4 | 1.8 | 1.8 | -3.8 | -2.6 | -4.4 | -3.4 | -4.6 | -3.4 | 2.0 | 2.4 | -0.2 | 0.6 | 0.0 | 0.2 |
| 95 | 5.0 | 5.0 | 1.1 | 0.9 | 1.4 | 1.0 | -5.8 | -6.3 | -5.0 | -4.9 | -5.7 | -5.1 | 1.0 | 0.8 | -0.4 | -0.4 | -0.7 | -0.4 |
| 90 | 4.6 | 4.8 | 0.6 | 0.8 | 1.2 | 0.6 | -6.6 | -7.9 | -5.3 | -6.6 | -6.8 | -6.2 | 0.8 | 0.4 | -1.0 | -0.8 | -1.4 | -1.0 |
| 80 | 4.1 | 3.6 | 0.2 | 0.2 | 0.4 | 0.0 | -10.4 | -9.3 | -6.4 | -7.0 | -7.4 | -7.3 | -0.4 | 0.1 | -1.2 | -1.2 | -2.4 | -1.6 |
| 75 | 1.7 | 3.4 | 0.2 | 0.1 | 0.3 | -0.2 | -10.8 | -9.6 | -6.4 | -7.6 | -7.6 | -7.6 | -0.6 | -0.5 | -1.4 | -1.4 | -2.4 | -1.6 |
| 50 | -1.3 | -0.3 | -1.1 | -1.1 | -1.4 | -1.8 | -13.2 | -12.8 | -9.1 | -10.4 | -8.9 | -9.9 | -3.8 | -2.9 | -3.5 | -3.6 | -3.1 | -3.2 |
| 25 | -4.6 | -2.9 | -3.0 | -3.6 | -2.6 | -3.0 | -18.6 | -16.3 | -10.4 | -11.8 | -11.5 | -11.5 | -6.4 | -5.0 | -5.4 | -5.5 | -5.0 | -5.3 |
| 20 | -5.3 | -5.2 | -4.1 | -4.2 | -2.6 | -3.8 | -20.4 | -17.3 | -11.5 | -11.8 | -12.1 | -11.9 | -7.2 | -5.8 | -5.4 | -6.0 | -5.3 | -6.1 |
| 10 | -6.9 | -5.8 | -4.8 | -4.6 | -4.2 | -5.3 | -26.7 | -25.5 | -12.0 | -12.8 | -13.7 | -12.2 | -10.8 | -6.9 | -6.0 | -6.2 | -6.0 | -7.1 |
| 5 | -7.8 | -6.0 | -5.4 | -5.2 | -5.1 | -6.1 | -28.5 | -29.0 | -12.3 | -14.0 | -13.8 | -12.4 | -14.0 | -10.1 | -6.9 | -6.4 | -7.2 | -7.4 |
| 0 | -9.6 | -6.4 | -6.8 | -5.8 | -6.8 | -7.4 | -28.6 | -30.4 | -12.6 | -14.2 | -14.8 | -13.2 | -14.4 | -12.6 | -8.0 | -8.2 | -8.4 | -9.2 |

*Note.* Table entry is percentile (across 32 conditions) of Monte Carlo estimate of $100\pi_k - 95$, where $\pi_k$ is coverage probability of confidence interval for $M_{\Gamma k}$.

95%, more than 10% below nominal in over half of the conditions, and occasionally as low as 30% below nominal. In no condition was the error method's CI coverage nearer nominal than that of its effect-size or cases counterparts, and it was usually at least 4% lower. Results for the error method will not be considered further. The effect-size and cases CIs are compared in more detail below along with their delta-method counterparts.

Table 18
*Percentiles of Difference in Absolute Deviation from Nominal Coverage Percentage Between Integral-Transformation and Taylor-Series 95% Confidence Intervals for $M_{\Gamma k}$, by Bootstrap Method*

| | Effect-size bootstrap | | | | | | Cases bootstrap | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| 100 | 0.6 | 0.6 | 0.4 | 0.8 | 0.6 | 0.6 | 0.8 | 0.4 | 0.8 | 0.4 | 0.4 | 0.4 |
| 95 | 0.5 | 0.5 | 0.3 | 0.2 | 0.4 | 0.3 | 0.4 | 0.3 | 0.2 | 0.2 | 0.4 | 0.4 |
| 90 | 0.4 | 0.4 | 0.2 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.2 | 0.2 | 0.4 | 0.4 |
| 80 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 75 | 0.2 | 0.2 | 0.0 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.0 | 0.2 | 0.2 | 0.2 |
| 50 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 25 | 0.0 | 0.0 | -0.2 | -0.2 | -0.2 | -0.2 | -0.1 | -0.1 | -0.2 | -0.2 | -0.2 | -0.2 |
| 20 | 0.0 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 | -0.2 |
| 10 | -0.2 | -0.2 | -0.4 | -0.2 | -0.4 | -0.2 | -0.2 | -0.2 | -0.4 | -0.4 | -0.4 | -0.4 |
| 5 | -0.3 | -0.3 | -0.4 | -0.4 | -0.4 | -0.2 | -0.4 | -0.4 | -0.4 | -0.5 | -0.5 | -0.5 |
| 0 | -0.6 | -0.6 | -0.4 | -0.6 | -0.4 | -0.4 | -0.6 | -0.6 | -0.6 | -0.8 | -0.8 | -0.6 |

*Note.* Table entry is percentile (across 32 conditions) of Monte Carlo estimate of $|100\pi_{k,\mathrm{TS2}} - 95| - |100\pi_{k,\mathrm{IT}} - 95|$, where $\pi_{k,\mathrm{TS2}}$ and $\pi_{k,\mathrm{IT}}$ are coverage probability of TS2 and IT confidence intervals for $M_{\Gamma k}$.

As for comparisons between the IT and TS2 estimators, Table 18 shows their CIs' difference in absolute deviation from nominal coverage percentage separately for the effect-size and cases methods. These differences were at most 0.8% in absolute value

and typically less than 0.2%, and neither estimator exhibited a systematic pattern of deviating from nominal more than the other. The IT CIs tended to exhibit coverage percentage nearer nominal than TS2 CIs for certain components of $\mathbf{M}_\Gamma$ with certain bootstrap methods (e.g., $M_{\Gamma 1}$ and $M_{\Gamma 2}$ with effect-size bootstrap, $M_{\Gamma 1}$ with cases bootstrap), but this was rather subtle, as was the apparent advantage of TS2 CIs over IT CIs for other combinations (e.g., $M_{\Gamma 3}$ with effect-size bootstrap, $M_{\Gamma 3}$ and $M_{\Gamma 4}$ with cases bootstrap). To simplify subsequent exposition, I will consider only the IT estimator.

Performance differences between effect-size and cases bootstrap CIs depended upon the condition. Table 19 shows percentiles of these two CIs' difference in coverage percentage and their difference in absolute deviation from nominal 95% (for IT only). Cases CIs usually yielded lower coverage percentage than their effect-size counterparts: This was most pronounced when $I = 10$ and $\xi_1 = 0.05^2$ (lower by over 7%), and reversals rarely exceeded 1%—most often when $I = 40$ and $\xi_1 = 0.20^2$. Because the effect-size CIs sometimes exceeded 95% coverage, however, the cases CI occasionally yielded coverage up to 5% nearer nominal in the sense of smaller absolute deviation from 95%. This potential advantage favoring the cases CI occurred almost exclusively for $M_{\Gamma 1}$ and $M_{\Gamma 2}$ in the eight conditions with $\eta = 0$ and $\xi_1 = 0.05^2$. Whether it is indeed an advantage depends partly upon other CI properties, such as width. Upon closer inspection, in these 8 conditions the effect-size method's estimated bootstrap variances for $\hat{M}_{\Gamma 1}$ and $\hat{M}_{\Gamma 2}$ were substantially too large—much more so than in the other 24 conditions.

