2011

The Impact of Multiple Endpoint Dependency on Homogeneity Measures in Meta-Analysis

Christopher Thompson

THE FLORIDA STATE UNIVERSITY

COLLEGE OF EDUCATION

THE IMPACT OF MULTIPLE ENDPOINT DEPENDENCY ON HOMOGENEITY MEASURES IN META-ANALYSIS

By

CHRISTOPHER THOMPSON

A Thesis submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Master of Science

Degree Awarded:
Summer Semester, 2011

The members of the committee approve the thesis of Christopher Thompson defended on June 13th, 2011.

_____
Betsy Becker
Professor Directing Thesis

_____
Yanyun Yang
Committee Member

_____
Daniel McGee
Committee Member

Approved:

_____
Betsy Becker, Chair, Educational Psychology and Learning Systems

The Graduate School has verified and approved the above-named committee members.

This work is dedicated to Lester and Elizabeth Scoggins

ACKNOWLEDGEMENTS

This thesis was in no way an independent endeavor. First and foremost, I would like to thank Dr. Becker for her endless insight and support during the entire process; your patience was treasured and your help was monumental. I would also like to thank the other members of my committee, Dr. Yang and Dr. McGee, for their supportive comments and suggestions. In addition, Sunny, Bernd, and Sanghyun provided much assistance with SAS and R syntax, as well as theoretical and grammatical considerations. My appreciation goes to the three of you, as well as the rest of the SynRG group. Not to place them too far down the list, but my parents also are deserving of gratitude for their continual multidimensional support. Last but far away from least, I would like to thank Kelly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Multivariate meta-analysis refers to the statistical analyses of a collection of studies where at least some studies provide multiple effect-size estimates that may or may not represent multiple constructs. Multiple endpoint studies typically involve research designs where individuals in one treatment group and one control group produce measures on multiple variables. These types of studies will likely lead to statistically dependent effect sizes. The dependence that arises from multiple endpoint studies in the meta-analytic framework has not been thoroughly studied. The main purpose of this thesis was to investigate the impact of dependence from multiple endpoint studies utilizing homogeneity measures commonly found in current meta-analyses.

This thesis is comprised of two sections: simulation and model estimation. Both sections replicated 3,000 meta-analyses for varying conditions. The simulation section varied study sample size, number of studies, between-outcomes correlation, and dependency structure. The model estimation section used generalized least squares estimation to analyze many of the same conditions. The standardized mean difference was the utilized effect-size estimator and Type I error rates of Q statistics were the primary unit of analysis.

Results showed that increased dependence among effect sizes is associated with increased Type I error rates of Q statistics. More specifically, under very dependent conditions, Type I error rates were significantly greater than their nominal levels regardless of within-study sample size and number of studies, sometimes with more than a twofold inflation. The model estimation section demonstrated that using generalized least squares estimation to account for multiple endpoint dependency maintains Type I error rates within nominal levels and is preferable to incorrectly assuming independence.

CHAPTER ONE

INTRODUCTION

> "Given a large mass of data, we can by judicious selection construct perfectly plausible unassailable theories—all of which, some of which, or none of which may be right."
>
> --Paul Arnold Srere (1925-1999), American biochemist

Meta-analysis is the statistical analysis of a large collection of results from individual studies for the purpose of integrating findings (Glass, 1976). Given the vast quantity of literature in fields such as medicine and the social-sciences, meta-analysis provides a method of organization and synthesis of research on related topics. This process is often completed using effect sizes, or quantitative representations of the magnitude of association between variables of interest found in studies (Hedges, 2007). This thesis utilized the standardized mean difference as the focal effect size. Formulas for computing asymptotic variances and covariances of effects, as well as related statistical quantities, are found in Gleser and Olkin (2009). A discussion of the standardized-mean-difference effect size as it relates to this thesis is provided in chapter two.

There are two fundamental types of meta-analysis: univariate meta-analysis and multivariate meta-analysis. Univariate meta-analysis refers to the statistical analysis of a collection of studies which provide distinct effect sizes that measure a single variable. This approach is often not realistic in the social-science and medical fields due to the complexity of research questions and study designs. Typically, multiple variables are of interest to the researcher. Rather, multivariate meta-analysis refers to the statistical analysis of a collection of studies where at least some studies provide multiple effect sizes that may or may not represent multiple constructs. Examples of existing multivariate meta-analyses are Becker (1990) and Caird, Willness, Steel, and Scialfa (2008).

Multivariate meta-analyses generally combine results from primary studies that utilize a one or more of three research designs. This thesis focused on the research design that involves one treatment group and one control group, where measures of multiple variables are supplied for each individual. Gleser and Olkin (2009) refer to studies with this type of design as *multiple*

endpoint studies. The two remaining types of designs, *multiple treatment groups* and *multiple time points*, are briefly discussed in chapter two.

Multivariate data in meta-analysis will likely lead to dependence among effect sizes. The degree and type of dependence depends on many factors, including but not limited to, the type of effect size (e.g., mean difference, correlation, odds ratio) and the type of multivariate data structure. For example, raw data from multiple treatment groups are usually independent. However, effect sizes from these studies are typically statistically dependent if treatments are compared to a common control group, because of a common standard deviation term (Kim & Becker, 2010). Alternatively, the dependence related to multiple endpoint studies arises from correlated raw data. The dependence between effect sizes from multiple treatment groups is discussed extensively by Kim and Becker (2010), whereas the dependence with respect to multiple endpoint effect sizes has not been thoroughly analyzed. The primary purpose of this thesis was to partially fill this gap in the meta-analytic literature by simulating dependent data under varying conditions, followed by an analysis of select homogeneity measures.

To investigate the impact of dependence from multiple endpoint studies this thesis utilized select homogeneity statistics and indices. While other indices and approaches are available to study this type of dependence, the rationale for restricting analyses to homogeneity measures involves the perceived applicability of the results. Mainly, homogeneity statistics and indices are fairly simple to calculate and are typically used as the primary means of choosing a meta-analytic model (fixed effects or random effects). For that reason, I presume that this thesis' findings will be useful for meta-analysts in a variety of fields.

CHAPTER TWO

LITERATURE REVIEW

This chapter begins with a description of multivariate data and dependence in meta-analysis. This is followed by a brief discussion of common statistical techniques for dealing with dependence in meta-analysis. Last, a review of the literature on select homogeneity statistics and indices in meta-analysis is provided.

2.1 Origins of Dependence in Meta-Analysis

Current social-science and medical researchers implement a wide variety of experimental designs, many of which can produce dependent effect sizes when combined with meta-analytic techniques. This diversity in experimental design produces several types of multivariate data which lead to dependence in meta-analysis. This chapter begins with an overview of three common experimental designs that lead to dependence in meta-analysis. Studies included in meta-analyses are often more complicated and intricate in design methodology than are described in this chapter. An example study is briefly discussed and referenced for each design.

2.1.1 Multiple Treatment Groups

Some experimental designs test multiple treatment groups simultaneously with a single control group. Although outcomes for these groups are statistically independent, the common control group that appears in the effect-size calculations will produce correlated effect sizes (Gleser & Olkin, 2009). "The more groups and contrasts involved, the more complex is the dependence in the multivariate data" (Becker, 2000, p. 502). For example, Hartung et al. (2002) studied mitoxantrone treatment options for patients with secondary progressive multiple sclerosis by administering a placebo to one group of patients (control group), 5 mg/m^2 of mitoxantrone to a second group of patients (first treatment group), and 12 mg/m^2 of mitoxantrone to a third group of patients (second treatment group). If this study were to be included in a hypothetical meta-analysis analyzing treatments for secondary multiple sclerosis, the presence of multiple treatment groups would produce two dependent effect sizes.

While it is theoretically possible for an experimental design to consist of multiple control groups which are compared to a single treatment group, this design is less predominant in the

social-science and medical literatures. Nevertheless, this parallel design would also produce dependence if several effect sizes from such a study were included in a meta-analysis.

2.1.2 Multiple Endpoints

Other experimental designs may involve a single treatment group and a single control group, and measure multiple outcomes, or endpoints, for each individual. Assuming measures on each participant are likely to be correlated, corresponding effect-size estimates for these measures will be correlated within studies (Gleser & Olkin, 2009). As an example, Huntley, Rasmussen, Villarubi, Sangtong, and Fey (2000) compared the mathematics achievement of students given the Core-Plus Mathematics Project curriculum to those given a conventional curriculum. This was accomplished by administering multiple tests of algebraic knowledge. If this study were to be included in a hypothetical meta-analysis on experimental mathematics curricula, the presence of multiple tests that produce multiple outcomes for each individual would produce dependent effect sizes. This type of multivariate data structure was the focus of this thesis.

More complex experimental designs incorporate both multiple treatment groups and multiple endpoints. For example, Barrett-Connor et al. (2002) looked at secondary data concerning the effectiveness of a specific osteoporosis medication, Raloxifene. Patients were initially given either a placebo, 60 mg/d (milligrams per day) of raloxifene (first treatment group), or 120 mg/d of raloxifene (second treatment group). The researchers then measured a series of cardiovascular and cerebrovascular events (endpoints), therefore creating a different type of multivariate data structure that merges multiple endpoints and multiple treatment groups. The use of this research design and the impact of its dependence in meta-analysis have not been studied extensively.

2.1.3 Multiple Time Points

A third type of multivariate data structure in meta-analysis occurs when individuals are measured on the same instrument during multiple time points, such that an effect size can be computed for each unique time point. This type of experimental design is especially common in research where there is a high demand for longitudinal studies. For example, Chae et al. (2005, p. 832) looked at the effectiveness of intramuscular electrical stimulation in reducing hemiplegic shoulder pain. Patients in each treatment group were administered a type of stimulation

(electrical or cuff-type sling) six hours a day, for six weeks. Self reports of existing pain were routinely administered at the beginning and end of treatments, as well as at 3, 6, and 12 months beyond the conclusion of treatment. If this study were to be included in hypothetical meta-analysis on shoulder pain treatments, the presence of multiple time points (self reports of pain) would produce dependent effect sizes.

2.2 Accounting for Dependence in Meta-Analysis

As discussed in the previous section, the presence of multivariate data in meta-analysis is quite common. Meta-analysts have four general options when faced with such types of data. These methods are not equivalent in accuracy or statistical quality; Becker (2000) and Hedges (2007) are sources for much of the following information.

The first and most simplistic option for the analyst is to ignore dependence entirely. In other words, the researcher assumes data independence. While intuitively this may seem surely detrimental to the meta-analysis, under certain circumstances the adverse effects may be minimal. Hedges (2007) states "it may not be too misleading if relatively few studies report more than one effect size" (p. 924). Put another way, if among k studies in a meta-analysis with p outcomes of interest, only m of k studies report multiple outcomes, where m is considerably lower than k, assuming independence might not be extensively damaging to analyses. As an example, Becker (1989) re-analyzed gender differences in science achievement. From a total of 30 effect sizes only two arose from common studies. In this instance it is expected to be permissible to proceed under the independence assumption due to the relatively low number of studies that contributed multiple effect sizes. As the number of studies that report multiple outcomes becomes larger, assuming independence becomes increasingly ill-advised. As an example of problematic results from ignoring dependence entirely, Hedges (2007) states that continuing under the independence assumption with positively-correlated dependent data leads to conservative estimates of standard errors for overall mean effect sizes. Researchers who are unsure about the severity or impact of dependent data can perform sensitivity analysis to assess these concerns. Greenhouse and Iyengar (2009) provide procedures for such analyses.

The second approach consists of partitioning dependent data into separate data sets, typically by outcome or dependent variable, followed by independent analyses. Becker (2000) explains that this approach is often deceptive because comparisons of results across data sets

(e.g., mean effect sizes) will still violate the assumption of independence due to the remaining presence of dependence. However, if these types of comparisons are not desired by the researcher, this approach is valid and functional. Also, the number of effect sizes in analyses can be drastically reduced depending on stratification criteria, which may lead to other issues.

The third approach is to mathematically condense or reduce the number of effect sizes in a particular study so that any given study does not introduce dependence into the meta-analysis. The most common application of this method uses simple summative methods, such as the arithmetic mean or median of all effect sizes for a given variable or construct, as an estimate. A similar option is to select the *best* effect-size estimate that the researcher deems most representative of the overall data. More intricate methods have been created to reduce dependent data in meta-analysis (e.g., Hedges & Olkin, 1985; Rosenthal & Rubin, 1986). Last, Marín-Martínez and Sánchez-Meca (1999) discuss some advantages and disadvantages of select data reduction methods. While several approaches are reasonable, all methods lead to data loss and often require revised formulas for common computations (e.g., standard error of fixed-effects weighted mean effect size), as discussed in Hedges (2007).

The fourth approach to dealing with dependence in meta-analysis is to statistically model the dependence. This is the most theoretically justifiable technique of dealing with dependence. Despite the existence of several methods, this continues to be a topic of interest for meta-analysis researchers. Hedges and Olkin (1985) present a multivariate approach which requires all studies in the meta-analysis to supply data for all outcomes under investigation. Raudenbush, Becker, and Kalaian (1988) extend the work of Hedges and Olkin (1985) by presenting a generalized least squares method that allows for variation in outcomes and correlational structures among studies (i.e., not all studies contribute results for all outcomes). Timm (1999) uses Finite Intersection Test procedures to test the equivalence of effect sizes and estimate an overall effect size. Nam, Mengersen, and Garthwaite (2003) propose several multivariate Bayesian models for meta-analysis. All of the above procedures require knowledge or estimates of correlation matrices among dependent effect sizes.

This requirement can be particularly difficult to satisfy if studies fail to provide adequate correlational information. Secondary resources, such as studies with similar measurement instruments or test manuals, can be helpful in estimating approximate values of within-study correlations. The impact of estimating or ignoring within-study correlation has not been

adequately studied, although Ishak, Platt, Joseph, and Hanley (2008) found treatment effect and heterogeneity estimates (i.e., effect-size variances under random-effects model) to not be strongly affected by imprecise estimations of study covariance matrices. Last, more recent multivariate modeling methods do not require knowledge of the correlational structure for dependent effect sizes, provided certain circumstances are satisfied (e.g., Hedges, Tipton, & Johnson, 2010).