Table 19
*Percentiles of Difference in Coverage Percentage and Absolute Deviation from Nominal Coverage Percentage Between Effect-Size and Cases 95% Bootstrap Confidence Intervals for $M_{\Gamma k}$*

| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
|---|---|---|---|---|---|---|
| | | | Coverage percentage | | | |
| 100 | 2.0 | 1.0 | 1.2 | 0.4 | 1.2 | 1.2 |
| 95 | 1.2 | 0.9 | 0.8 | 0.2 | 0.8 | 0.9 |
| 90 | 0.2 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 |
| 80 | -1.0 | -0.3 | 0.2 | -0.2 | -0.4 | -0.2 |
| 75 | -1.8 | -1.4 | -0.2 | -0.2 | -0.6 | -0.5 |
| 50 | -3.2 | -2.7 | -1.3 | -1.3 | -1.3 | -1.2 |
| 25 | -5.0 | -4.9 | -2.7 | -3.7 | -4.0 | -3.0 |
| 20 | -5.2 | -5.5 | -3.9 | -4.3 | -4.9 | -4.0 |
| 10 | -5.4 | -6.6 | -5.3 | -5.0 | -6.0 | -5.7 |
| 5 | -6.1 | -7.3 | -5.7 | -5.7 | -6.2 | -6.2 |
| 0 | -6.6 | -7.6 | -7.4 | -6.4 | -7.4 | -6.2 |
| | | | Absolute deviation from nominal 95% | | | |
| 100 | 6.6 | 7.6 | 5.8 | 5.6 | 7.4 | 6.2 |
| 95 | 6.1 | 6.7 | 4.8 | 4.7 | 5.6 | 5.1 |
| 90 | 5.4 | 5.2 | 4.4 | 4.4 | 5.0 | 4.6 |
| 80 | 4.4 | 2.8 | 3.4 | 3.6 | 2.8 | 3.4 |
| 75 | 3.3 | 2.5 | 2.4 | 2.7 | 2.6 | 3.0 |
| 50 | 1.5 | 0.4 | 0.9 | 0.9 | 1.2 | 1.0 |
| 25 | -2.1 | -1.4 | -0.2 | 0.2 | 0.3 | 0.2 |
| 20 | -3.0 | -2.6 | -0.4 | 0.0 | 0.0 | 0.0 |
| 10 | -3.8 | -3.8 | -0.6 | -0.2 | -0.6 | -0.7 |
| 5 | -4.0 | -4.8 | -1.1 | -0.3 | -1.1 | -1.1 |
| 0 | -4.6 | -5.0 | -1.2 | -0.6 | -1.2 | -1.6 |

*Note.* Table entry in top panel is percentile (across 32 conditions) of Monte Carlo estimate of $100(\pi_{k,C} - \pi_{k,ES})$, where $\pi_{k,C}$ and $\pi_{k,ES}$ are coverage probability of cases and effect-size bootstrap confidence intervals for $M_{\Gamma k}$ based on integral transformation; bottom panel, $|100\pi_{k,C} - 95| - |100\pi_{k,ES} - 95|$.

**4.4.2. Bootstrap versus delta method.** Both the effect-size and cases CIs performed substantially better than their delta-method counterparts for $M_{\Gamma 1}$ and $M_{\Gamma 2}$

when $\eta = 0$, but for other components or in other conditions the delta-method CIs tended to outperform their bootstrap counterparts. (Comparisons between delta-method versus bootstrap results are subject to greater sampling error than comparisons between bootstrap results, because the latter are based on the same simulated meta-analytic data, which were independent of Study 1's data.) Table 20 shows percentiles of the difference in absolute deviation from nominal CI coverage percentage between the delta-method CI and each bootstrap CI. This difference favored the bootstrap methods substantially for

Table 20
*Percentiles of Difference in Absolute Deviation from Nominal Coverage Percentage Between Delta-Method and Bootstrap 95% Confidence Intervals for $M_{\Gamma k}$, by Bootstrap Method*

| | Effect-size bootstrap | | | | | | Cases bootstrap | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentile | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | $\Gamma_5$ | $\Gamma_6$ |
| 100 | 2.8 | 2.0 | 4.2 | 3.4 | 1.8 | 3.8 | 6.0 | 7.2 | 6.2 | 5.6 | 8.2 | 5.2 |
| 95 | 2.0 | 1.7 | 2.6 | 1.8 | 1.7 | 2.1 | 4.6 | 4.8 | 4.8 | 5.4 | 5.7 | 4.8 |
| 90 | 0.6 | 1.6 | 2.2 | 1.2 | 1.6 | 1.4 | 4.1 | 3.8 | 4.4 | 4.4 | 5.3 | 4.6 |
| 80 | -0.2 | 0.0 | 1.5 | 0.9 | 1.2 | 1.2 | 2.6 | 2.0 | 3.4 | 3.2 | 3.6 | 3.5 |
| 75 | -0.4 | -0.4 | 1.2 | 0.6 | 1.1 | 0.7 | 2.1 | 1.8 | 3.1 | 2.8 | 2.8 | 3.2 |
| 50 | -14.5 | -8.7 | 0.2 | 0.0 | 0.2 | 0.1 | -16.4 | -10.1 | 1.9 | 1.5 | 1.8 | 1.2 |
| 25 | -43.7 | -42.0 | -0.3 | -0.8 | -0.2 | -1.0 | -42.1 | -41.3 | 0.5 | 0.0 | 0.5 | 0.2 |
| 20 | -53.6 | -51.1 | -0.8 | -1.1 | -0.2 | -1.0 | -56.3 | -53.8 | 0.2 | -0.2 | 0.2 | 0.2 |
| 10 | -60.4 | -60.3 | -1.0 | -1.8 | -0.8 | -1.6 | -60.5 | -60.9 | -0.2 | -1.3 | -0.4 | -0.2 |
| 5 | -62.2 | -63.7 | -1.5 | -2.4 | -1.0 | -1.6 | -62.5 | -63.0 | -1.0 | -1.7 | -0.6 | -0.5 |
| 0 | -65.2 | -66.4 | -3.4 | -3.0 | -2.2 | -2.2 | -63.2 | -65.6 | -3.4 | -2.0 | -0.8 | -2.8 |

*Note.* Table entry is percentile (across 32 conditions) of Monte Carlo estimate of $|100\pi_{k,\mathrm{BS}} - 95| - |100\pi_{k,\mathrm{DM}} - 95|$, where $\pi_{k,\mathrm{BS}}$ and $\pi_{k,\mathrm{DM}}$ are coverage probability of bootstrap and delta-method confidence intervals for $M_{\Gamma k}$ based on IT estimator.

$M_{\Gamma 1}$ and $M_{\Gamma 2}$ in about half of the conditions (i.e., when $\eta = 0$) but otherwise tended to favor the delta method more often and by somewhat larger amounts—more so for the cases bootstrap (sometimes over 5% and up to 8%) than the effect-size bootstrap (usually less than 3% and always less than 5%).

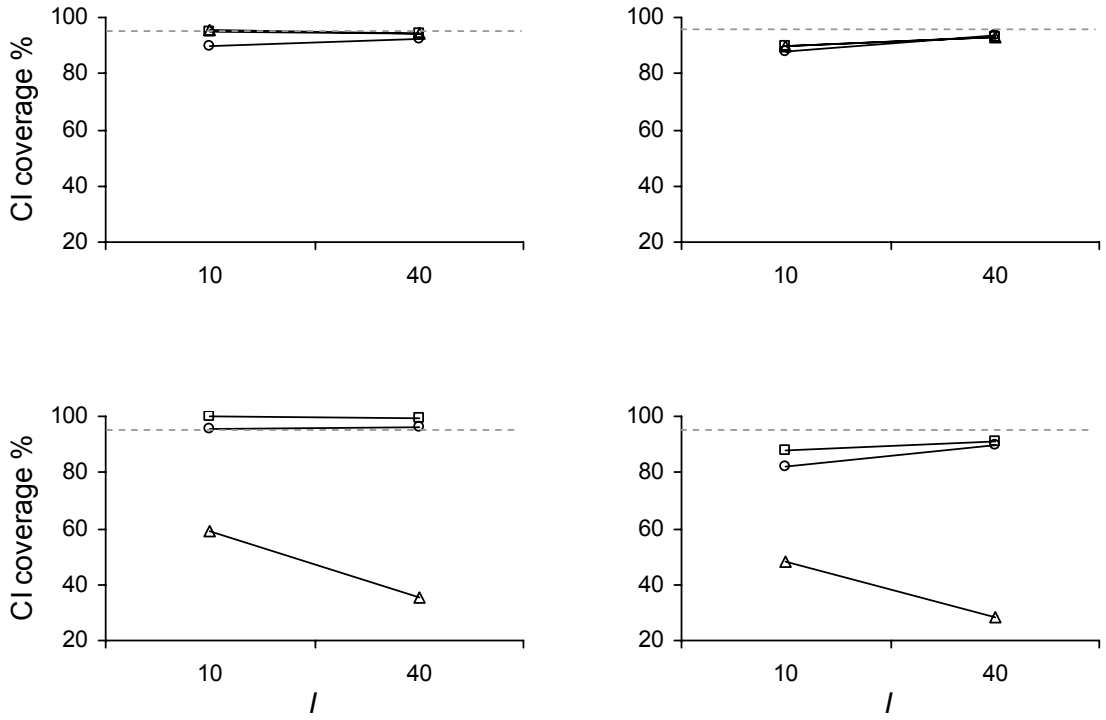


*Figure 5*. Coverage percentage for three methods' CIs for $M_{\Gamma 1}$, averaged over $\xi_2$ and $\phi$ for each combination of other design factors. Methods: delta method (triangle, △), effect-size bootstrap (square, □), cases bootstrap (circle, ○). Design factors: *I* (*x*-axis), $\xi_1$ ($0.05^2$ in left panels, $0.20^2$ in right panels), $\eta$ (0.8 in top panels, 0.0 in bottom panels). Reference lines at 95%.

By way of more detail regarding the influence of design factors, the four line plots in Figure 5 show coverage percentage for the delta-method and both bootstrap CIs for $M_{\Gamma 1}$ at each combination of *I*, $\eta$, and $\xi_1$, averaged over the four combinations of $\xi_2$ and

$\phi$—correlation factors that least affected CI coverage. (CIs for $M_{\Gamma 2}$ yielded similar patterns.) Most striking is the delta method's abysmal performance when $\eta = 0$, especially when $I = 40$. Also apparent is the typically lower (but sometimes nearer nominal) coverage for cases CIs than effect-size CIs. Figure 6 shows a similar set of four line plots for $M_{\Gamma 6}$, which typifies patterns for $M_{\Gamma 3}$, $M_{\Gamma 4}$, and $M_{\Gamma 5}$. Although all three

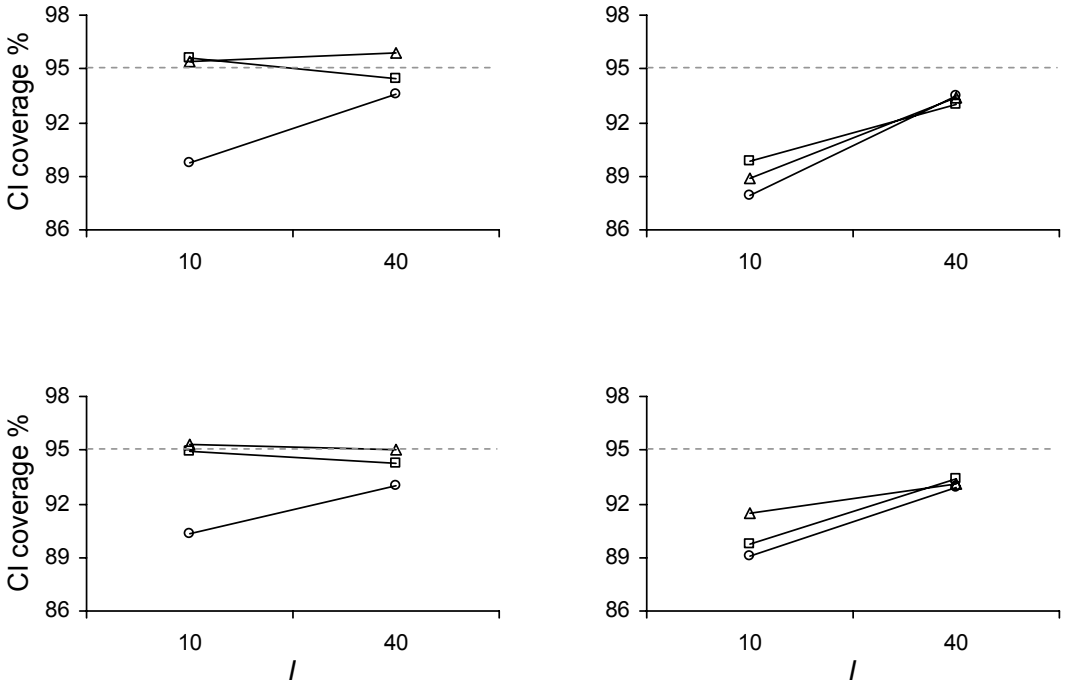


*Figure 6*. Coverage percentage for three methods' CIs for $M_{\Gamma 6}$, averaged over $\xi_2$ and $\phi$ for each combination of other design factors. Methods: delta method (triangle, △), effect-size bootstrap (square, □), cases bootstrap (circle, ○). Design factors: $I$ ($x$-axis), $\xi_1$ ($0.05^2$ in left panels, $0.20^2$ in right panels), $\eta$ (0.8 in top panels, 0.0 in bottom panels). Reference lines at 95%.

methods' CIs tended to cover $M_{\Gamma 6}$ less than nominal 95%, especially when $I = 10$, this was less pronounced for the delta method and effect-size bootstrap when $\xi_1 = 0.05^2$. This is most likely due to the combined effects of using standard-normal quantiles to construct CIs—making them too narrow with smaller $I$—and the latter two methods' heavier reliance on $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ for their standard errors (recall that $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ overestimates variances in $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ substantially when $\xi_1 = 0.05^2$ and underestimates them somewhat when $\xi_1 = 0.20^2$).

### 4.5. Summary of Monte Carlo Studies

Studies 1 and 2 constitute a preliminary assessment of the proposed estimation and inference techniques applied to a six-component function of ideal generic ESs in a five-factor simulation design. This assessment treated the function components as six separate scalars and focused on bias and MSE for point estimators and coverage probability for CIs. In this section I summarize each study's key findings.

**4.5.1. Study 1.** For most of the six component functions in most of Study 1's 243 conditions, standardized bias—and hence the contribution of bias to MSE—for the IT $\hat{M}_{\Gamma k}$ and the TS2 $\breve{M}_{\Gamma k}$ was negligibly small with no clear tendency to be positive or negative, showed little association with design factors, and tended to be only slightly larger than that for $\hat{M}_{\Theta j}$. Also, the IT and TS2 estimators of $M_{\Gamma k}$ yielded similar standardized bias and MSE, with neither showing a clear advantage for any function. The two quadratic functions, $\Gamma_1 = \Theta_1^2$ and $\Gamma_2 = \Theta_1\Theta_2$, yielded notable exceptions: Standardized bias for their IT and TS2 mean estimators was substantially more positive when $\eta = M_{\Theta j} = 0$ with small $\xi_1 = \Sigma_{\Theta jj}$ than in all other conditions, in which it was

comparable to standardized bias for other components of $\mathbf{M}_{\boldsymbol{\Gamma}}$. This aberrant performance occurs largely because when $M_{\Theta j} = 0$ the means of $\Gamma_1$ and $\Gamma_2$ depend mainly on $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$, whose diagonal is estimated with large positive bias when $\Sigma_{\Theta jj}$ is small.