2.3 Standardized-Mean-Difference Effect Size

In order to compare quantities and make valid interpretations of statistics in meta-analysis, study results must be reported on identical scales. Multiple types of effect-size measures are currently available in the literature; Huberty (2002) presents a brief history and description of some available indices. Several of these indices are raw (presented in the original score metric of the primary study) while others are standardized. However, raw effect sizes are only an option to researchers if all the studies in the meta-analysis use the same scale (Bond, Wiitala, & Richard, 2003; Borenstein, 2009). Standardized effect sizes provide a solution to this issue by transforming effect sizes into a common metric. Becker (2000) states that although two studies might use different scales, if the construct underlying those scales represented the 'same' view of a construct, effect sizes could be compared across studies. This thesis only considered the standardized-mean-difference effect size.

Though Glass (1976) is often considered a seminal article in the meta-analytic literature, it is hardly the origin of the mean-difference effect size (e.g., Cohen, 1962). The aim of the mean-difference effect size is to quantify the magnitude and direction of a group difference on a specific variable. The population standardized-mean-difference effect size can be written as

$$\delta_{ij} = \frac{\mu_{ij}^{\mathrm{T}} - \mu_{ij}^{\mathrm{C}}}{\sigma_{ij}}, i = 1, \dots, k; j = 1, \dots, p\,, \tag{1}$$

where δ_{ij} is the population standardized-mean-difference effect-size parameter for the j^{th} outcome measure in the i^{th} study, given population treatment and control means μ_{ij}^{T} and μ_{ij}^{C}, respectively, and a population within-groups standard deviation, σ_{ij}. Glass (1976) proposes a sample estimator, g, for the population standardized-mean-difference effect size,

$$g_{ij} = \frac{\bar{Y}_{ij}^{\mathrm{T}} - \bar{Y}_{ij}^{\mathrm{C}}}{S_{ij}}, \tag{2}$$

where g_{ij} is the sample standardized-mean-difference effect size for the j^{th} outcome measure in the i^{th} study, $\bar{Y}_{ij}^{\mathrm{T}}$ and $\bar{Y}_{ij}^{\mathrm{C}}$ are the sample treatment and control means for the j^{th} outcome of the i^{th} study that estimate μ_{ij}^{T} and μ_{ij}^{C}, respectively, and S_{ij} is a sample standard deviation that estimates σ_{ij}. Glass (1976) calculates S_{ij} as the standard deviation of the control group but a more widely implemented approach calculates S_{ij} as a pooled standard deviation.

Hedges (1981) showed Glass' *g* to be biased, especially given studies consisting of small sample sizes. More specifically,

$$\mathrm{E}(g_{ij}) = \left(\frac{\Gamma\left(\frac{m_{ij}}{2}\right)}{\sqrt{\frac{m_{ij}}{2}}\,\Gamma\left(\frac{m_{ij}-1}{2}\right)} \right)^{-1} \delta_{ij}, \tag{3}$$

where $m_{ij} = n_i^{\mathrm{T}} + n_i^{\mathrm{C}} - 2$, n_i^{T} and n_i^{C} are sample sizes of the treatment and control groups for the i^{th} study, respectively, $\mathrm{E}(g_{ij})$ is the expected value of g_{ij}, and $\Gamma\left(\frac{m_{ij}}{2}\right)$ is the gamma function with parameter $\frac{m_{ij}}{2}$. Furthermore, Hedges (1981) provides an approximation of the standardized-mean-difference effect size, corrected for small sample bias:

$$d_{ij} = g_{ij}\left(1 - \frac{3}{4m_{ij} - 1}\right), \tag{4}$$

where d_{ij} is the unbiased sample standardized-mean-difference effect size.

The variance of an effect size is a useful measure of dispersion and is required for many summative statistics (e.g., mean effect sizes) in meta-analysis. The exact variance of the mean-difference effect size is complex but is typically well-approximated by

$$v_{ij} = \frac{n_i^{\mathrm{T}} + n_i^{\mathrm{C}}}{n_i^{\mathrm{T}} n_i^{\mathrm{C}}} + \frac{d_{ij}^2}{2(n_i^{\mathrm{T}} + n_i^{\mathrm{C}})}, \tag{5}$$

where v_{ij} is the sample standardized-mean-difference effect-size variance for the j^{th} outcome of the i^{th} study (Hedges, 1981).

Last, the choice not to include the j^{th} subscript for n_i^{T} and n_i^{C} terms assumes that all individuals in treatment and control groups in the i^{th} study provide scores for all outcomes, hence there is no reason to denote specific outcomes when considering group membership since $n_{i1}^{\text{T}} = n_{i2}^{\text{T}} = \cdots = n_{ip}^{\text{T}}$, and similarly for n_i^{C}. While this assumption is fairly realistic, unexpected situations may arise that can result in the rejection of this assumption. Two examples might include mortality of patients in medical studies and student attrition in educational studies.

2.4 Homogeneity Statistics and Indices

Statistical heterogeneity occurs when true effects differ among studies (Higgins & Thompson, 2002). Homogeneity analysis is a set of procedures which aim to determine if a group of effect sizes are derived from a common population that can be represented by a distinct population effect size. In a homogeneous distribution, the dispersion of the effect sizes around their mean will be no greater than what is expected from sampling error alone (Lipsey & Wilson, 2001). If the aforementioned definition does not hold, the data are said to be heterogeneous. The introduction of systematic or random error components is required to model the data. Homogeneity analysis is a typical initial step in many meta-analyses, mainly due to its influence on calculations and interpretations of numerous values and statistics that follow.

The choice of an overall model (fixed-effects or random-effects) can be based on the decision regarding homogeneity in the observed data. This thesis examined three homogeneity measures: Q statistic, Birge's ratio, and I^2. Analyses focused on the Q statistic since it is the more commonly reported homogeneity measure in meta-analyses. Also, Birge's ratio and I^2 are merely functions of the Q statistic. This is not an exhaustive list of available homogeneity statistics in the meta-analytic literature. Some visual methods of homogeneity analysis include the use of normal probability plots, as proposed by Hardy and Thompson (1998), and the use of Galbraith plots, as shown in Galbraith (1988). Another index to assess homogeneity is R, which is "the ratio of the standard error of the underlying mean from a random-effects meta-analysis to the standard error of a fixed-effects meta-analytic estimate" (Higgins & Thompson, 2002, p. 1539). This particular index is rarely used in the meta-analytic literature.

2.4.1 *Q* Statistic

Although the *Q* statistic was published prior to the formulation of meta-analysis (see Cochran, 1954), one of the first uses of *Q* for the purpose of meta-analysis was in Hedges (1982). This thesis utilizes two forms of the *Q* statistic: univariate and multivariate. While the underlying interpretation is essentially the same between the two forms, the computation of the statistic does change. The univariate form was utilized in the simulation section while the multivariate form was utilized in the model estimation section. I introduce both *Q* statistic forms in this section and leave further explanation (particularly for the multivariate form) for later chapters.

The sample univariate *Q* statistic is

$$Q = \sum_{i=1}^{k} \left[\frac{\left(d_i - \bar{d}\right)^2}{v_i} \right] \tag{6}$$

where d_i is the i^{th} sample standardized-mean-difference effect size, v_i is the fixed-effects variance of the i^{th} effect size, $\bar{d}$ is the fixed-effects weighted mean effect size, and *k* is the number of independent studies, as shown in Hedges and Olkin (1985). The above equation tests the homogeneity of effect sizes across all studies for a single outcome.

The univariate *Q* statistic follows an asymptotic chi-square distribution with $k - 1$ degrees of freedom under the fixed-effects model. Under the random-effects model, *Q* follows a non-central chi-square distribution which has a different degrees of freedom (Biggerstaff & Tweedie, 1997; Higgins & Thompson, 2002). Although this thesis incorporated two outcomes, separate outcome data sets were analyzed in the simulation section which permitted the use of Equation 6. The calculation of *Q* by way of Equation 6 is only valid for the simulation section. The process of converting multivariate data to univariate data is discussed in chapter three.

The null hypothesis of the statistical test that is paired with the univariate *Q* statistic states that all observed effects arise from a single population effect size. In notational form, the univariate *Q* statistic tests the null hypothesis:

$$\mathrm{H}_0\colon \delta_1 = \delta_2 = \cdots = \delta_i = \delta\ .$$

The rejection of the null hypothesis suggests the presence of heterogeneity among effect sizes, at which point the adoption of a fixed-effects model with no predictors would be inappropriate. A random-effects model, which attempts to account for between-studies variation, is a typical option for meta-analysts with heterogeneous data. When the null hypothesis of the *Q* test is rejected, the summative effect-size statistic, $\bar{d}$, is re-calculated to account for heterogeneity among effect sizes and is interpreted as a 'mean effect size' rather than a 'common effect size,' as is typically interpreted under a fixed-effects model.

The multivariate *Q* statistic is

$$Q_{\mathrm{M}} = \left(\mathbf{d} - \mathbf{X}\hat{\boldsymbol{\delta}}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{d} - \mathbf{X}\hat{\boldsymbol{\delta}}\right), \tag{7}$$

where all terms are extensively defined later in section 3.3.1. The multivariate *Q* statistic follows an asymptotic chi-square distribution with $\dim \mathbf{d}(p) - \dim \boldsymbol{\delta}(p)$ degrees of freedom, where $\mathbf{d}(p)$ and $\boldsymbol{\delta}(p)$ are also defined in section 3.3.1. The null hypothesis of the statistical test that is paired with the multivariate *Q* statistic states that all observed effects across all outcomes arise from a single population effect-size estimate. In notational form, the multivariate *Q* statistic tests the null hypothesis:

$$\mathrm{H}_0\colon\ \delta_{11} = \cdots = \delta_{ij} = \cdots = \delta\ \forall\ i, j : i = 1, \cdots, k;\ j = 1, \cdots, p.$$

Both forms of the *Q* statistic test for homogeneity across effect sizes. The main difference is that the multivariate form tests for homogeneity across all outcomes in addition to all effect sizes.

There has been some speculation with respect to efficiency, power, and other statistical qualities of the *Q* statistic. Hardy and Thompson (1998) investigated the power of *Q* under multiple conditions. Their simulation showed that *Q* has the most power when the total information among studies, which they define as the summation of inverse variance weights $\left(I = \sum_{i=1}^{k} 1/v_i\right)$, is maximized. Parallel to that finding, the power of *Q* is greatly reduced when a small number of studies comprise large proportions of the total information. In other words, the vast amount of precision that is introduced by so-called "super-studies", which possess huge sample sizes and thus have large weight, is not as crucial as an even allocation of weights when attempting to achieve high power for the *Q* statistic. Harwell (1997) simulated *Q* under a variety of conditions, varying number of studies, form of non-normally distributed data, and considering

unequal variances. His results showed that Q will typically have a nominal Type I error rate and a minimal Type II error rate under the conditions of independently normally distributed data with equal variances, at least 40 observations per study, and a moderate within-study sample size to number of studies ratio. When the number of observations per study drops below 40, Type I error rates will often be conservative. Small ratios of within-study sample size to number of studies, N_i/k, intensify this issue. This occurs when a meta-analysis has a large number of studies, the majority of which contain exceptionally small sample sizes.

The statistical properties of Q have been well-studied under a variety of conditions, all of which assume independence. This thesis analyzed the behavior of the Q statistic under dependent conditions. Since Q only describes the presence of heterogeneity (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006), other indices are available to quantify the extent and impact of heterogeneity among effect sizes.

2.4.2 Birge's Ratio

Another homogeneity measure used in this thesis was Birge's ratio, originally proposed by Raymond Birge for comparing observed error to probable error in the pure sciences (Birge, 1932). When applied to meta-analysis, Birge's ratio is

$$R_B = \frac{Q}{df_Q}, \tag{8}$$

where R_B is the Birge's ratio and df_Q is the degrees of freedom associated with the Q statistic. Birge's ratio is an estimate of the between-studies variation versus the within-studies sampling error (Konstantopoulos & Hedges, 2009). Values approximately equal to one signify that the between-studies and within-studies variances are roughly equal. As R_B increases, the presence of variation beyond within-study sampling error increases. Higgins and Thompson (2002) propose a symmetric Wald-type uncertainty interval for R_B:

$$R_B \pm Z\left(\frac{\alpha}{2}\right)\frac{\sqrt{v_Q}}{df_Q}. \tag{9}$$

Written in terms of the univariate Q statistic, Equation 9 becomes

$$\frac{Q}{k-1} \pm Z\left(\frac{\alpha}{2}\right)\frac{\sqrt{2(k-1)}}{k-1}, \tag{10}$$

where $Z\left(\frac{\alpha}{2}\right)$ is the point beyond which the standard normal distribution has probability equal to $\frac{\alpha}{2}$ and v_Q is the variance of the Q statistic. Despite the existence of an uncertainly interval, Equation 9 is not commonly used in published meta-analyses. This thesis only acknowledges the proposed uncertainly interval and does not make use of Equations 9 or 10.

Higgins and Thompson (2002) describe sample p values of the Q statistic: $p = .1$, $p = .05$, and $p = .01$ associated with $R_B = 1.28$, $R_B = 1.37$, and $R_B = 1.55$, respectively, given a sample of 10 studies. These values deviate a bit when the number of studies in a meta-analysis increases. These are rough comparisons proposed by Higgins and Thompson which arose from simulation rather than statistical theory.

2.4.3 I^2

The last method of homogeneity analysis used in this thesis was I^2, which describes "the percentage of total variation across studies that is due to heterogeneity rather than chance" (Higgins, Thompson, Deeks, & Altman, 2003, p. 558). The logic behind this index stipulates that the expected variation when effect sizes are homogenous (df_Q) is to be subtracted from the homogeneity statistic so that what remains is excessive variation due to heterogeneity. This quantity can be derived from Birge's ratio, such that

$$I^2 \quad = \frac{R_B - 1}{R_B} \rightarrow \frac{\frac{Q - df_Q}{df_Q}}{\frac{Q}{df_Q}} \rightarrow \frac{Q - df_Q}{Q}. \tag{11}$$

If Q is less than its degrees of freedom, I^2 is defined as zero:

$$I^2 = max\left\{0, \left(\frac{Q - df_Q}{Q}\right)100\%\right\}. \tag{12}$$

A set of non-rigorous interpretations of I^2 are *no variation*, *low*, *moderate*, and *high variation* for values 0, 25, 50, and 75, respectively (Higgins et al., 2003). Due to the ill-defined interpretation and similar sampling distribution to that of Q, Shadish and Haddock (2009)

suggest that I^2 only be used in conjunction with Q and not as an independent homogeneity measure. Some researchers (e.g., Montori, Leung, Walter, & Guyatt, 2005) choose to calculate I^2 as a proportion rather than a percentage. This linear transformation does not change the underlying meaning of the index.