Estimators of variances in $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ also tended to perform similarly to their $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ counterparts, but their standardized bias was larger than for estimators of $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\mathbf{M}_{\boldsymbol{\Gamma}}$ and exhibited more associations with design factors and more variation among components of $\boldsymbol{\Gamma}$. The most prominent and consistent pattern was that $\hat{\Sigma}_{\Gamma kk}$ and $\breve{\Sigma}_{\Gamma kk}$ exhibited substantial positive bias when $\xi_1 = 0.05^2$ but notable negative bias when $\xi_1 = 0.20^2$, emulating associations between $\xi_1$ and bias in $\widetilde{\Sigma}_{\Theta jj}$. Two patterns unique to particular components of $\boldsymbol{\Gamma}$ arose: Estimators of $\Sigma_{\Gamma 11}$ and $\Sigma_{\Gamma 22}$ (i.e., for quadratic functions of $\boldsymbol{\Theta}$) yielded markedly more positive standardized bias in some conditions where $\eta = 0$, and estimators of $\Sigma_{\Gamma 22}$ and $\Sigma_{\Gamma 44}$ (i.e., for functions of two components of $\boldsymbol{\Theta}$) yielded more standardized bias in one direction or the other when $\xi_2$ and $\phi$ were dissimilar. In some conditions the IT and TS2 estimators of $\Sigma_{\Gamma kk}$ differed notably in bias or MSE for the non-quadratic components of $\boldsymbol{\Gamma}$, mainly with larger $\xi_1$. These differences favored IT in some conditions but TS2 in others, and the more biased estimator tended to yield lower MSE.

CIs for $M_{\Gamma k}$ also performed much like their $M_{\Theta j}$ counterparts for most components of $\boldsymbol{\Gamma}$ in most conditions: They yielded similar coverage percentage, including poor performance in some conditions due largely to bias in $\widetilde{\Sigma}_{\Theta jj}$ and essentially treating $\widetilde{\Sigma}_{\Theta jj}$ as known (by ignoring its sampling variance in the delta method and using standard-

normal quantiles). As with point estimators of $M_{\Gamma k}$ and $\Sigma_{\Gamma kk}$, in some conditions CIs for $M_{\Gamma 1}$ and $M_{\Gamma 2}$ performed substantially differently than those for other components: CI coverage percentage for $M_{\Gamma 1}$ and $M_{\Gamma 2}$ was far below nominal 95% when $\eta = 0$, mainly because for these quadratic functions of $\boldsymbol{\Theta}$ the delta method severely underestimates the sampling variance of $\overline{M}_{\Gamma k}$ when $M_{\Theta j}$ is near a critical point of $\overline{M}_{\Gamma k}$ as a function of $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$. Finally, the IT and TS2 CIs did not exhibit notable differences in coverage.

**4.5.2. Study 2.** Each of the three bootstrap methods considered here was used to construct standard-normal CIs with bootstrap standard errors based on $B = 30$ bootstrap samples. The most consistent finding from comparisons among bootstrap methods is that in all 32 conditions the error bootstrap yielded somewhat to substantially lower coverage percentage than its effect-size and cases counterparts. Cases CIs usually yielded lower coverage percentage than effect-size CIs, but occasionally this lower percentage was nearer nominal 95%—primarily for the two quadratic functions ($\Gamma_1$ and $\Gamma_2$) when $\eta = M_{\Theta j} = 0$ with small $\xi_1 = \Sigma_{\Theta jj}$. Coverage percentage did not differ notably between IT and TS2 CIs for any bootstrap method.

Both the effect-size and cases CIs performed substantially better than their delta-method counterparts for $M_{\Gamma 1}$ and $M_{\Gamma 2}$ with $\eta = 0$, when the latter's coverage percentage was far below nominal. For $M_{\Gamma 1}$ and $M_{\Gamma 2}$ when $\eta = 0.8$ or for other components of $\boldsymbol{\Gamma}$, however, delta-method CIs tended to outperform their bootstrap counterparts, more so for the cases bootstrap than the effect-size bootstrap.

# 5. GENERAL DISCUSSION

The present thesis was motivated by a problem that arises with some frequency in applications of meta-analysis: Meta-analytic results are obtained by analyzing ESs in a metric chosen for convenience or to conform to assumptions of standard meta-analytic models, but one wishes to express certain of these results (e.g., point estimates or confidence regions for lower-order moments of ES parameters) in a different metric. My primary aims were to propose practically feasible techniques to accomplish such re-expression in a fairly general situation—vector-valued functions of multivariate ESs under a random-effects model—and evaluate selected aspects of these techniques' performance using Monte Carlo simulations. It is hoped that these proposed techniques and initial Monte Carlo evaluation thereof serve as a proof of concept and motivate further work on this and related problems. I begin this chapter by highlighting contributions of this thesis to meta-analytic methodology. I then comment on several limitations of the proposed techniques and the Monte Carlo studies, along with remarks on potentially fruitful directions for future research in this domain.

## 5.1. Overview of Contributions

Simply put, the statistical problem addressed in this thesis is to begin with estimates of the mean and covariance matrix of $\boldsymbol{\Theta}$, a random multivariate ES in a metric suitable for meta-analysis, and obtain estimates of the mean and covariance of $\boldsymbol{\Gamma} \equiv g(\boldsymbol{\Theta})$, a possibly vector-valued function of $\boldsymbol{\Theta}$. More specifically, we begin with $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ as meta-analytic estimators of $\mathbf{M}_{\boldsymbol{\Theta}} \equiv \mathrm{E}(\boldsymbol{\Theta})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} \equiv \mathrm{Cov}(\boldsymbol{\Theta})$, respectively, as well as

$\hat{\mathbf{M}}_{\Theta}$) as an estimate of Cov($\hat{\mathbf{M}}_{\Theta}$), and we wish to estimate $\mathbf{M}_{\Gamma} \equiv E(\boldsymbol{\Gamma})$ and $\boldsymbol{\Sigma}_{\Gamma} \equiv$ Cov($\boldsymbol{\Gamma}$). In this section I briefly overview the techniques proposed herein to accomplish this re-expression, including point estimators of $\mathbf{M}_{\Gamma}$ and $\boldsymbol{\Sigma}_{\Gamma}$ as well as inferences on $\mathbf{M}_{\Gamma}$, and draw some conclusions about their performance on the basis of findings from the Monte Carlo studies.

**5.1.1. Proposed techniques.** The proposed techniques consisted of estimation and inference procedures. The IT estimation approach involves an integral transformation of $\hat{\mathbf{M}}_{\Theta}$ and $\tilde{\boldsymbol{\Sigma}}_{\Theta}$ whereby $\mathbf{M}_{\Gamma}$ and $\boldsymbol{\Sigma}_{\Gamma}$ are estimated as their defining multidimensional integrals, whose integrands are the product of either $g(\boldsymbol{\Theta})$ or $[g(\boldsymbol{\Theta}) - \mathbf{M}_{\Gamma}][g(\boldsymbol{\Theta}) - \mathbf{M}_{\Gamma}]^{T}$ and the density of $\boldsymbol{\Theta}$ (e.g., multivariate normal). Implementing this approach essentially entails substituting $\hat{\mathbf{M}}_{\Theta}$ and $\tilde{\boldsymbol{\Sigma}}_{\Theta}$ for this latter density's parameters and evaluating the integrals (e.g., by Monte Carlo methods) to obtain the estimators $\hat{\mathbf{M}}_{\Gamma}$ and $\hat{\boldsymbol{\Sigma}}_{\Gamma}$.

By comparison, the TS2 estimation approach involves approximating $g(\boldsymbol{\Theta})$ by a second-order Taylor polynomial and taking this polynomial's expected value and covariance matrix as approximations of the corresponding moments of $\boldsymbol{\Gamma}$. These approximating moments can be expressed in terms of $\mathbf{M}_{\Theta}$ and $\boldsymbol{\Sigma}_{\Theta}$, with certain distributional assumptions required for the covariance matrix. Implementing this approach entails computing derivatives for gradient vectors and Hessian matrices and substituting $\hat{\mathbf{M}}_{\Theta}$ and $\tilde{\boldsymbol{\Sigma}}_{\Theta}$ for their estimands to obtain the estimators $\breve{\mathbf{M}}_{\Gamma}$ and $\breve{\boldsymbol{\Sigma}}_{\Gamma}$.