Equations 8 and 11 show that both Birge's ratio and I^2 are functions of the Q statistic. Therefore, I treated the Q statistic as the focus of statistical investigation; results involving Birge's ratio and I^2 served as secondary focal points of the thesis. The next chapter discusses methods used in the simulation and model estimation sections.

CHAPTER THREE

METHODS

To analyze the impact of multiple endpoint dependency I focused on homogeneity measures, as discussed in chapter two. This process was completed in two sections: simulation and model estimation. Both sections involved a large number of independently generated meta-analyses under varying conditions. The model estimation section utilized generalized least squares estimation, which accounts for multiple endpoint dependency through a sample variance-covariance matrix. The simulation section used dependency structures to create and manipulate dependence during the iteration process; this is discussed thoroughly in section 3.2. Both sections analyzed the impact of dependence on homogeneity measures. All analytical portions of this thesis were conducted in SAS 9.2 and graphics were created in SPSS 19 and R.

3.1 Data Generation

A 4 x 4 correlation matrix, **r**, was constructed directly in the SAS environment:

$$\mathbf{r} = \begin{bmatrix} 1 & \rho_{12}^{T} & 0 & 0 \\ \rho_{12}^{T} & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_{12}^{C} \\ 0 & 0 & \rho_{12}^{C} & 1 \end{bmatrix}, \tag{13}$$

where ρ_{12}^{T} and ρ_{12}^{C} are population between-outcomes correlation coefficients for outcome 1 and outcome 2 for treatment and control groups, respectively. The condition $\rho_{12}^{T} = \rho_{12}^{C}$ was imposed for all conditions in both the simulation and model estimation sections. In other words, since there were multiple values ρ_{12}^{T} and ρ_{12}^{C}could take on (see Tables 3.1 and 3.2), a restriction was created so that $\rho_{12}^{T}$and ρ_{12}^{C} were equivalent across all other conditions. More intuitively written, $\rho_{12}^{T} = \rho_{12}^{C} = \rho_{12} = \rho$, where ρ can also be interpreted as the correlation between outcome 1 and outcome 2. The composition of **r** is such that values within treatment and control groups for outcomes were correlated to a known degree ρ, while values for all other grouping combinations (e.g., treatment group for outcome 1 and control group for outcome 1) had no structured correlation (approximate independence). Four data points were randomly sampled from the standard normal distribution and placed into a 4 x 1 vector, **u**. A Cholesky decomposition was

performed on **r**, the transpose of which was multiplied by **u** to create a 4 x 1 column vector, **y**. In notational form,

$$\mathbf{y}_{4x1} = \text{chol}(\mathbf{r})'_{4x4} * \mathbf{u}_{4x1}, \tag{14}$$

where $\text{chol}(\mathbf{r})'$ is the transpose of the Cholesky decomposition of the matrix **r**. See Appendix A for more information regarding the Cholesky decomposition.

This procedure was completed *n* times, where *n* represents the number of observations in a single study in a meta-analysis. This thesis assumed that *n* did not vary among multiple outcomes. The transpose of each iteration, $\mathbf{y}'$, were stacked vertically, the end result of which was an *n* x 4 matrix, **Y**. The elements of each **Y** matrix represent the outcomes for one study in a meta-analysis, as shown in Figure 3.1.

Y_{i11}^{T}	Y_{i21}^{T}	Y_{i11}^{C}	Y_{i21}^{C}
Y_{i12}^{T}	Y_{i22}^{T}	Y_{i12}^{C}	Y_{i22}^{C}
⋮	⋮	⋮	⋮
Y_{i1n}^{T}	Y_{i2n}^{T}	Y_{i1n}^{C}	Y_{i2n}^{C}

Figure 3.1: Raw-Data Matrix.

Note that Y_{ijm}^{T} represents the m^{th} observation of the j^{th} outcome of the i^{th} study, where $m = 1, \ldots, n, j = 1, 2,$ and $i = 1, \ldots, k$ for the treatment group and Y_{ijm}^{C} is defined similarly for the control group. The goal of the data generation process was to create correlated raw data within experimental scenarios (treatment or control). With respect to Figure 3.1, columns 1 and 2 were correlated and column 3 and 4 were correlated, both to the same approximate degree. All other combinations of the four columns possessed very low correlations, which only arose by chance. This process was completed *k* times to simulate *k* studies in a single meta-analysis. That is, each replicated meta-analysis had *k* raw-data vectors **Y**.

3.2 Simulation Method

Simple arithmetic means and standard deviations were calculated for each column of **Y**. Next, pooled standard deviations were calculated using

$$S_{ij_{\text{pooled}}} = \sqrt{\frac{(n_i^{\text{T}} - 1)(S_i^{\text{T}})^2 + \left(n_i^{\text{C}} - 1\right)\left(S_{ij}^{\text{C}}\right)^2}{n_i^{\text{T}} + n_i^{\text{C}} - 2}} \overset{n_i^{\text{T}} = n_i^{\text{C}}}{\Longrightarrow} \sqrt{\frac{\left(S_{ij}^{\text{T}}\right)^2 + \left(S_{ij}^{\text{C}}\right)^2}{2}}, \tag{15}$$

where $S_{ij_{\text{pooled}}}$ is the sample pooled standard deviation for the j^{th} outcome of the i^{th} study and S_{ij}^{T} and S_{ij}^{C} are sample treatment and control group variances, respectively. Next, unbiased standardized-mean-difference effect sizes were calculated using

$$d_{ij} = \left(1 - \frac{3}{4(n_i^{\text{T}} + n_i^{\text{C}}) - 1}\right)\frac{\bar{Y}_{ij}^{\text{T}} - \bar{Y}_{ij}^{\text{C}}}{S_{ij_{\text{pooled}}}} \overset{n_i^{\text{T}} = n_i^{\text{C}}}{\Longrightarrow} \left(1 - \frac{3}{8n_i - 9}\right)\frac{\bar{Y}_{ij}^{\text{T}} - \bar{Y}_{ij}^{\text{C}}}{S_{ij_{\text{pooled}}}}, \tag{16}$$

where d_{ij} is the unbiased standardized-mean-difference effect-size estimate for the j^{th} outcome of the i^{th} study, $\bar{Y}_{ij}^{\text{T}}$ and $\bar{Y}_{ij}^{\text{C}}$ are sample treatment and control group means for the j^{th} outcome of the i^{th} study, respectively, and n_i is the generic sample size of the treatment and control groups in the i^{th} study. These effect sizes were arrayed in a k x 2 matrix. Each effect-size estimate had a sample variance equal to

$$v_{ij} = \frac{n_i^{\text{T}} + n_i^{\text{C}}}{n_i^{\text{T}} n_i^{\text{C}}} - \frac{d_{ij}^2}{2(n_i^{\text{T}} + n_i^{\text{C}})} \overset{n_i^{\text{T}} = n_i^{\text{C}}}{\Longrightarrow} \frac{2}{n_i} - \frac{d_{ij}^2}{4n_i}. \tag{17}$$

Equations 15 though 17 provide equations when treatment and control groups are equivalent in size. This thesis incorporated this restraint throughout all conditions.

Variances from Equation 17 were arrayed in a matrix with the same dimensionality of the effect-size matrix previously mentioned. Weights of effect-size estimates, W_{ij}, were calculated as inverses of respective effect-size variances. In notational form:

$$W_{ij} = \frac{1}{v_{ij}}, \tag{18}$$

where W_{ij} is the weight associated with the effect-size estimate for the j^{th} outcome of the i^{th} study. The aim of the abovementioned calculations was to obtain Q statistics for each outcome in a meta-analysis. Therefore, the fixed-effects weighted mean effect size is required:

$$\bar{d}_j = \frac{\sum_{i=1}^{k} W_{ij} d_{ij}}{\sum_{i=1}^{k} W_{ij}}, \tag{19}$$

where $\bar{d}_j$ is the fixed-effects weighted mean effect size for the j^{th} outcome. Note that the upper bound of the summation is *k*, which implies that all studies in the meta-analysis were assumed to have provided effect-size estimates for both outcomes. This does not necessarily have to be the case in practice. The equation for the univariate *Q* statistic for the j^{th} outcome was calculated as:

$$Q_j = \sum_{i=1}^{k} W_{ij}\left(d_{ij} - \bar{d}_j\right)^2. \tag{20}$$

Q statistics were calculated for both outcomes in a single meta-analysis of *k* studies. Using results from Equation 20, Birge's ratio and I^2 indices were calculated as follows (respectively):

$$R_B = \frac{Q}{k-1} \tag{21}$$

and

$$I^2 = \frac{Q - k + 1}{Q}. \tag{22}$$

This process was replicated 3,000 times for all simulation conditions (see Table 3.1).

Table 3.1: Simulation Conditions.

Dependency Structure	Between–Outcomes Correlation (ρ)	Number of Studies (*k*)	Sample Size (*n*)
Independent	.50	24	20
Moderately Dependent	.75	48	100
Very Dependent	.99	96	180
			260

Four conditions varied throughout the simulation process. First, individual study sample sizes, *n*, took on four values that represented sizes ranging from small to large: 20, 100, 180, and 260. Next, the number of studies within a meta-analysis, *k*, took-on three values: 24, 48, and 96.

The between-outcomes correlation, ρ, also took on three values: .50, .75, and .99. The process of obtaining the between-outcomes correlation was discussed in section 3.1. The last condition in the simulation section was the dependency structure, which is explained below. There were a total of 108 conditions in the simulation section.

The dependency structure condition was the main focus of the simulation section. I sought to investigate homogeneity measures under conditions with dependent data; therefore I created data with three different dependency structures: *Independent*, *Moderately Dependent*, and *Very Dependent*. Figures 3.2a – 3.2c show visual representations of these structures and are explained below. First I consider Figure 3.2a, which is labeled the "Independent" structure.

$d_{1,1}$	$d_{1,2}$
⋮	⋮
⋮	$d_{b,2}$
⋮	$d_{b+1,2}$
⋮	⋮
$d_{c,1}$	$d_{c,2}$
$d_{c+1,1}$	$d_{c+1,2}$
⋮	⋮
$d_{a,1}$	⋮
$d_{a+1,1}$	⋮
⋮	⋮
$d_{k,1}$	$d_{k,2}$

Figure 3.2a: Independent Structure. $a = \frac{3k}{4}, b = \frac{k}{4}, c = \left(\frac{k}{2}\right)$.

In Figure 3.2a, effect sizes *within* columns are independent, whereas pairs of effect sizes *across* columns are correlated with a population correlation ρ. This implies that the correlation between $d_{1,j}$ and $d_{2,j}$ is zero, whereas the correlation between $d_{i,1}$ and $d_{i,2}$ is approximately equal to ρ. Effect sizes from multiple outcomes within a study are expected to be correlated whereas effect sizes for the same outcome across different studies are expected to be uncorrelated. Since effect sizes in univariate meta-analysis are typically analyzed within outcomes (within columns of effect sizes in this example), Figure 3.2a could produce two independent sets of data analyses. Figure 3.2a was considered the baseline structure, and the

remaining two structures were derived directly from the independent result. Next, I consider Figure 3.2b, which has been labeled as the "Very Dependent" structure.

$d_{1,1}$	$d_{c+1,1}$
⋮	⋮
⋮	⋮
⋮	⋮
⋮	⋮
$d_{c,1}$	$d_{k,1}$
$d_{1,2}$	$d_{c+1,2}$
⋮	⋮
⋮	⋮
⋮	⋮
⋮	⋮
$d_{c,2}$	$d_{k,2}$

Figure 3.2b: Very Dependent Structure. $c = \left(\frac{k}{2}\right)$.

Here I created a data set where $k/2$ pairs of effects $\left(d_{i,1}, d_{i,2}\right)$ are analyzed together. Operationally, the upper one-half of effect sizes in column 2 were switched with the lower one-half of effect sizes in column 1. Then the effects in column 1 were analyzed, this resulted matrix has the largest possible proportion of correlated effect sizes within each column. Contextually, this would be equivalent to univariately analyzing a group of effect sizes where half of the effect sizes are correlated with the other half. Figure 3.2b displays an increased proportion of shaded and non-shaded pairs of effect sizes within columns compared to Figure 3.2a. Last, I consider Figure 3.2c, which has been labeled as the "Moderately Dependent" structure.

$d_{1,1}$	$d_{a+1,1}$
⋮	⋮
⋮	$d_{k,1}$
⋮	$d_{b+1,2}$
⋮	⋮
⋮	⋮
⋮	⋮
⋮	⋮
$d_{a,1}$	⋮
$d_{1,2}$	⋮
⋮	⋮
$d_{b,2}$	$d_{k,2}$

Figure 3.2c: Moderately Dependent Structure. $a = \frac{3k}{4}, b = \frac{k}{4}$.

Recalling that effect sizes are typically analyzed for each particular outcome, this third effect-size matrix was created by moving one-fourth of the effect sizes from column 2 into column 1, creating a somewhat elevated level of dependence in column 1. Essentially, half of the effect sizes in each column are paired (thus dependent), while the other half are independent. Specifically, $(d_{1,1}, d_{1,2}), \ldots, (d_{b,1}, d_{b,2})$ are dependent values in the first column that will be treated as if they are independent for analysis. This is precisely what Figure 3.2c intends to represent. Though the number of studies varied, all possible values of *k* were multiples of four (see Table 3.1). This allowed for the first fourth of effect sizes in column 2 to be switched with the last fourth of the effect sizes in column 1 for all conditions using the moderately dependent structure. The label "Moderately Dependent" represents the level of expected dependence. Since half of the effect sizes within columns are correlated, the term "moderate" seemed applicable. Figure 3.2c displays a limited proportion of shaded and non-shaded pairs within columns, which is more than are present in Figure 3.2a but less than Figure 3.2b.

The independent structure was the baseline structure for all conditions in the simulation section. When a specific set of conditions involved either the moderately dependent or very dependent structure, the baseline structure was manipulated directly in SAS to resemble the desired dependency structure. A variance-covariance matrix was created to correspond with each reorganized data matrix, as described below. All remaining calculations and procedures in the

simulation section accounted for revised effect size and effect-size variance matrices. Even though the independent structure does not contribute any dependence into analyses, it is still considered one of three dependency structures for continuity of terminology.

Although the dependence in the data is artificially created, this kind of dependence can be compared to what happens in meta-analyses. The decision by a meta-analyst to convert multivariate data to univariate data prior to analysis essentially treats dependent data as if it were independent. As an example, Becker (1989) re-examined the literature on gender and science achievement. The 30 studies in produced 31 effect sizes. However, only 29 of the 31 effect sizes arose from independent samples. The author omitted each of the two dependent effect sizes in turn, but another option would have been to perform a univariate meta-analysis using all 31 effect sizes. If this were done, the structure of the effect sizes would have been similar to the first column of Figure 3.2c, with slightly less dependence present. A single column of effect sizes would have been analyzed, but two of the effect sizes would be correlated.

Although the dependency structure and between-outcomes conditions can be viewed as related in the sense that they both have dependent data, the between-outcomes correlation is only a required condition for the creation of the real focus of the simulation section, the dependency structure. The two conditions remain separate; the between-outcomes correlation refers to a relationship between effect-sizes prior to univariate analyses whereas the dependency structure refers to the relationship among effect-sizes during univariate analysis.

The simulation section replicated 3,000 meta-analyses, each providing two Q statistics (one for each outcome). Since the data used to calculate the two sets of Q statistics was sampled randomly and showed almost identical results, only Q statistics from the first outcome were analyzed. Below is the format of Q statistic vector:

$$\begin{bmatrix} Q_{1,1} \\ Q_{1,2} \\ \vdots \\ Q_{1,3000} \end{bmatrix}.$$

Using $Q \dot{\sim} \chi^2_{k-1}$ (Hedges, 1982), frequencies of Q statistics were calculated and compared to expected frequencies specified by $\chi^2_{k-1;\,.95}$, which were used to calculate empirical rejection rates across all conditions. Proportions of Q statistics greater than their appropriate critical value were calculated and compared to the 95% confidence interval for proportions to

assess statistical significance of Type I error rates. The standard error of the proportion, given significance level α = .05, for Q is

$$\text{SE} = \sqrt{\frac{\alpha(1-\alpha)}{N_{\text{meta}}}} = \sqrt{\frac{.05(1-.05)}{3000}} \approx .004\,, \tag{23}$$

where $N_{\text{meta}}$is the number of replicated meta-analyses. We expect the Type I error rates to fall in the interval $.05 \pm 1.96(.004) = [.042, .058]$ in approximately 95% of the instances.

3.3 Model Estimation Method

The univariate approach that was used to analyze multivariate data in section 3.2 ignored existing dependence. Therefore, a multivariate meta-analysis approach which accounts for dependence was considered. The generalized least squares (GLS) estimation method, as shown in Raudenbush, Becker, and Kalaian (1988) and Gleser and Olkin (2009), was used in contrast to the univariate approach from the simulation section. The GLS approach directly accounts for the dependence among effect sizes using an effect-size variance-covariance matrix. All conditions besides the dependency structures discussed in section 3.2 were considered in the model estimation portion. A new condition, raw-data correlation type, was used and is discussed later. I begin with an overview of GLS as it is typically applied to multivariate meta-analysis. This is followed by a description of procedures used to estimate homogeneity measures.

3.3.1 GLS for Meta-Analysis

Suppose a typical linear regression model is of the form

$$\mathbf{d} = \mathbf{X}\boldsymbol{\delta} + \boldsymbol{\varepsilon}\,, \tag{24}$$

where $\mathbf{d}$ is a matrix with dimensionality *pk* x 1, $\mathbf{X}$ is a matrix of dimensionality *pk* x *p* whose elements are $x_{m,n} \in \{0,1\}$, $\boldsymbol{\delta}$ is matrix of dimensionality *p* x 1 whose elements are the common effect-size parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a matrix of dimensionality *pk* x 1 whose elements are differences between the observed values and model estimates. The elements in $\boldsymbol{\delta}$ are referred to as common effect-size parameters since a fixed-effects approach was assumed. The elemental structure of $\mathbf{X}$, typically referred to as the design matrix, allows the analyst to incorporate an incomplete effect-size structure into the model. Both the design and dimensionality of $\mathbf{X}$ vary

according to specific effect sizes obtained from each study in a multivariate meta-analysis. An expanded form of the linear regression model relevant to this thesis is

$$\begin{bmatrix} d_{11} \\ d_{12} \\ \vdots \\ d_{k1} \\ d_{k2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{2k} \end{bmatrix}. \tag{25}$$

One basic form of estimation is ordinary least squares (OLS) estimation. A key assumption of OLS is $\boldsymbol{\varepsilon} \backsim N(0, \sigma_\varepsilon^2)\ \forall\ \varepsilon_i \in \boldsymbol{\varepsilon} \mid i: 1, \ldots, k$. In other words, the elements of the error matrix are assumed to follow a normal distribution with a mean of zero and common variance. Also, OLS assumes the elements of the error matrix to be uncorrelated. Given the satisfaction of all assumptions, the OLS coefficients are given by

$$\widehat{\boldsymbol{\delta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{d} \tag{26}$$

$$\Leftrightarrow$$

$$\begin{bmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \\ 0 & 1 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \\ 0 & 1 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} d_{11} \\ d_{12} \\ \vdots \\ d_{k1} \\ d_{k2} \end{bmatrix}, \tag{27}$$

where $\widehat{\boldsymbol{\delta}}$ is a matrix of dimensionality *p* x 1 whose elements are the maximum likelihood estimates of the common effect-size parameters (Gross, 2003).

The preceding discussion took the typical linear regression form and applied it to the meta-analysis context. When the assumptions of OLS are violated, other estimation procedures are available. GLS allows for heterogeneity and correlation among error variances. This estimation method is able to produce accurate estimates of the parameters by incorporating a variance-covariance matrix, as shown by Raudenbush, Becker, and Kalaian (1988):

$$\widehat{\boldsymbol{\delta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{d} \tag{28}$$

such that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma_k} \end{bmatrix} \tag{29}$$

and

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma^2_{\delta_{i1}} & cov(\delta_{i1}, \delta_{i2}) \\ cov(\delta_{i1}, \delta_{i2}) & \sigma^2_{\delta_{i2}} \end{bmatrix} : i \in [1, k] \,. \tag{30}$$

The elements of $\widehat{\boldsymbol{\delta}}$ are the best linear unbiased estimators of the common effect-size parameters, $\boldsymbol{\Sigma}$ is a block-diagonal matrix whose elements are population variance-covariance matrices of effect sizes for k studies, $\sigma^2_{\delta_{i1}}$ and $\sigma^2_{\delta_{i2}}$ are the population effect-size variances for outcomes 1 and 2 from the i^{th} study, respectively, and $cov(\delta_{i1}, \delta_{i2})$ is the population effect-size covariance from the i^{th} study. The expanded form of the GLS estimator is similar to the OLS estimator and is not shown. Although $\boldsymbol{\Sigma}_i$ is typically unknown, it is well-approximated by

$$\mathbf{V}_i = \begin{bmatrix} v_{i1} & cov(d_{i1}, d_{i2}) \\ cov(d_{i1}, d_{i2}) & v_{i2} \end{bmatrix}, \tag{31}$$

where $\mathbf{V}_i$ is the sample effect-size variance-covariance matrix for the i^{th} study, v_{i1} and v_{i2} are the sample effect-size variances for outcomes 1 and 2, respectively, and $cov(d_{i1}, d_{i2})$ is the sample effect-size covariance from the i^{th} study (Raudenbush, Becker, & Kalaian, 1988). Effect-size variance-covariance matrices were arrayed in a single effect-size variance-covariance matrix by way of direct summation:

$$\mathbf{V} = \oplus_{\mathrm{i=1}}^{\mathrm{k}} \mathbf{V}_i \,, \tag{32}$$

where $\mathbf{V}$ is the block-diagonal variance-covariance matrix for a single meta-analysis. In a block-diagonal matrix the elements on the main diagonal are square matrices ($\mathbf{V}_i$) and all other elements are zero (Schott, 2005).

A notable difference between the simulation and model estimation procedures is the requirement of an effect-size covariance estimate. This calculation requires an estimate of the raw-data correlation, which is discussed later. The population effect-size covariance for the standardized-mean-difference effect size is

$$cov(\delta_{i1}, \delta_{i2}) = \left(\frac{1}{n_i^{\mathrm{T}}} + \frac{1}{n_i^{\mathrm{C}}}\right)\rho + \left(\frac{\delta_{i1}\delta_{i2}}{2(n_i^{\mathrm{T}} + n_i^{\mathrm{C}})}\right)\rho^2 \,, \tag{33}$$

where ρ is the Pearson product-moment correlation parameter for the observed values (Gleser & Olkin, 2009). Since this population parameter is typically not available, it is well-approximated by the sample effect-size covariance,

$$cov(d_{i1}, d_{i2}) = \left(\frac{1}{n_i^{\mathrm{T}}} + \frac{1}{n_i^{\mathrm{C}}}\right) r + \left(\frac{d_{i1} d_{i2}}{2(n_i^{\mathrm{T}} + n_i^{\mathrm{C}})}\right) r^2 , \tag{34}$$

where r is the sample Pearson product-moment correlation for the raw data. Last, the multivariate Q statistic was calculated:

$$Q_{\mathrm{M}} = (\mathbf{d} - \mathbf{X}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{d}])'\mathbf{V}^{-1}(\mathbf{d} - \mathbf{X}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{d}]) \tag{35}$$

$$\Leftrightarrow$$

$$Q_{\mathrm{M}} = \left(\mathbf{d} - \mathbf{X}\hat{\boldsymbol{\delta}}\right)'\mathbf{V}^{-1}\left(\mathbf{d} - \mathbf{X}\hat{\boldsymbol{\delta}}\right). \tag{36}$$

The notation of Q_{M} denotes a multivariate Q statistic, versus the univariate Q statistic from section 3.2. The degrees of freedom for Q_{M} is equal to the difference between $\dim \mathbf{d}(p)$ and $\dim \boldsymbol{\delta}(p)$. The term $\dim \mathbf{d}(p)$ is the dimensionality of the observed effect-size vector, as a function of the outcomes; the term $\dim \boldsymbol{\delta}(p)$ is the dimensionality of the effect size parameter vector, as a function of the outcomes. Contextually, $\dim \mathbf{d}(p)$ is the total number of effect sizes that are to be included in the analysis and $\dim \boldsymbol{\delta}(p)$ is the number of effect-size parameters to be estimated. When all studies provide information for all outcomes, as was assumed in this thesis, the degrees of freedom can be revised as

$$\dim \mathbf{d}(p) - \dim \boldsymbol{\delta}(p) = kp - p = p(k-1) . \tag{37}$$

As with the univariate Q statistic, Q_{M} allows for the calculation of other homogeneity indices, mainly Birge's ratio and I^2. The calculation of Birge's ratio and I^2 indices are identical to Equation 21 and Equation 22, respectively, besides the revised degrees of freedom for the multivariate case (see Equation 37).

3.3.2 Model Estimation Procedure

The initial steps of the model estimation procedure emulated procedures used in the simulation section. Data generation procedures were identical to those shown in section 3.1. Resulting raw-data matrices remained of the form shown in Figure 3.1. Standardized-mean-

difference effect sizes and their variances were calculated using Equations 16 and 17, respectively. However, the use of dependency structures was not required. The Pearson product-moment correlation was calculated for the raw data. More specifically, the correlation of the following two data vectors was calculated:

$$\mathbf{Y}_{i1} = \begin{bmatrix} Y_{i11}^{\mathrm{T}} \\ Y_{i12}^{\mathrm{T}} \\ \vdots \\ Y_{i1n}^{\mathrm{T}} \\ Y_{i11}^{\mathrm{C}} \\ Y_{i12}^{\mathrm{C}} \\ \vdots \\ Y_{i1n}^{\mathrm{C}} \end{bmatrix} \tag{38}$$

and

$$\mathbf{Y}_{i2} = \begin{bmatrix} Y_{i21}^{\mathrm{T}} \\ Y_{i22}^{\mathrm{T}} \\ \vdots \\ Y_{i2n}^{\mathrm{T}} \\ Y_{i21}^{\mathrm{C}} \\ Y_{i22}^{\mathrm{C}} \\ \vdots \\ Y_{i2n}^{\mathrm{C}} \end{bmatrix}, \tag{39}$$

where $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$ are the observed data vectors for outcomes 1 and 2 for the i^{th} study, respectively. This approach correlates values Y_{i1n}^{T} with Y_{i2n}^{T} and Y_{i1n}^{C} with Y_{i2n}^{C}. A discussion regarding an alternative approach that was used to estimate the raw-data correlation is presented later. This correlation value was used to calculate the effect-size covariance using Equation 34. Effect-size variance and covariance estimates were arrayed in a variance-covariance matrix identical to forms discussed in section 3.2.

This process was completed for *k* studies in a single meta-analysis. Effect-size estimates were stacked vertically, alternating by outcome. This procedure differs from the methodology used in the simulation section. Resulting effect-size matrices for a single meta-analysis were of the form:

$$\begin{bmatrix} d_{11} \\ d_{12} \\ \vdots \\ d_{k1} \\ d_{k2} \end{bmatrix}.$$

The common effect-size estimates were not of interest in this thesis, but were required for the calculation of Q_{M} (see Equation 36).

Following with the simulation section, the above procedures were replicated 3,000 times under multiple conditions, each providing a single Q_{M}, Birge's ratio, and I^2 estimates. Table 3.2 describes the conditions in the model estimation section.

Table 3.2: Model Estimation Conditions.

Raw-Data Correlation	Between–Outcomes Correlation (ρ)	Number of Studies (k)	Sample Size (n)
Hypothetical (ρ)	.50	24	20
Empirical (r)	.75	48	100
	.99	96	180
			260

The between-outcomes correlation, number of studies, and individual study sample size conditions remain unchanged from the simulation section. However, the dependency structure was no longer applicable. Recall that the simulation section treated multivariate data as univariate data during analysis. This required that dependence be artificially imposed in analyses. The model estimation section analyzes data using multivariate methodology (i.e., the data is never partitioned into univariate data sets). This method incorporates the dependence directly into estimation using effect-size variance-covariance matrices.

An additional condition, raw-data correlation, was evaluated. This correlation refers to the Pearson product-moment correlation that is calculated for the raw data and is used when calculating effect-size covariances (see Equation 34). Raw-data correlations are not well-reported in the social-science literature and are often difficult to obtain. By utilizing what I refer to as the *hypothetical* raw-data correlation, I am supposing that I do not have access to sample-specific correlations but have the ability to estimate their approximate values. For example, Becker (1990) looked at the effect of coaching on Scholastic Aptitude Test (SAT) mathematics and SAT verbal outcomes. She set pretest-posttest correlations for both the SAT mathematics

and SAT verbal scores to $r = .88$, as access to the complete SAT data was not readily available. The hypothetical condition referred to the raw-data correlation ρ, which is identical to the between-outcomes correlation. Since data was generated to be correlated to an approximate degree of ρ, it is reasonable to use this value as an "estimate" of the raw-data correlation if one were to assume the sample correlation to be unavailable.