The essence of the IT and TS2 approaches is relatively easy to describe to non-technical audiences who are likely to use them, and they are fairly easy to implement in realistic circumstances, especially if simple numerical methods are used to evaluate integrals and derivatives. (Although many potential users will not have resources to program the computations, this barrier to usage will be reduced substantially if the techniques are implemented in accessible software.) Furthermore, because both the IT and TS2 approaches are applied directly to $\hat{\mathbf{M}}_{\Theta}$ and $\widetilde{\mathbf{\Sigma}}_{\Theta}$, neither requires access to the original meta-analytic data. This is especially valuable when re-expressing results from a previous meta-analysis for which the original data are not available but $\hat{\mathbf{M}}_{\Theta}$ and $\widetilde{\mathbf{\Sigma}}_{\Theta}$ are—a likely scenario when results from a meta-analytic review are used for purposes its authors did not anticipate, such as evaluating novel models whose parameters depend on $\mathbf{\Theta}$, conveying findings to diverse groups of stakeholders, or incorporating review outcomes into decision analyses to guide policy.

As for inferences on $\mathbf{M}_{\Gamma}$, the primary approach I proposed relies on the multivariate delta method to estimate an approximation of $\text{Cov}(\hat{\mathbf{M}}_{\Gamma})$ or $\text{Cov}(\breve{\mathbf{M}}_{\Gamma})$ from $\hat{\mathbf{M}}_{\Theta}$, $\widetilde{\mathbf{\Sigma}}_{\Theta}$, and $\hat{\text{Cov}}(\hat{\mathbf{M}}_{\Theta})$. More specifically, this delta-method application uses a Taylor polynomial of order one to approximate either $\hat{\mathbf{M}}_{\Gamma}$ or $\breve{\mathbf{M}}_{\Gamma}$ as a function of $\hat{\mathbf{M}}_{\Theta}$, treating $\widetilde{\mathbf{\Sigma}}_{\Theta}$ as known. This approximation involves derivatives for gradient vectors, which in many applications will be best obtained numerically. The resulting covariance matrix for $\hat{\mathbf{M}}_{\Gamma}$ or $\breve{\mathbf{M}}_{\Gamma}$ may then be used to construct confidence regions for or test hypotheses

about one or more components of $\boldsymbol{\Gamma}$. Like the IT and TS2 estimators themselves, this delta-method approach does not require the original meta-analytic data. To facilitate its application, authors of meta-analytic reviews should be encouraged to report $\hat{\text{Cov}}(\hat{\mathbf{M}}_{\Theta})$ or equivalent results (instead of, e.g., only standard errors for $\hat{\mathbf{M}}_{\Theta}$ elements).

Finally, as an alternative to delta-method inference I briefly described bootstrap procedures to estimate the sampling distribution of $\hat{\mathbf{M}}_{\Gamma}$ or $\breve{\mathbf{M}}_{\Gamma}$, which in turn can be used to construct confidence regions for or test hypotheses about $\mathbf{M}_{\Gamma}$ by either obtaining a covariance matrix or using other approaches (e.g., percentiles of bootstrap distribution). I specifically described four different bootstrap resampling methods, each of which is an extension of previously proposed bootstrap techniques for meta-analysis. These bootstrap approaches may outperform their delta-method counterparts in certain respects under some conditions, especially when features of one's data sufficiently violate key assumptions of the delta method. Although these potential advantages come at the cost of additional computational burden as well as requiring the original data, for some users these barriers will be outweighed by higher confidence in the validity of inferences.

**5.1.2. Conclusions from Monte Carlo findings.** Both Monte Carlo studies of the proposed techniques were summarized in the previous chapter. On the basis of those findings, some tentative conclusions about these techniques are warranted. It is worth bearing in mind that the present Monte Carlo studies' simulated data conformed closely to the standard multivariate random-effects model (e.g., normal $\mathbf{T}_i$ with known $\boldsymbol{\Psi}_{\mathbf{T}i}$) and satisfied certain assumptions required by some of the proposed techniques (e.g., normal

$\boldsymbol{\Theta}$). In a sense, these data represent an ideal scenario under which most of the proposed techniques should perform at their best. In subsequent sections I elaborate on this limitation and others and suggest additional simulation studies.

Broadly speaking, the proposed techniques seem to perform reasonably well in a variety of conditions. Namely, the point estimators for $M_{\boldsymbol{\Gamma}k}$ and $\Sigma_{\boldsymbol{\Gamma}kk}$ and CIs for $M_{\boldsymbol{\Gamma}k}$ often performed nearly as well as their counterparts for $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$, in terms of standardized bias for point estimators and coverage probability for CIs. Although bias per se may be of only limited interest, standardized bias is of greater interest in that it reflects the contribution of bias to MSE. I am not aware of a convenient way to compare MSE itself between estimators of different functions or in different metrics. At any rate, this similar performance in the $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ metrics of point estimators and CIs indicates that one can apply the proposed techniques to finite meta-analytic samples without degrading markedly the selected statistical properties I examined.

That said, the performance of these estimators and CIs was sometimes mediocre or poor. The most troubling cases, however, were usually isolated to certain conditions, either for a particular meta-analytic task or for certain components of $\boldsymbol{\Gamma}$. For example, estimators of both $\Sigma_{\boldsymbol{\Theta}jj}$ and $\Sigma_{\boldsymbol{\Gamma}kk}$ exhibited substantial positive standardized bias when $\Sigma_{\boldsymbol{\Theta}jj}$ was small, but this was usually not more pronounced for $\boldsymbol{\Gamma}$ than for $\boldsymbol{\Theta}$. More problematic for the proposed transformation techniques was its unusually poor performance for the quadratic functions $\Gamma_1$ and $\Gamma_2$ when $M_{\boldsymbol{\Theta}j} = 0$: problematically large positive bias in point estimators of $M_{\boldsymbol{\Gamma}1}$ and $M_{\boldsymbol{\Gamma}2}$ (mainly with smaller $\Sigma_{\boldsymbol{\Theta}jj}$) and in point estimators of $\Sigma_{\boldsymbol{\Gamma}11}$ and

$\Sigma_{\Gamma 22}$ (mainly with larger $\Sigma_{\Theta jj}$), as well as egregiously low coverage probability for CIs for $M_{\Gamma 1}$ and $M_{\Gamma 2}$, especially with more studies and larger $\Sigma_{\Theta jj}$. These pronounced exceptions for $\Gamma_1$ and $\Gamma_2$ serve as reminders that the proposed techniques must be used with caution for some types of functions in certain conditions.

Regarding the two estimation methods I have proposed, under the situations studied here the choice between IT versus TS2 does not seem to be important for point estimation of $M_{\Gamma k}$ or CI construction for $M_{\Gamma k}$ using either delta-method or bootstrap standard errors. These two methods' similar performance is not surprising for point and interval estimators of $M_{\Gamma 1}$ and $M_{\Gamma 2}$, for which IT and TS2 differed mainly by minor numerical methods (e.g., integration and differentiation strategies); indeed, they performed nearly identically for these mean estimators as well as for point estimators of $\Sigma_{\Gamma 11}$ and $\Sigma_{\Gamma 22}$. However, more noticeable differences between IT and TS2 arose for point estimators of $\Sigma_{\Gamma 33}$ through $\Sigma_{\Gamma 66}$, favoring IT in some situations but TS2 in others. Potential users for whom estimation of $\Sigma_{\Gamma kk}$ is critical are advised to consider carefully the function being estimated and the conditions at hand when choosing between these estimators. Note, too, that in some applications one might wish to choose between the IT and TS2 estimators on the basis of statistical or other properties not considered here, such as operating characteristics of particular hypothesis tests, computational speed and burden, or ease of implementation for specific types of users.

The choice among methods for CI construction is not straightforward. Delta-method CIs for $M_{\Gamma k}$ perform reasonably well in many situations where CIs for $M_{\Theta j}$

perform well, which indicates that this approach is often appropriate for re-expressing results for $\boldsymbol{\Theta}$ hyperparameters as CIs in the $\boldsymbol{\Gamma}$ metric. In some conditions, however, even CIs for $M_{\Theta j}$ perform poorly, such as with few studies and larger $\Sigma_{\Theta jj}$, in which case it is difficult to justify delta-method CIs even though their poor performance may not be due to the delta method per se. Furthermore, delta-method CIs for $M_{\Gamma k}$ are clearly inappropriate for some functions in particular conditions, such as for $\Theta_j\Theta_l$ when $M_{\Theta j} = M_{\Theta l} = 0$, as demonstrated by their poor performance for $\Gamma_1 = \Theta_1^2$ and $\Gamma_2 = \Theta_1\Theta_2$.

It is not clear whether bootstrap CIs can be recommended as a general-purpose alternative to delta-method CIs. Although all three bootstrap methods examined in the present simulations performed better than the delta-method CIs in the latter's most adverse situations (i.e., for $M_{\Gamma 1}$ and $M_{\Gamma 2}$ when $M_{\Theta j} = 0$), only two of them—the effect-size and cases bootstrap methods—maintained acceptable coverage probability in these situations. The error bootstrap performed unacceptably poorly in most situations. Moreover, aside from the above situation involving quadratic functions of $\boldsymbol{\Theta}$, the effect-size and cases bootstrap CIs did not perform better than their delta-method counterparts for most functions and conditions considered herein, and in some of these situations one or both of these bootstrap methods performed markedly worse than the delta method, especially the cases bootstrap. As I elaborate on below, recommending or discouraging particular CI methods on the basis of the present simulations would be premature, largely because numerous variants are available that might improve CI performance substantially—especially for bootstrap methods.

## 5.2. Limitations and Future Directions

The techniques I have proposed address some important tasks likely to be of interest when re-expressing meta-analytic results in another metric, and they do so for a fairly general case likely to encompass many potential users' situations. Two Monte Carlo studies suggested that these procedures perform acceptably well for a number of functions in a wide variety of conditions, with some notable exceptions. Nevertheless, the proposed techniques are limited in important ways, as are the Monte Carlo studies. In this section I comment on several such limitations and offer several thoughts on future research to address these limitations and other matters.

**5.2.1. Proposed techniques.** Some aspects of the proposed techniques' unacceptable performance in some situations might be remedied by various minor changes to the techniques as implemented in the simulations. For example, the EM-GLS estimators of $\mathbf{M}_\Theta$, $\mathbf{\Sigma}_\Theta$, and the former's sampling covariance matrix seem to exhibit certain deficiencies in some conditions, most notably the biased estimation of $\mathbf{\Sigma}_\Theta$. Other approaches to estimating these quantities (e.g., see Section 2.1.3) could in turn improve estimation and inference for $\mathbf{M}_\Gamma$ and $\mathbf{\Sigma}_\Gamma$ by the proposed techniques. Another potentially useful change is to construct CIs using Student-*t* quantiles (or *F* for multivariate confidence regions) instead of standard-normal (or chi-squared) quantiles, which could be especially helpful with fewer studies (e.g., Harbord et al., 2007; Hartung & Knapp, 2001). One could also investigate a number of alternative strategies for secondary estimation tasks (e.g., other estimators for ES parameters, $\boldsymbol{\theta}_i$, used to estimate $\mathbf{u}_i$ and $\mathbf{e}_i$ in the error bootstrap; number of bootstrap samples, *B*) or for numerical methods (e.g.,

evaluation of multidimensional integrals for IT estimators, evaluation of derivatives for TS2 estimators and delta-method covariance matrices).

More substantial changes to the proposed techniques may also improve performance on the statistical tasks considered herein. One notable limitation of the proposed IT estimators is their assumption that $\boldsymbol{\Theta}$ is normal. In some situations one might wish to consider other distributional forms for $\boldsymbol{\Theta}$, such as parametric univariate or multivariate families governed by $\mathbf{M}_{\boldsymbol{\Theta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$, or other hyperparameters. Estimating the density of $\boldsymbol{\Theta}$ by nonparametric methods may also be feasible (e.g., Louis, 1984). The CTS2 estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ also depends on normality of $\boldsymbol{\Theta}$; alternative estimators that rely on weaker distributional assumptions, such as quasi-normality, might be worth pursuing. In a similar vein, for some functions a simpler first-order Taylor series approximation—only the first term on the right-hand sides of Equations 13 and 19—might perform reasonably well in some conditions and require fewer assumptions for estimating $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$.

As for substantial changes to inference techniques, the proposed delta-method approach is limited by its treatment of $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ as fixed and known in approximating the covariance matrix for an estimator of $\mathbf{M}_{\boldsymbol{\Gamma}}$. Besides the use of Student-*t* or *F* quantiles as a somewhat ad hoc approach to address this limitation, another approach is to incorporate the covariance matrix of both $\hat{\mathbf{M}}_{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$ into the delta-method approximation (perhaps without covariances involving $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\Theta}}$). Bootstrap inference may also incorporate uncertainty about $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ into inferences on $\mathbf{M}_{\boldsymbol{\Gamma}}$, but in conditions where this uncertainty is large (e.g., with few studies) alternative methods for constructing CIs may work better

than symmetric CIs based on bootstrap standard errors (e.g., equal-tail percentiles, bias-corrected and accelerated percentiles). For some classes of functions, such as most monotonic scalar functions of one component of $\boldsymbol{\Theta}$, approaches such as the delta-method and bootstrap may be bypassed by simply applying the MIT or MTS2 directly to endpoints of a CI for $\mathbf{M}_{\boldsymbol{\Theta}}$.

It is worth noting other statistical tasks that are potentially of interest but not handled by the proposed techniques. Some that seem most likely to arise in practice include inferences on $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ or on both $\mathbf{M}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$ jointly (e.g., confidence regions or tests) and prediction regions for $\boldsymbol{\gamma}_{I+1}$ (i.e., $\boldsymbol{\gamma}$ for the next study; see, e.g., Harbord et al., 2007, for examples in the case of bivariate logits). In most practical circumstances $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} = \mathbf{0}$ iff $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} = \mathbf{0}$, so testing $H_0$: $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} = \mathbf{0}$ also tests $H_0$: $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} = \mathbf{0}$. Hence, confidence regions for $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$—such as several authors have promoted in simpler cases (e.g., Biggerstaff & Tweedie, 1997; Hardy & Thompson, 1996; Knapp, Biggerstaff, & Hartung, 2006; Tian, 2008; Viechtbauer, 2007)—are likely to be of more interest than tests. Perhaps bootstrap procedures for inferences about $\mathbf{M}_{\boldsymbol{\Gamma}}$ could be adapted for inferences about $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}$.

It would also be valuable to extend the proposed techniques to models that include study-level covariates to account for heterogeneity in ES parameters (e.g., Kalaian & Raudenbush, 1996; Berkey et al., 1998). A more dramatic modification of the proposed techniques would be to adopt a Bayesian approach to multivariate random-effects models, such as Prevost et al. (2007) proposed for the case of correlation matrices or Nam et al. (2003) presented for a more general case of generic ESs. Despite their

additional complexity (e.g., choice of prior distributions) and computational demand, Bayesian techniques may be well-suited to address some of the proposed techniques' limitations (e.g., assumed distributional form for $\boldsymbol{\Theta}$, neglected uncertainty about $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$).

Finally, a number of issues that might arise in specific applications remain to be addressed. For example, it would be valuable to delineate systematically classes of functions for which the proposed techniques are most appropriate and to identify scenarios in which otherwise permissible functions may lead to problematic performance, such as quadratic functions of $\boldsymbol{\Theta}$ near critical points. The proposed techniques may handle a larger class of functions with certain modifications (e.g., delta method based on second-order Taylor polynomial). On a related note, the proposed techniques may require adaptations when used with particular choices of $\boldsymbol{\Theta}$ or $g$. For example, when $\boldsymbol{\Theta}$ represents a matrix of Pearson correlations or Fisher $z$-transforms, the support for multivariate-normal $\boldsymbol{\Theta}$ (i.e., $\boldsymbol{\Theta} \in \mathbb{R}^J$) would extend outside the space of admissible correlation matrices, which is theoretically dubious and may cause computational problems (e.g., integrating over $\boldsymbol{\Theta}$ for IT estimators). Similarly, for bounded functions symmetric CIs based on delta-method or bootstrap standard errors may yield inadmissible endpoints (e.g., for $\Gamma_1$, $\Gamma_3$, $\Gamma_4$, and $\Gamma_6$ in the present Monte Carlo studies).