Conversely, the *empirical* raw-data correlation supposes that the true correlation is readily available. The value of the raw-data correlation under the empirical condition is calculated as the Pearson product-moment correlation, as discussed earlier in this section. I compared sets of results between these conditions and assessed differences.

CHAPTER FOUR

RESULTS

The goal of this thesis was to assess the impact of multiple endpoint dependency on select homogeneity measures. Type I error rates of Q statistics, as well as mean Birge's ratio and I^2 values, were analyzed to achieve this goal. The focus of the results section is on the Type I error rates of Q statistics. The methodology used in the simulation section assumed independence when dependence was present under varying conditions (see Table 3.1); the methodology used in the model estimation section utilized many of the same conditions as the simulation section (see Table 3.2) while accounting for existing dependence in the data. Results for all homogeneity measures under all conditions are visually represented by scatterplot matrices with reference lines and confidence intervals when applicable.

4.1 Simulation Results

All procedures in sections 3.1 and 3.2 were performed as described. Figure 4.1 shows Type I errors rates of univariate Q statistics for all conditions. The y-axis represents the Type I error rates, the x-axis represents individual study sample sizes, row panels represent the number of studies within a meta-analysis, column panels represent the between-outcomes correlation, and markers represent dependency structures. The dotted lines denote the 95% confidence interval described in section 3.2; this interval displays nominal levels of Type I error rates. Results are discussed by dependency structure.

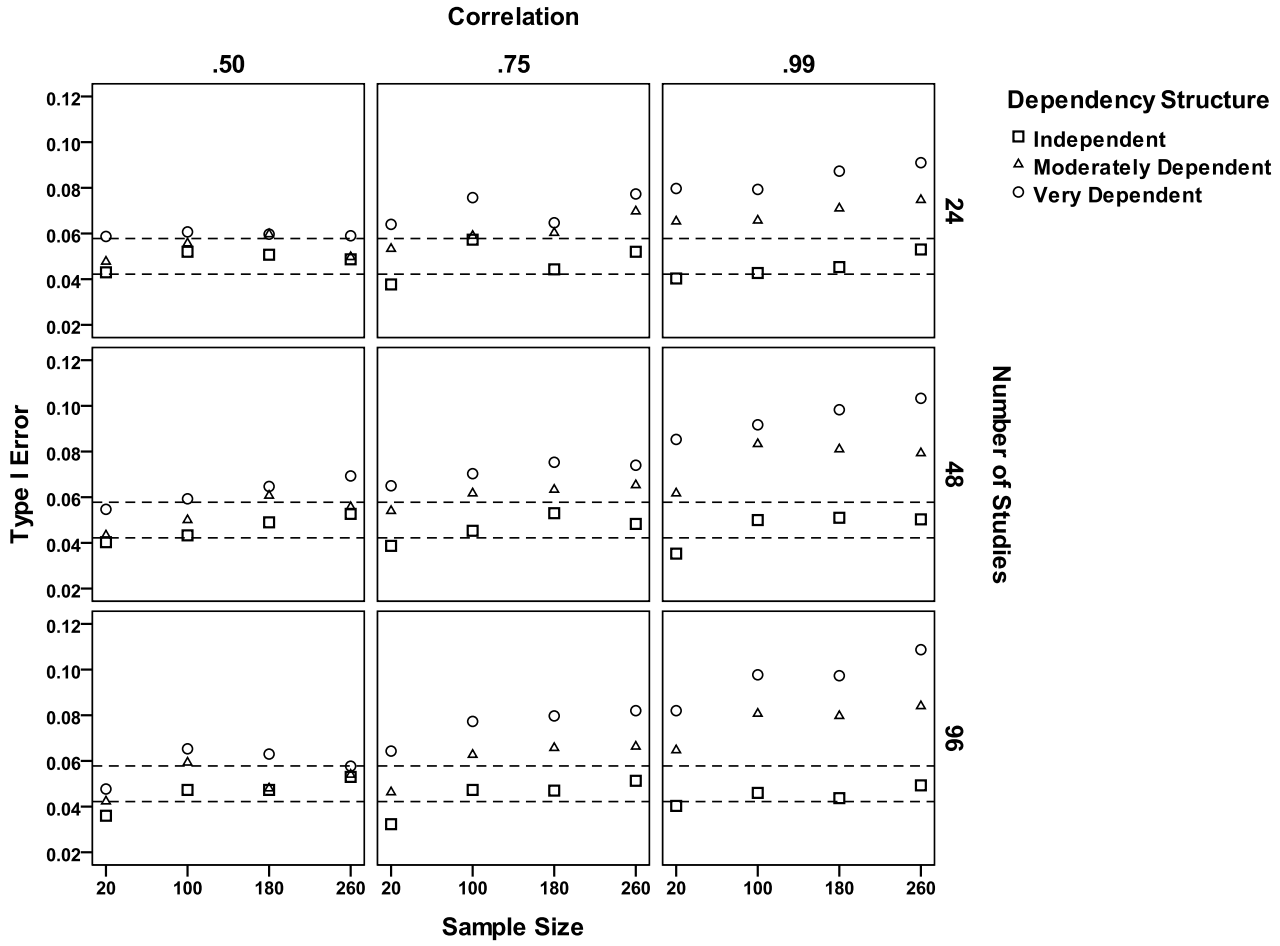

Figure 4.1: Type I Error Rates for Univariate *Q* Statistics.

Throughout all conditions under the independent structure (indicated by □), Type I error rates were never elevated above the nominal level of .05. When error rates were statistically significant, the error rate was below the nominal level (e.g., $\rho = .50$, $k = 96$, $n = 20$). For many cases where $n = 20$, conditions with independent structures have Type I error rates that are significantly lower than their nominal value. Harwell (1997) discusses how the *Q* statistic tends to have a low Type I error rate when study sample sizes are small, therefore this result is expected. Recall that the independent structure is considered a baseline structure because no dependence is present in the model. All results, whether statistically significant or within nominal levels, were consistent with previous findings regarding the *Q* statistic, indicating an adequate baseline structure from which dependency structure manipulation was appropriate.

Overall, conditions with moderately dependent structures (indicated by Δ) tended to produce Type I error rates larger than those for similar conditions under the independent

structure. When ρ = .50, the majority of the Type I error rates were within nominal levels. As the between-outcomes correlation increased, Type I error rates increased, often substantially. When the between-outcomes correlation was very strong (ρ = .99), all Type I error rates for conditions under the moderately dependent structure were significantly larger than their nominal values, the largest of which was .082.

Trends of the Type I error rates (nominal or significant) for conditions under the very dependent structure (indicated by ○) did not fluctuate much from the results previously mentioned regarding the moderately dependent structure. However, significance levels were more pronounced. The highest Type I error rate for these conditions was .109, more than twice the size of the expected Type I error rate of .05. As with the moderately dependent structure, all combinations of conditions with very dependent structures and strong between-outcomes correlations (ρ = .99) produced Type I error rates significantly higher than their nominal levels. It appears that when outcomes are highly-correlated and analyzed together, Type I error rates become inflated regardless of individual study sample size or the number of studies. Next I briefly discuss the mean Birge's ratio and I^2 results from the simulation section.

Figure 4.2 shows mean Birge's ratio values from the 3,000 replicated meta-analyses for all conditions. The y-axis represents the mean Birge's ratio, the x-axis represents individual study sample sizes, row panels represent the number of studies within a meta-analysis, column panels represent the between-outcomes correlation, and markers represent dependency structures. The dotted line denotes a mean Birge's ratio equal to one, the expected value under the null model.

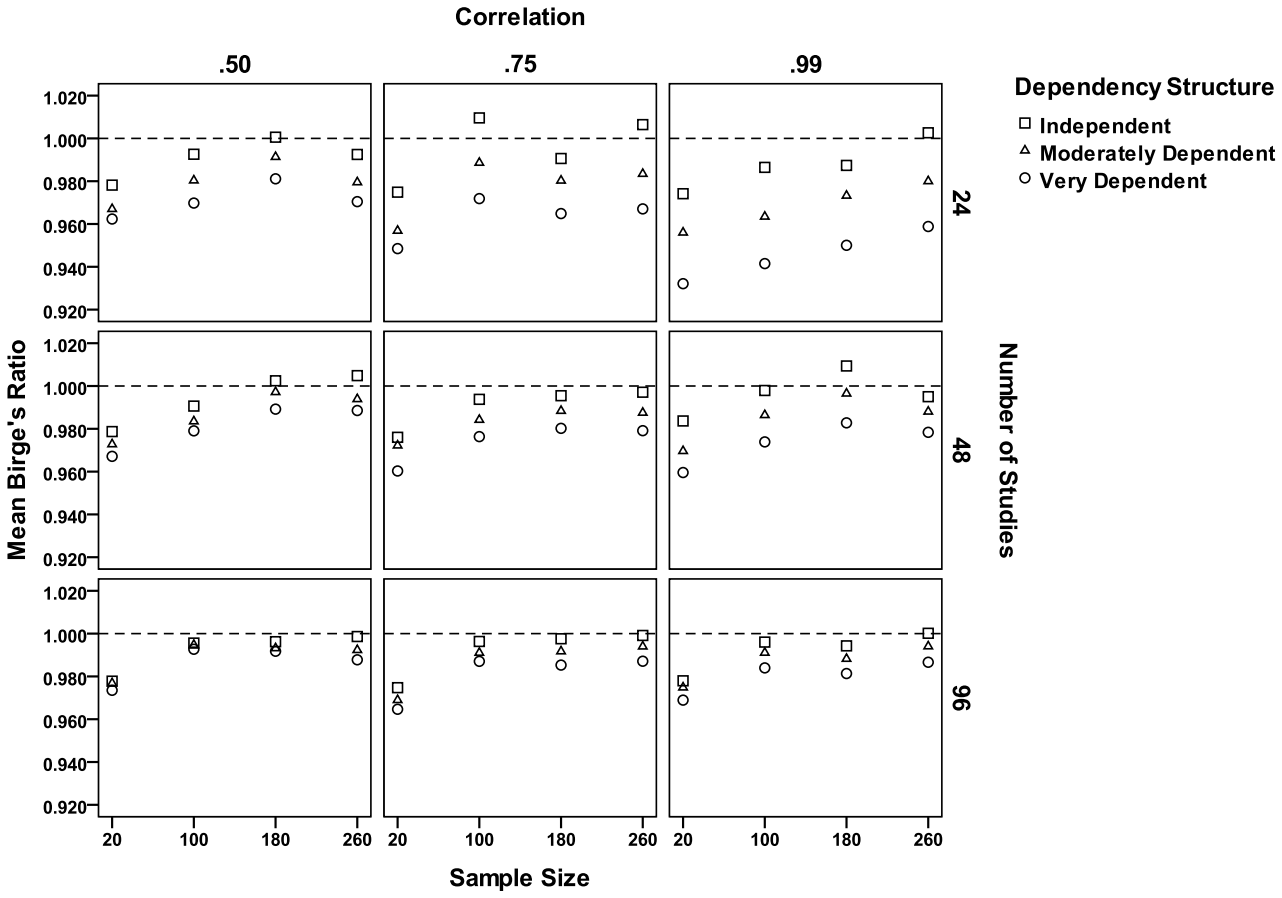

Figure 4.2: Mean R_B (Simulation Section).

It is evident that the majority of mean Birge's ratios are less than one, which implies that, on average, the Q statistic for a given condition was typically less than its respective degrees of freedom. Smaller Q statistics coincide with increased amounts of homogeneity. There was an overwhelming presence of homogeneity among the data in all conditions. In particular, an increase in dependence (denoted by the dependency structure) corresponded with a small decrease in mean Birge's ratio values, which represents an increase in effect-size homogeneity. Figure 4.3 shows 95% confidence intervals for mean Birge's ratio results from the simulation section for a select number of cases (where ρ = .99).

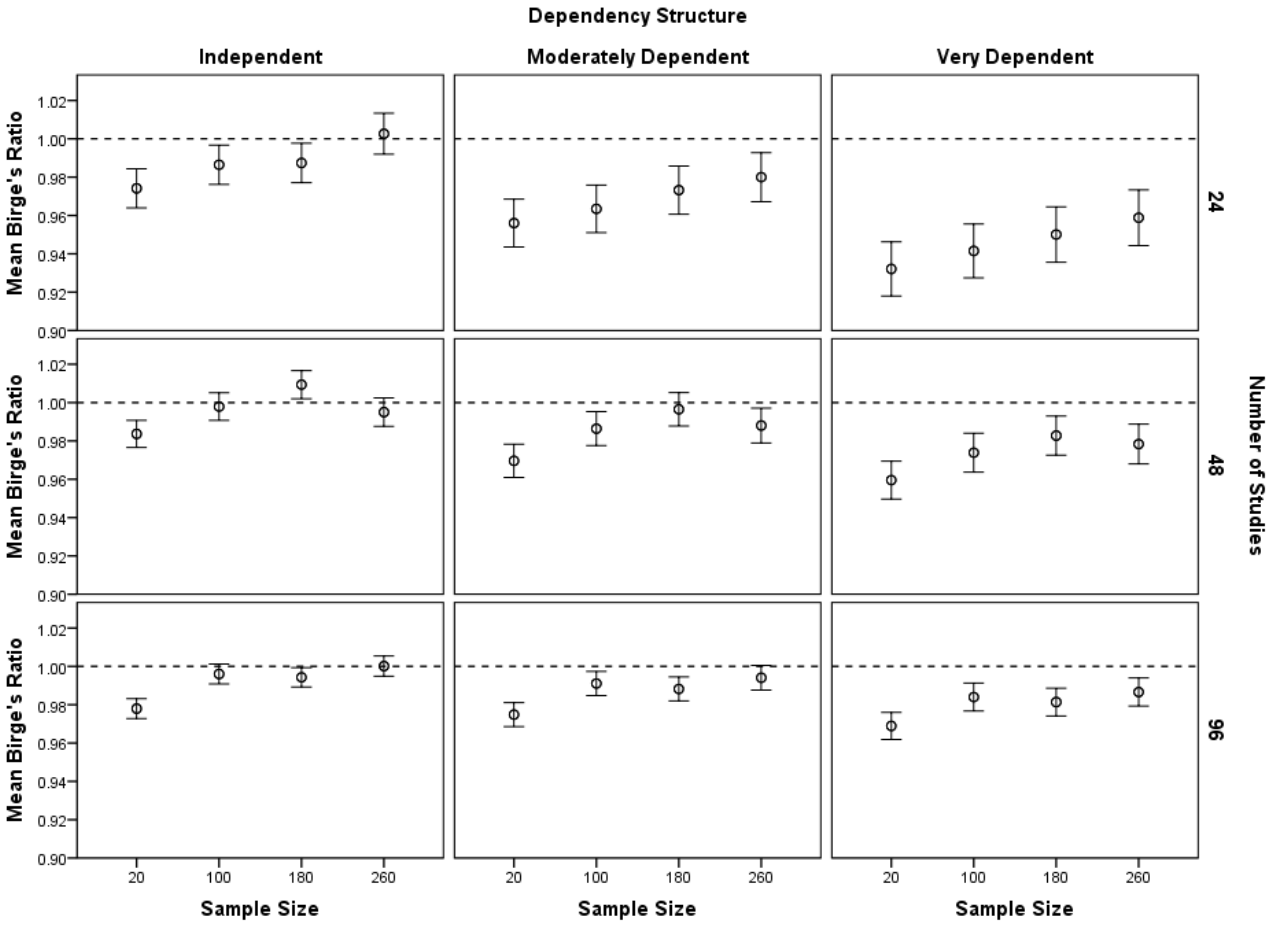

Figure 4.3: Mean R_B (Simulation Section) with 95% Confidence Intervals where $\rho = .99$.