**5.2.2. Monte Carlo studies.** As with most Monte Carlo studies, several choices made largely for convenience limited the generalizability of results to potential users' situations. One such choice is the three (or two) particular levels of each design factor: $I$, $\eta$, $\xi_1$ (especially its ratio with $\boldsymbol{\Psi}_{\mathbf{T}ijj}$), $\xi_2$, and $\phi$. Some applications may fall too far outside

of the design space to warrant useful extrapolation of the present findings. Another limitation is that ideal generic ES data were generated to conform exactly to the multivariate random-effects model and to the $\boldsymbol{\Theta} \sim N$ assumption underlying certain proposed techniques. More informative for practical applications would be simulations that use realistic ES metrics, such as correlations, SMDs, or logits. For many of these, $\boldsymbol{\Psi}_{\mathbf{T}i}$ and the sampling distribution's shape depend on $\boldsymbol{\theta}_i$ in addition to sample size(s), and some of them may require changes to the proposed techniques (e.g., to handle inadmissible values of $\boldsymbol{\theta}_i$ or $\mathbf{t}_i$). The impact of non-normal distributions of $\boldsymbol{\Theta}$ is also of interest. Other features of the simulated data that might influence some techniques' performance include the number of $\boldsymbol{\Theta}$ components ($J$); the patterns of values in $\mathbf{M}_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$, which were conveniently simple in the present simulations; and aspects of within-study sample size(s). As alluded to in Chapter 2, a wide variety of functions $g$ could also be investigated, varying in both the number ($K$) and nature of component functions. Identifying classes of functions for which certain techniques perform similarly would be valuable. (It might be interesting to examine whether the proposed techniques' performance notably improves or deteriorates if applied to each component of $\boldsymbol{\Gamma} \equiv g(\boldsymbol{\Theta})$ separately instead of collectively as a vector.)

The present Monte Carlo studies included only the proposed techniques, and even then only particular versions of them (e.g., central 95% CIs based on standard-normal quantiles). In addition to examining other inference procedures (e.g., 1-sided confidence bounds, multivariate confidence regions, tests of hypotheses about $\mathbf{M}_{\boldsymbol{\Gamma}}$) as well as some

of the minor and major variants of the proposed techniques mentioned in the previous section, one could investigate competing techniques that may be easier without marked detriments in performance. Among these are direct analysis of observed functions (e.g., EM-GLS applied to $\mathbf{g}_i = g(\mathbf{t}_i)$, as described in Section 3.1), first-order Taylor series approximation instead of TS2 (e.g., $g(\hat{\mathbf{M}}_{\Theta})$ as an estimator of $\mathbf{M}_{\Gamma}$), EM-GLS with diagonal $\boldsymbol{\Psi}_{\mathbf{T}i}$ or $\boldsymbol{\Sigma}_{\Theta}$ (e.g., Hafdahl, 2001; Becker, 2009), or estimation of and inferences on $\mathbf{M}_{\Gamma}$ using the known $\boldsymbol{\Sigma}_{\Theta}$—unrealistic in practice, but potentially informative about the impact of estimating $\boldsymbol{\Sigma}_{\Theta}$. For users interested in the homogeneous fixed-effects case (see Section 3.5.1), simulations under that model would be valuable. These would be considerably simpler to implement than in the random-effects case (e.g., M. W.-L. Cheung & Chan, 2005; Furlow & Beretvas, 2005; Hafdahl, 2001, 2007).

One might also consider additional evaluation criteria neglected in the present simulations. For instance, evaluate multivariate properties of point or interval estimators may be of interest, such as by using a multivariate measure of standardized bias for an estimator of $\mathbf{M}_{\Gamma}$ or $\boldsymbol{\Sigma}_{\Gamma}$. Also, a meaningful way to compare MSE between corresponding estimators in the $\Theta$ and $\Gamma$ metrics (e.g., estimators of $\mathbf{M}_{\Theta}$ and $\mathbf{M}_{\Gamma}$) would permit assessing the proposed techniques' influence on efficiency. CIs could also be evaluated in other ways, such as by their typical width or variability in width. Deriving these or other statistical properties analytically would be preferable to simulation evidence, but analytic derivations for meta-analytic procedures tend to be intractable except in unrealistically simplified scenarios (e.g., ideal generic ESs, $\boldsymbol{\Psi}_{\mathbf{T}i} = \boldsymbol{\Psi}_{\mathbf{T}} \; \forall \; i$, known $\boldsymbol{\Sigma}_{\Theta}$).

**5.2.3. Accessibility for applied researchers.** Many potential users of the techniques I have proposed are applied researchers who are unlikely to understand and implement the techniques without further support. For example, many such researchers in the behavioral, social, or health sciences have taken only one or two graduate courses in applied statistics and are not proficient in any programming language. To increase such researchers' interest in and responsible usage of these techniques, at least three types of additional resources would be valuable. First, didactic presentations (e.g., articles, conference papers, Internet resources) that include real-data examples and emphasize practical decisions and interpretations will help applied researchers appreciate the proposed techniques' value and provide guidance in using them. For instance, Hafdahl (2009b, 2009d) used meta-analytic data from studies in social-cognitive psychology and sports psychology to demonstrate some of the proposed techniques applied to path models as functions of correlation matrices—one of the most popular uses of multivariate meta-analysis to date (though bivariate meta-analysis of true- and false-positive rates for diagnostic tests is a competitor).

Second, providing user-friendly software that implements the proposed techniques will greatly reduce computational barriers for many applied researchers. Such software might consist of freely available executable programs, scripts or macros for popular statistical or mathematical software (e.g., SAS, SPSS, R, Matlab), interactive Internet applets, or modules incorporated into existing meta-analysis software (for reviews see, e.g., Bax, Yu, Ikeda, & Moons, 2007; Sterne, Egger, & Sutton, 2001). A

major challenge of producing such software is to incorporate a variety of ES metrics and functions to suit the needs of diverse potential users.

Third, software that could run Monte Carlo simulations tailored to one's data and desired statistical tasks would facilitate choosing among variants of the techniques and deciding whether they work sufficiently well in one's unique circumstances. As an example of this idea, the Mplus software package for structural equation modeling includes a Monte Carlo feature to permit custom simulations. Such software could also be made available in several different ways, and it could be used for a variety of tasks such as power analysis, sample-size planning, and sensitivity analyses.

## REFERENCES


Abrams, K., & Jones, D. R. (1995). Meta-analysis and the synthesis of evidence. IMA *Journal of Mathematics Applied in Medicine & Biology, 12,* 297-313.

Ades, A. E. (2003). A chain of evidence with mixed comparisons: Models for multi-parameter synthesis and consistency of evidence. *Statistics in Medicine, 22,* 2995-3016.

Altman, D. G., & Deeks, J. J. (2002). Meta-analysis, Simpson's paradox, and the number needed to treat. *BMC Medical Research Methodology, 2,* 3.

Arends, L. R., & Voko, Z., & Stijnen, T. (2003). Combining multiple outcome measures in a meta-analysis: An application. *Statistics in Medicine, 22,* 1335-1353.

Arends, L. R., Hunink, M. G. M., & Stijnen, T. (2008). Meta-analysis of summary survival curve data. *Statistics in Medicine, 27,* 4381-4396.

Ballesteros, J. (2005). Orphan comparisons and indirect meta-analysis: A case study on antidepressant efficacy in dysthymia comparing tricyclic antidepressants, selective serotonin reuptake inhibitors, and monoamine oxidase inhibitors by using general linear models. *Journal of Clinical Psychopharmacology, 25,* 127-131.

Bax, L., Yu, L. M., Ikeda, N., & Moons, K. G. M. (2007). A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology, 7,* 40.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical & Statistical Psychology, 41,* 257-278.

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics, 17,* 341-362.

Becker, B. J. (1995). Corrections to "Using results from replicated studies to estimate linear models". *Journal of Educational and Behavioral Statistics, 20,* 100-102.

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499-525). San Diego, CA: Academic Press.

Becker, B. J. (2009). Model-based meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 377-395). New York: Russell Sage Foundation.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357-381). New York: Russell Sage Foundation.

Becker, B. J., & Wu, M.-J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science, 22,* 414-429.

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

Biggerstaff, B. J., & Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine, 16,* 753-768.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8,* 406-418.

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

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

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society-Series C, 52,* 431-443.

Cates, C. J. (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis. *BMC Medical Research Methodology, 2,* 1.

Chalmers, I., Hedges, L., V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the Health Professions, 25,* 12-37.

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods, 10,* 40-64.

Cheung, S. F. (2000). Examining solutions to two practical issues in meta-analysis: Dependent correlations and missing data in correlation matrices. *Dissertation Abstracts International, 61*(08), 4469B. (UMI No. AAI9984691).

Cohen, J. (1982). Set correlation as a general multivariate data-analytic method. *Multivariate Behavioral Research, 17,* 301-341.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Conn, V. S., Hafdahl, A. R., Brown, S. A., & Brown, L. M. (2008). Meta-analysis of patient education interventions to increase physical activity among chronically ill adults. *Patient Education & Counseling, 70,* 157-172.

Conn, V. S., Hafdahl, A. R., Cooper, P. S., Brown, L. M., & Lusk, S. L. (2009). Meta-analysis of workplace physical activity interventions. *American Journal of Preventive Medicine, 37,* 330-339.

Conn, V. S., Hafdahl, A. R., Cooper, P. S., Ruppar, T. M., Mehr, D. R., & Russell, C. L. (2009). Interventions to improve medication adherence among older adults: Meta-analysis of adherence outcomes among randomized controlled trials. *Gerontologist, 49,* 447-462.

Conn, V. S., Hafdahl, A. R., Minor, M. A., & Nielsen, P. J. (2008). Physical activity interventions among adults with arthritis: Meta-analysis of outcomes. *Seminars in Arthritis and Rheumatism, 37,* 307-316.

Conn, V. S., Hafdahl, A. R., Porock, D. C., McDaniel, R., & Nielsen, P. J. (2006). A meta-analysis of exercise interventions among people treated for cancer. *Supportive Care in Cancer, 14,* 699-712.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317-327.

DerSimonian, R., & Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials, 28,* 105-114.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7,* 177-188.

Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin, 116,* 509-511.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). New York: Russell Sage Foundation.

Furlow, C. F., & Beretvas, S. N. (2005). Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. *Psychological Methods, 10,* 227-254.

Games, P. A., & Hedges, L. V. (1987). Multifactor analyses on proportions, variances, correlations, and standardized mean differences for independent groups. *Journal of Experimental Education, 56,* 15-23.

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

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York: Russell Sage Foundation.

Gleser, L. J., & Olkin, I. (2000). Meta-analysis for 2 × 2 tables with multiple treatment groups. In D. A. Stangl & D. A. Berry (Eds.), *Meta-analysis in medicine and health policy* (pp. 179-189). New York: Marcel Dekker.

Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology, 123,* 203-208.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mawah, NJ: Lawrence Erlbaum.

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3,* 339-353.

Hafdahl, A. R. (2001). Multivariate meta-analysis for exploratory factor analytic research (Doctoral dissertation, The University of North Carolina at Chapel Hill, 2001). *Dissertation Abstracts International, 62*(08), 3843B.

Hafdahl, A. R. (2004, June). *Refinements for random-effects meta-analysis of correlation matrices.* Paper presented at the meeting of the Psychometric Society, Monterey, CA.

Hafdahl, A. R. (2007). Combining correlation matrices: Simulation analysis of improved fixed-effects methods. *Journal of Educational and Behavioral Statistics, 32,* 180-205.

Hafdahl, A. R. (2008). Combining heterogeneous correlation matrices: Simulation analysis of fixed-effects methods. *Journal of Educational and Behavioral Statistics*, 33, 507-533.

Hafdahl, A. R. (2009a). Improved Fisher *z* estimators for univariate random-effects meta-analysis of correlations. *British Journal of Mathematical and Statistical Psychology, 62,* 233-261.

Hafdahl, A. R. (2009b, May). Meta-analysis for functions of dependent correlations. In A. R. Hafdahl (Chair), *Advances in meta-analysis for multivariable linear models.* Invited symposium presented at the meeting of the Association for Psychological Science, San Francisco, CA.

Hafdahl, A. R. (2009c). Random-effects meta-analysis of correlations: Monte Carlo evaluation of mean estimators. *British Journal of Mathematical and Statistical Psychology.* Advance online publication. doi:10.1348/000711009X431914..

Hafdahl, A. R. (2009d, July). *Meta-analysis for functions of effect sizes.* Paper presented at the meeting of the Society for Research Synthesis Methodology, Seattle, WA.

Hafdahl, A. R., & Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods, 14,* 24-42.

Hamza, T. H., van Houwelingen, H. C., & Stijnen, T. (2008). The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology, 61,* 41-51.

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P., & Sterne, J. A. C. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics, 8,* 239-251.

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine, 15,* 619-629.

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine, 20,* 1771-1782.

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203-217.

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

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational & Psychological Measurement, 60,* 543-563.

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

Kalaian, H. A. (1994). A multivariate mixed linear model for meta-analysis. *Dissertation Abstracts International, 55*(12), 3689A.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1,* 227-235.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Knapp, G., Biggerstaff, B. J., & Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal, 48,* 271-285.

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry, 59,* 990-996.

Law, K. S. (1995). The use of Fisher's *Z* in Schmidt-Hunter–type meta-analyses. *Journal of Educational and Behavioral Statistics, 20,* 287-316.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Lorig, K., Feigenbaum, P., Regan, C., Ung, E., Chastain, R. L., & Holman, H. R. (1986). A comparison of lay-taught and professional-taught arthritis self-management courses. *Journal of Rheumatology, 13,* 763-767.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association, 79,* 393-398.

Lu, G., & Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine, 23,* 3105-3124.

Mason, C., Allam, R., & Brannick, M. T. (2007). How to meta-analyze coefficient-of-stability estimates: Some recommendations based on Monte Carlo studies. *Educational and Psychological Measurement, 67,* 765-783.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361-365.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7,* 105-125.

Nam, I.-S., Mengersen, K., & Garthwaite, P. (2003). Multivariate meta-analysis. *Statistics in Medicine, 22,* 2309-2333.

Olkin, I., & Finn, J. (1990). Testing correlated correlations. *Psychological Bulletin, 108,* 330-333.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin, 118,* 155-164.

Olkin, I., & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ikeda, T. Hayakawa, H. Hudimoto, M. Okamoto, M. Siotani, & S. Yamamoto (Eds.), *Essays in probability and statistics: A volume in honor of Professor Junjiro Ogawa* (pp. 235-251). Wakaba, Tokyo: Shinko Tsusho.

Prevost, A. T., Mason, D., Griffin, S., Kinmonth, A.-L., Sutton, S., & Spiegelhalter, D. (2007). Allowing for correlations between correlations in random-effects meta-analysis of correlation matrices. *Psychological Methods, 12,* 434-450.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295-315). New York: Russell Sage Foundation.

Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology, 58,* 982-990.

Richards, J. M., Jr. (1982). Standardized versus unstandardized regression weights. *Applied Psychological Measurement, 6,* 201-212.

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., & Thompson, J. R. (2007). Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology, 7,* 3.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166-169.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529-540.

Schott, J. R. (1997). *Matrix analysis for statistics.* New York: Wiley.

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

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 257-277). New York: Russell Sage Foundation.

Sinclair, J. C., & Bracken, M. B. (1994). Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology, 47,* 881-889.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245-251.

Sterne, J. A. C., Egger, M., & Sutton, A. J. (2001). Meta-analysis software. In M. Egger, G. D. Smith, & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 336-346). London: BMJ Publishing Group.

Thompson, K. N., & Schumacker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Educational & Behavioral Statistics, 22,* 109-117.

Tian, L. (2008). Inferences about the between-study variance in meta-analysis with normally distributed outcomes. *Biometrical Journal, 50,* 248-256.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational & Psychological Measurement, 58,* 6-20.

Van den Noortgate, W., & Onghena, P. (2005). Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods, 37,* 11-22.

van Houwelingen, H. C., Zwinderman, K. H., & Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine, 12,* 2273-2284.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational & Behavioral Statistics, 25,* 101-132.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26,* 37-52.

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials.* Chichester, West Sussex: John Wiley & Sons.