The confidence intervals in Figure 4.3 show that although results varied across dependency structure levels as previously mentioned, many of the discrepancies were not significantly different. In a few instances the mean Birge's ratio was greater than one (e.g., $\rho = .99$, $k = 48$, $n = 180$), but only occurred with independent structures.

For all results shown in Figure 4.2 it is important to account for the scale of the y-axis. Although points may seem to vary (e.g., in the upper right cell), the range of the y-axis is not large at all. For example, despite the visual distance between $R_B = .940$ and $R_B = .960$, contextually these mean Birge's ratios are extremely similar. The standard deviations of Birge's ratio values across all conditions (not shown) were roughly in the interval [.10, .40], which does indicate a moderate degree of variability. An increase in dependency structure corresponded with a slight increase in the characteristic mean Birge's ratio standard deviation.

Figure 4.4 shows mean I^2 values from the 3,000 replicated meta-analyses for all conditions. The y-axis represents the mean I^2, the x-axis represents individual study sample sizes, row panels represent the number of studies within a meta-analysis, column panels represent the between-outcomes correlation, and markers represent dependency structures.

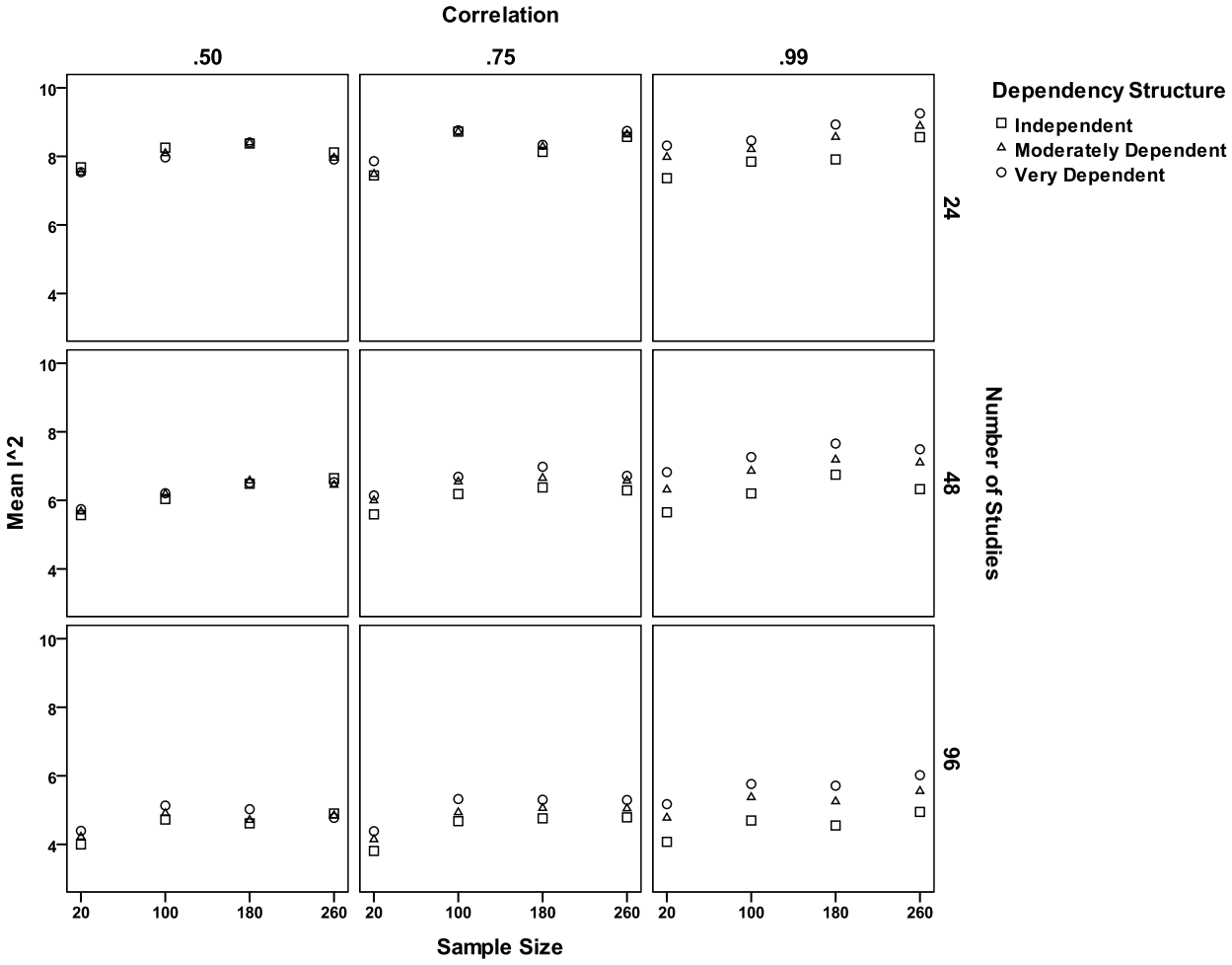

Figure 4.4: Mean I^2 (Simulation Section).

Mean I^2 results were similar to those for mean Birge's ratio values. For moderate levels of between-outcomes correlation (ρ = .50) there was virtually no distinction among mean I^2 values with respect to dependency structure, controlling for sample size and number of studies. However, increases in between-outcomes correlation showed slight discrepancies in mean I^2 values for dependency structures, again controlling for sample sizes and number of studies. The discrepancies displayed slightly elevated mean I^2 values for stronger dependency structures.

Figure 4.5 shows 95% confidence intervals for mean I^2 results from the simulation section for a select number of cases (where ρ = .99).

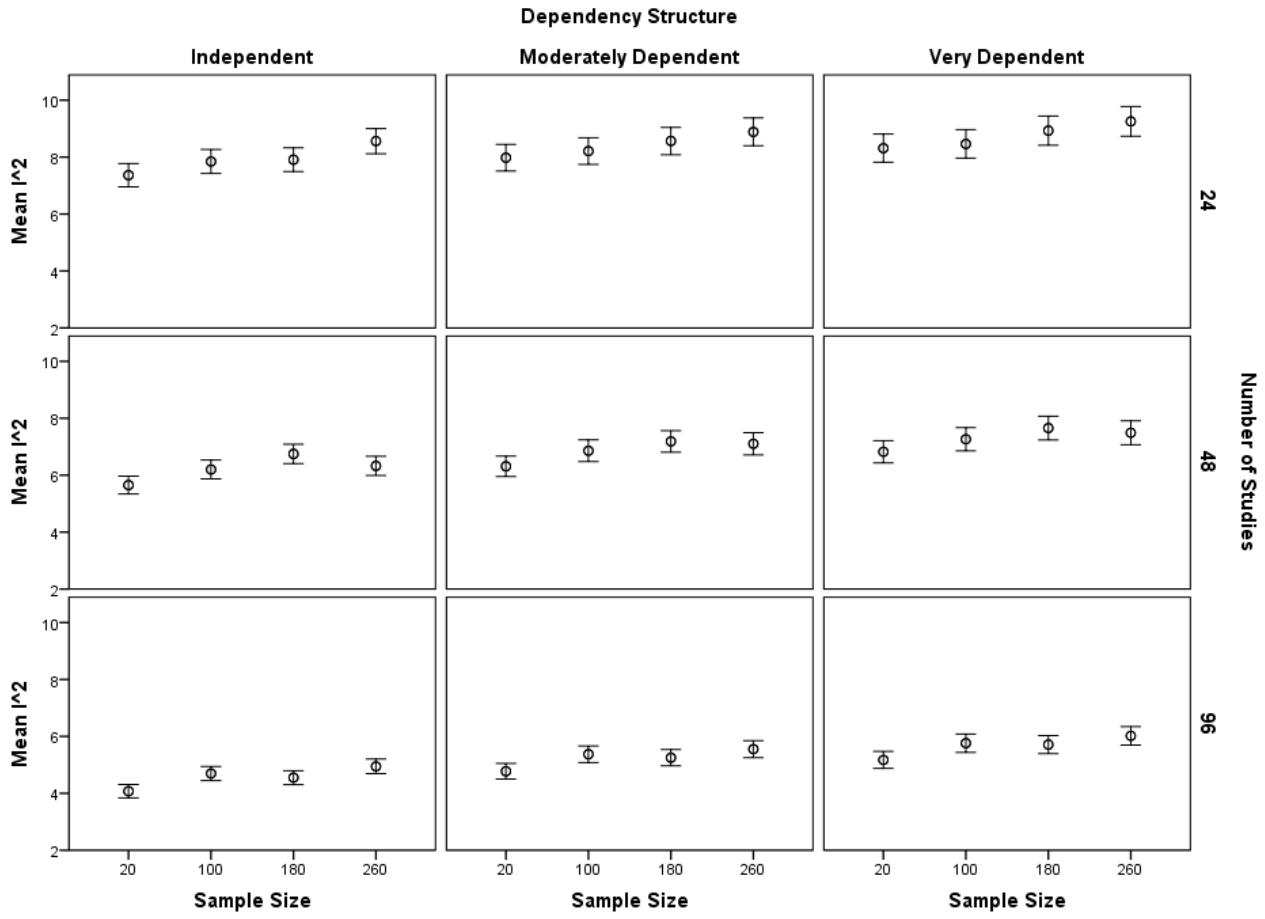

Figure 4.5: Mean I^2 (Simulation Section) with 95% Confidence Intervals where ρ = .99.

As with Figure 4.3, the confidence intervals in Figure 4.5 show that although results varied across dependency structure levels, many of the discrepancies were not significantly different.

Similar to the mean Birge's ratio results, the scale of the y-axis is crucial. Although points may seem to vary (e.g., in the lower right cell), the range of the y-axis is not large at all. For example, despite the visual distance between I^2 = 4 and I^2 = 6, contextually these mean I^2 values are extremely similar. The standard deviations of I^2 values across all conditions (not shown) were roughly in the interval [7, 14], which does indicate a moderate degree of variability. Across conditions, I^2 values were very small according to the interpretations presented by Higgins et al. (2003). The largest I^2 value was lower than 10, which corresponds to minimal

heterogeneity among effect sizes. This suggests that I^2 was not impacted as much as the Q statistic was by the introduction of dependence into univariate meta-analysis. A similar interpretation can be made regarding the mean Birge's ratio results previously discussed.

Furthermore, the results for the mean Birge's ratio and mean I^2 values seem contradictory to those shown by the Q statistics. The mean Birge's ratio and I^2 values indicated that increases in dependence corresponded to slight increases in effect-size homogeneity. Contradictorily, the Q statistic results indicated that increases in dependence corresponded with increased Type I error rates, which implies a stronger presence of heterogeneity. The results for mean Birge's ratio and I^2 values can be explained by increases in dependence involving increases in effect-size correlation, which are indicative of stronger homogeneity among data. The univariate and multivariate Q statistics follow asymptotic chi-square distributions under the fixed-effects model. Simulation results displayed mean Q statistics that were lower than their respective degrees of freedom and increased levels of activity in the tail of the distributions. Walsh (1947) showed that hypothesis tests involving chi-square statistics have increased Type I error rates associated with increases in intraclass correlation, which is the central concept surrounding increases in the dependency structure condition.

4.2 Model Estimation Results

The purpose of the model estimation section was to represent most of the same conditions from the simulation section while accounting for dependence using GLS methodology. For this reason, the dependency structure condition discussed in section 3.2 was not implemented in the model estimation section.

All procedures in sections 3.1 and 3.3, as well as applicable procedures from section 3.2, were performed as described. Figure 4.6 shows Type I error rates of multivariate Q for all conditions. The y-axis represents the Type I error rates, the x-axis represents individual study sample sizes, row panels represent the number of studies within a meta-analysis, column panels represent the between-outcomes correlation, and markers represent raw-data correlation types. The dotted lines denote the 95% confidence interval described in section 3.2; this interval displays nominal levels of Type I error rates.

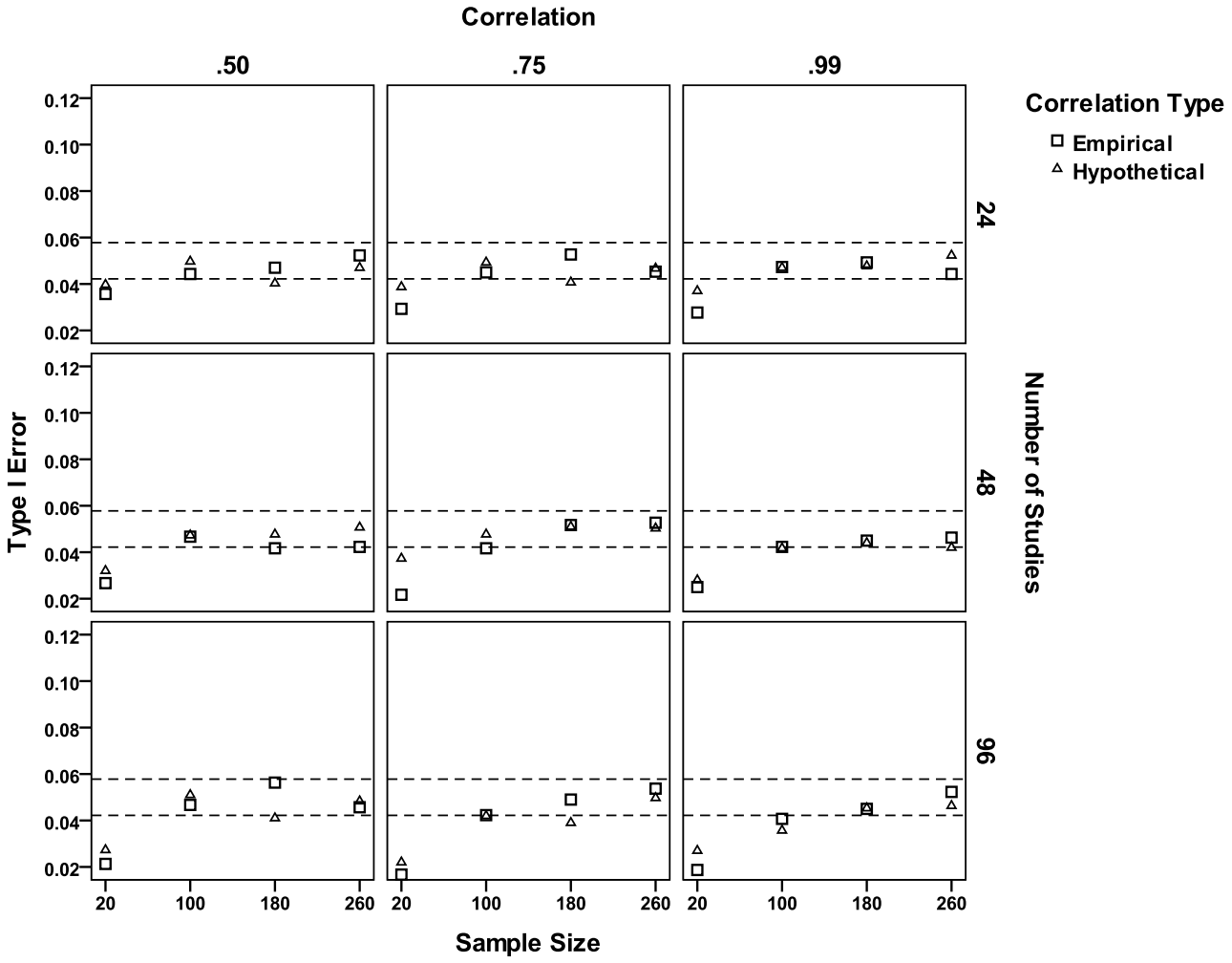

Figure 4.6: Type I Error Rates for Multivariate *Q* Statistics.

Results in Figure 4.6 are different than those for the univariate *Q* statistics (see Figure 4.1). Except in instances where conditions had small sample sizes ($n = 20$), the vast majority of Type I error rates were within nominal levels. For small sample sizes, the Type I error rate of the *Q* statistic tends to be low (Harwell, 1997). Overall results for nominal Type I error rates across most conditions were as expected, because a theoretically justifiable method was used to model dependence rather than wrongfully assuming independence.

The comparison of conditions with respect to raw-data correlation type proved interesting. Recall that the empirical raw-data correlation conditions (indicated by □) utilized Pearson product-moment correlation coefficients while the hypothetical raw-data conditions (indicated by ○) utilized between-outcomes correlation values (see Table 3.2). For some

conditions (e.g., ρ = .99, k = 48, n = 100) the Type I error rates were virtually indistinguishable between raw-data correlation types. Although other conditions (e.g., ρ = .75, k = 48, n = 20) produced Type I error rates that were visually distinguishable, these discrepancies might be attributed to sampling error. Also, these discrepancies seemed more predominant with small sample sizes. Therefore, for select conditions with larger sample sizes, the model estimation section procedures were re-run with 10,000 replications. Results (not shown) confirmed that the discrepancies between results for empirical and hypothetical conditions were primarily present with small sample sizes; conditions with larger sample sizes showed exceedingly similar results. Next I briefly discuss the mean Birge's ratio and I^2 results from the model estimation section.

Figure 4.7 shows mean Birge's ratio values from the 3,000 replicated meta-analyses for all conditions. The y-axis represents the mean Birge's ratio, the x-axis represents individual study sample sizes, row panels represent the number of studies within a meta-analysis, column panels represent the between-outcomes correlation, and markers represent raw-data correlation types. The dotted line denotes a mean Birge's ratio equal to one.

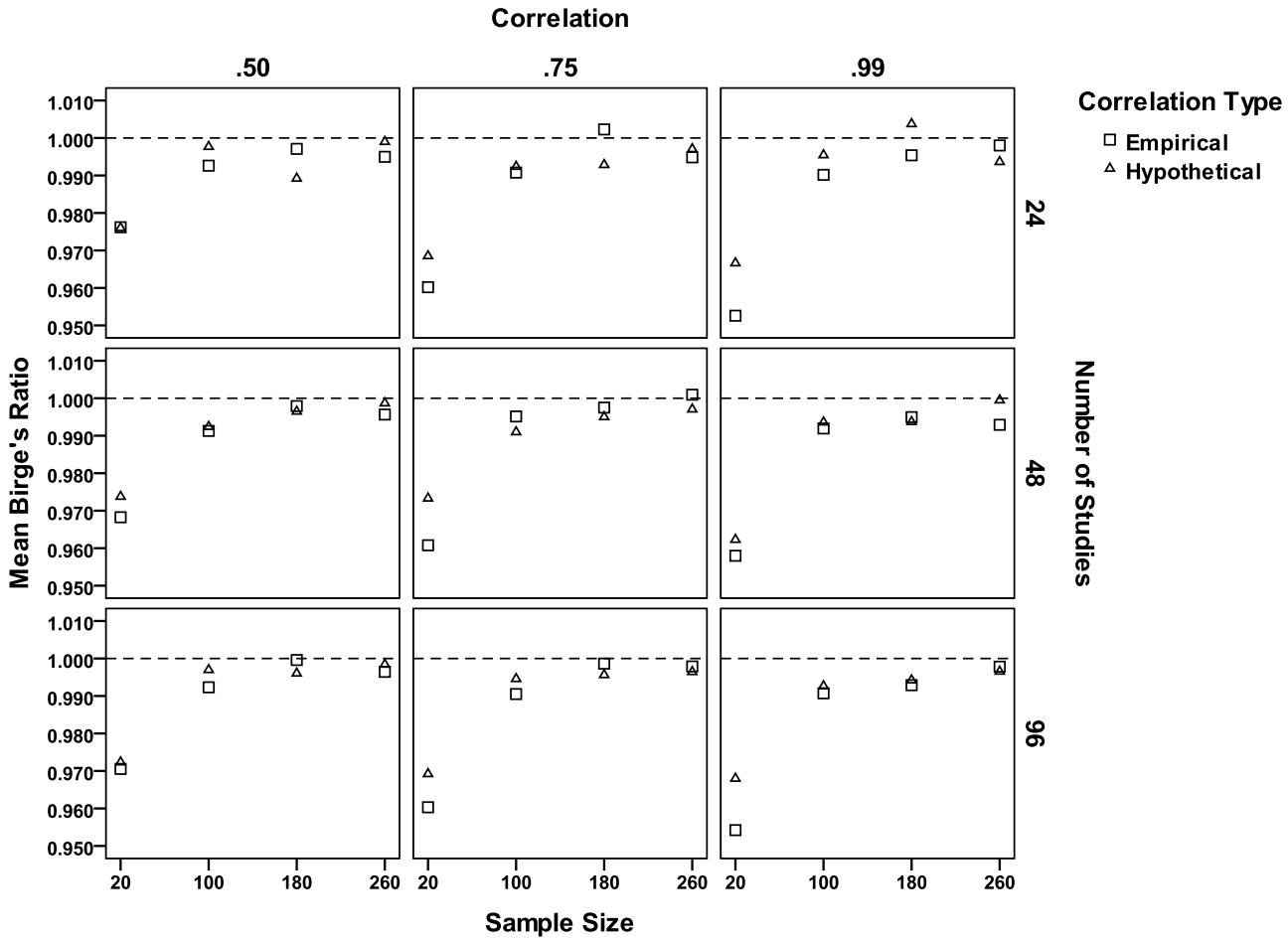

Figure 4.7: Mean R_B (Model Estimation Section).

Results shown in Figure 4.7 are similar to those from the simulation section (see Figure 4.2). A somewhat curvilinear trend existed as sample sizes increased across conditions. This trend seems to be invariant to the between-outcomes correlation condition, indicating that a decrease in sample size corresponded to a slight increase in homogeneity. Figure 4.8 shows 95% confidence intervals for mean Birge's ratio results from the model estimation section for a select number of cases (where ρ = .75).

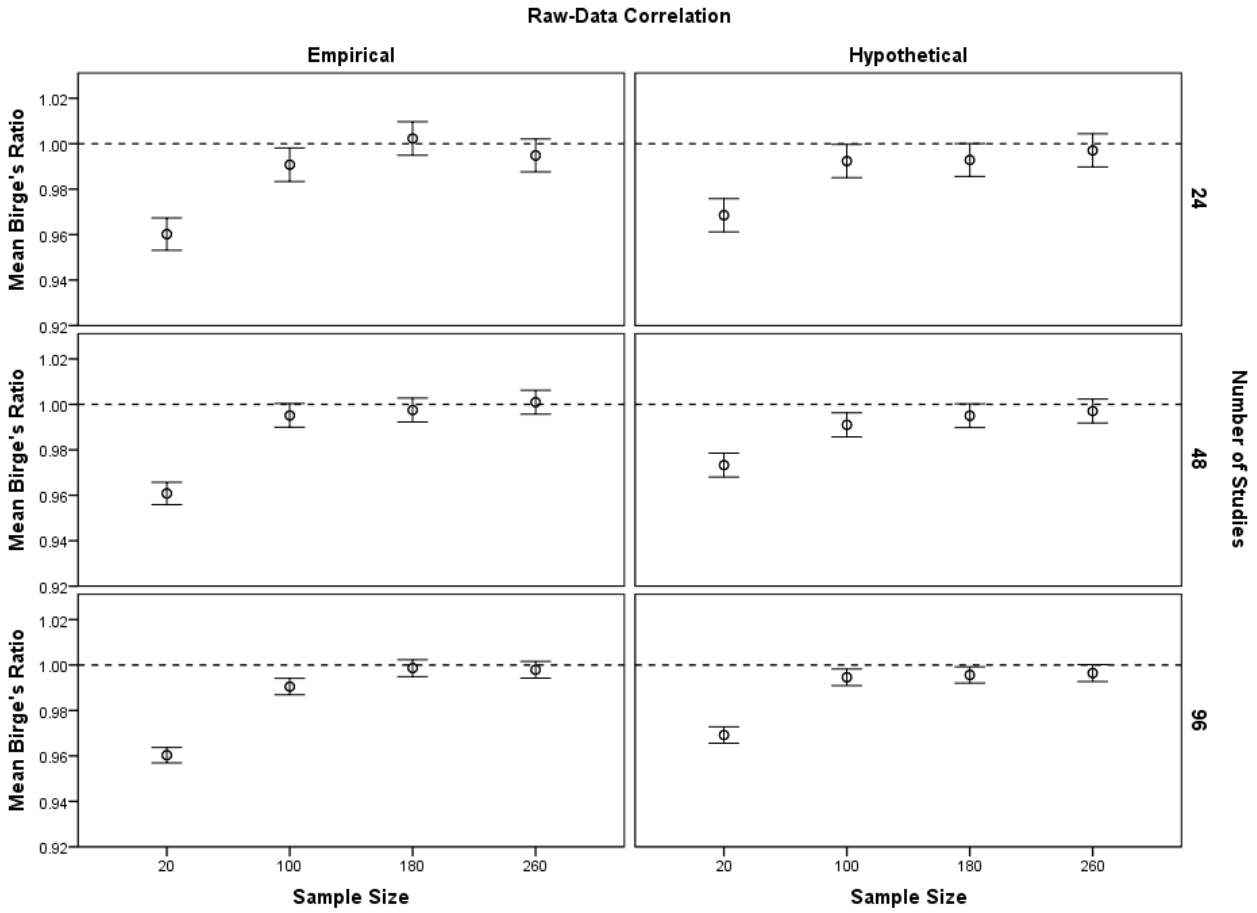

Figure 4.8: Mean R_B (Model Estimation Section) with 95% Confidence Intervals where $\rho = .75$.

The confidence intervals in Figure 4.8 show that although results varied across raw-data correlation levels, many of the discrepancies were not significantly different. As with multivariate Q statistic results, when the number of replications was increased to 10,000 results (not shown) across raw-data correlation levels were virtually identical when study sample sizes were not small. As with Figure 4.2, it is important to consider the condensed scale of the y-axis in Figure 4.7 when interpreting results. The standard deviations of Birge's ratio values across all conditions (not shown) were roughly in the interval [.09, .21], which does not indicate much variability.

Figure 4.9 shows mean I^2 values from the 3,000 replicated meta-analyses for all conditions. The y-axis represents the mean I^2, the x-axis represents individual study sample sizes, row panels represent the number of studies within a meta-analysis, column panels represent the between-outcomes correlation, and markers represent raw-data correlation types.

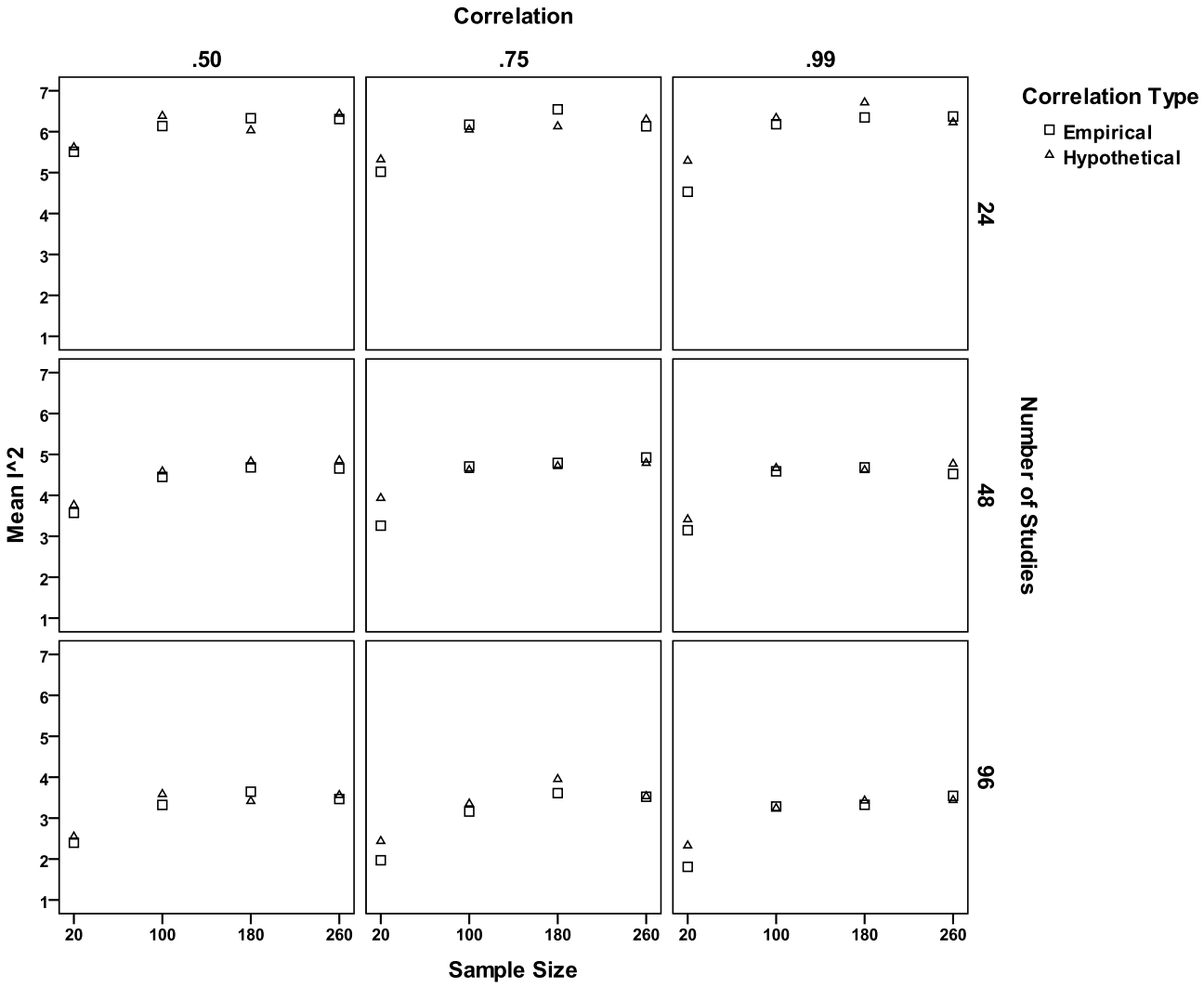

Figure 4.9: Mean I^2 (Model Estimation Section).

The results for mean I^2 values are similar to those for mean Birge's ratio values. A decrease in sample size corresponded to a slight increase in homogeneity. There were some slight discrepancies among results with respect to raw-data correlation type. Figure 4.10 shows 95% confidence intervals for mean I^2 results from the model estimation section for a select number of cases (where ρ = .75).

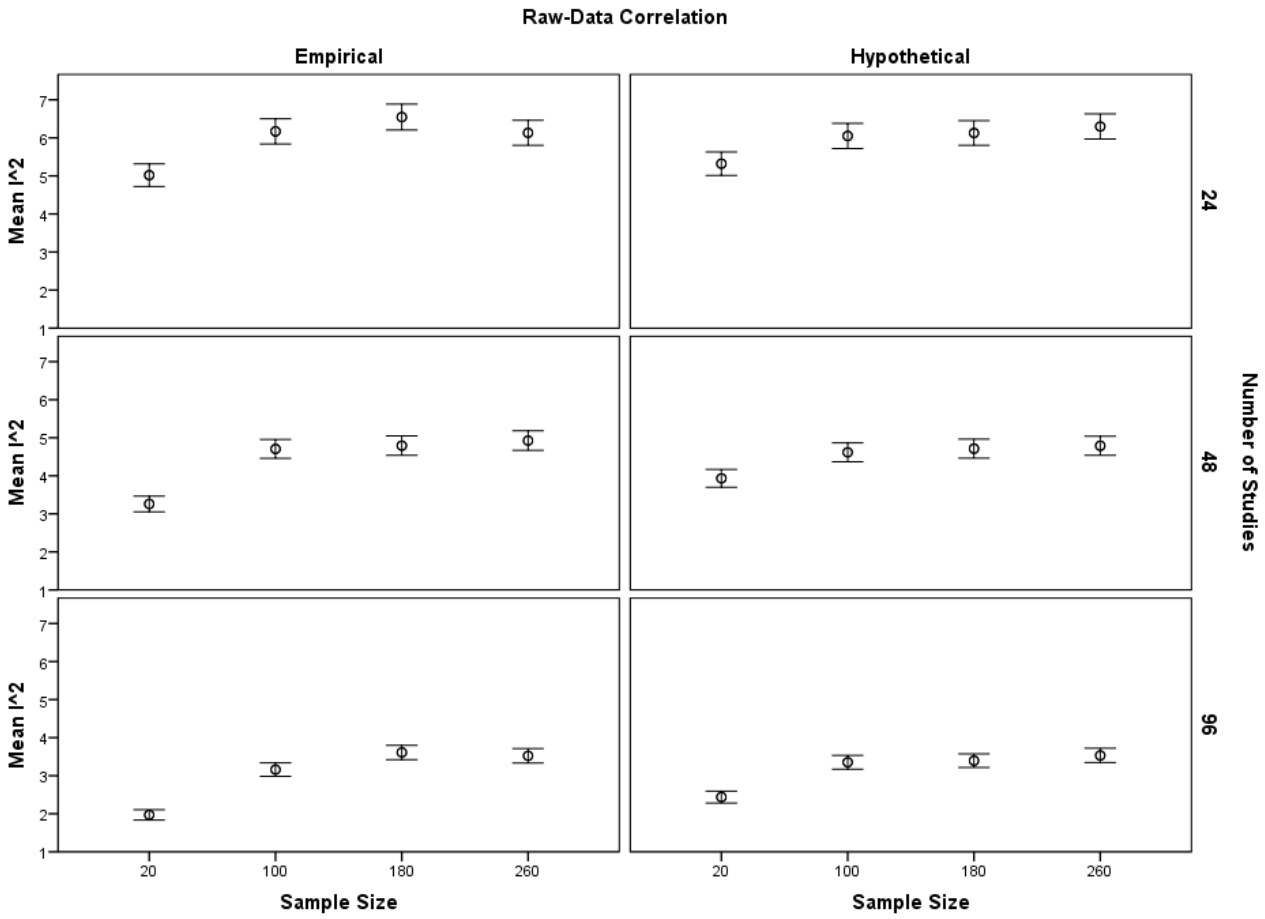

Figure 4.10: Mean I^2 (Model Estimation Section) with 95% Confidence Intervals where ρ = .75.

Similar to the Birge's ratio results from the model estimation section (see Figure 4.8), the confidence intervals in Figure 4.10 show that although results varied across raw-data correlation levels, many of the discrepancies were not significantly different. When the number of replications was increased to 10,000 results (not shown) across raw-data correlation levels were virtually identical when study sample sizes were not small. As with Figure 4.4, the scale of the y-axis in Figure 4.9 is crucial when interpreting results. The standard deviations of I^2 values across all conditions (not shown) were roughly in the interval [3, 9], which does not indicate much variability.

CHAPTER FIVE

CONCLUSIONS

The chapter begins with a summary of overall goals and a discussion of results for this thesis. A review of key results and how they may be informative to meta-analysts is presented. Last, some limitations are considered.

5.1 Discussion

The main purpose of this thesis was to investigate the impact of dependence that arises from multiple endpoint studies utilizing select homogeneity measures. Two different methods of meta-analysis were considered: a univariate approach, which ignores dependence, and a multivariate approach, which explicitly accounts for dependence. Conditions for the simulation section included study sample size, number of studies, between-outcomes correlation, and dependency structure (see Figures 3.2a – 3.2c). Conditions for the model estimation section included study sample size, number of studies, between-outcomes correlation, and raw-data correlation type. Outcomes for both sections were Type I error rates of Q statistics, mean Birge's ratio, and mean I^2 values. Emphasis was placed on the Type I error rates of Q statistics.

Results from the simulation section showed that when outcomes are highly correlated and analyzed together, Type I error rates of Q statistics become inflated. This result seemed to be relatively invariant to individual study sample size or number of studies. Mean Birge's ratio and I^2 results from the simulation section showed that as dependency levels increased, the two indices tended to display slightly more homogeneity. However, discrepancies due to increased dependency structures typically not significantly different and overall trends were only faintly evident. Overall results for these two indices seemed to not be affected by increased levels of dependence

Results from the model estimation section provided evidence that the GLS method of analyzing multivariate data proposed by Raudenbush, Becker, and Kalaian (1988) successfully accounts for dependence. Nominal levels of Type I error rates for Q statistics were present throughout all conditions, except when study sample sizes were small. Results for mean Birge's ratio and I^2 values were similar to those for the simulation section. Last, the comparison of

empirical and hypothetical raw-data correlation conditions showed minimal differences, except for conditions with small sample sizes. This comparison was evident only when the number of meta-analysis replications was increased to 10,000. A lower number of replications (3,000) showed minor discrepancies between the two methods of raw-data-correlation estimation.

The work of this thesis adds to the current discussion of dependence in the field of meta-analysis. Methods from the simulation section showed how homogeneity measures are affected when multiple endpoint data is treated univariately when performing meta-analysis. Researchers should refrain from implementing this approach because homogeneity results are likely to be affected. Instead, the model estimation section of this thesis showed that the GLS method of analyzing dependent effect sizes appears to successfully account for dependence present with multiple endpoint data. Type I error rates for Q statistics were typically within nominal levels for most conditions. Overall analyses from this thesis also showed that when studies with small sample sizes ($n \leq 20$) are included in meta-analyses, homogeneity measures do not perform nominally, regardless of the presence of dependence.

Last, a minor portion of this thesis investigated the comparison of empirical versus hypothetical raw-data correlations when used in the calculation of the multivariate Q statistics. Although this comparison should be studied more extensively, results showed that discrepancies in Q, mean Brige's ratio, and mean I^2 values were minimal. This supports the idea of using estimates of raw-data correlation when true correlations are not available. Discrepancies between results based on population and estimated correlations were larger with small study sample sizes.

5.2 Limitations

The scope of this thesis is not without its limitations. The issue of multiple endpoint dependency extends beyond the effect on homogeneity measures. Other methodological concerns and statistics were not considered. Also, both hypothetical groups used to calculate effect sizes were assumed to have equal sample sizes across all studies in a meta-analysis. While this is not a realistic assumption, the robustness of the results to this assumption has yet to be determined. Last, effect-size estimates were randomly generated such that the average effect size would equal to zero in both hypothetical groups. It is possible that other effect-size magnitudes, as well as differences in group effect-size magnitudes, may have effects on results.

APPENDIX A

CHOLESKY DECOMPOSITION

Let **A** be an *n* x *n* nonnegative definite matrix. There exists an *n* x *n* lower triangular matrix **L** with nonnegative diagonal elements such that $\mathbf{A} = \mathbf{L}\mathbf{L}'$ (Schott, 2005). The purpose of implementing a Cholesky decomposition was to generate correlated data. Chapter three discussed the correlation matrix, **r**, which is re-written as Equation A1

$$\mathbf{r} = \begin{bmatrix} 1 & \rho_{12}^{\mathrm{T}} & 0 & 0 \\ \rho_{12}^{\mathrm{T}} & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_{12}^{\mathrm{C}} \\ 0 & 0 & \rho_{12}^{\mathrm{C}} & 1 \end{bmatrix}. \tag{A1}$$

In this context, the goal was to generate observations for two outcomes from treatment and control groups which are correlated to a known degree (ρ) while all other grouping combinations are approximately uncorrelated. Suppose **A** is of the form

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}. \tag{A2}$$

The Cholesky decomposition allows us to re-write **A** into lower and upper 4 x 4 triangular matrices, **L** and $\mathbf{L}'$, respectively,

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \tag{A3}$$

and

$$\mathbf{L}' = \begin{bmatrix} l_{11} & l_{12}^{*} & l_{13}^{*} & l_{14}^{*} \\ 0 & l_{22} & l_{23}^{*} & l_{24}^{*} \\ 0 & 0 & l_{33} & l_{34}^{*} \\ 0 & 0 & 0 & l_{44} \end{bmatrix}, \tag{A4}$$

where $\mathbf{A} = \mathbf{L}\mathbf{L}'$. Below is algebraic derivation of the Cholesky decomposition for a 4 x 4 matrix:

$$\mathbf{L}(\mathbf{A}) = \begin{bmatrix} \sqrt{a_{11}} & 0 & 0 & 0 \\ \dfrac{a_{21}}{\sqrt{a_{11}}} & \sqrt{a_{22} - \dfrac{a_{21}^2}{a_{11}}} & 0 & 0 \\ \dfrac{a_{13}}{\sqrt{a_{11}}} & \dfrac{a_{32} - \dfrac{a_{13}a_{21}}{a_{11}}}{\sqrt{a_{22} - \dfrac{a_{21}^2}{a_{11}}}} & \sqrt{a_{33} - \dfrac{a_{13}^2}{a_{11}} - \dfrac{\left(a_{32} - \dfrac{a_{13}a_{21}}{a_{11}}\right)^2}{a_{22} - \dfrac{a_{21}^2}{a_{11}}}} & 0 \\ \dfrac{a_{41}}{\sqrt{a_{11}}} & \dfrac{a_{42} - \dfrac{a_{41}a_{21}}{a_{11}}}{\sqrt{a_{22} - \dfrac{a_{21}^2}{a_{11}}}} & \dfrac{a_{43} - \dfrac{a_{41}a_{13}}{a_{11}} - \dfrac{\left(a_{42} - \dfrac{a_{41}a_{21}}{a_{11}}\right)(a_{32} - a_{13})}{a_{22} - \dfrac{a_{21}^2}{a_{11}}}}{\sqrt{a_{33} - \dfrac{a_{13}^2}{a_{11}} - \dfrac{\left(a_{32} - \dfrac{a_{13}a_{21}}{a_{11}}\right)^2}{a_{22} - \dfrac{a_{21}^2}{a_{11}}}}} & \sqrt{a_{44} - \dfrac{a_{41}^2}{a_{11}} - \dfrac{\left(a_{42} - \dfrac{a_{41}a_{21}}{a_{11}}\right)^2}{a_{22} - \dfrac{a_{21}^2}{a_{11}}} - \left(\dfrac{a_{43} - \dfrac{a_{41}a_{13}}{a_{11}} - \dfrac{\left(a_{42} - \dfrac{a_{41}a_{21}}{a_{11}}\right)(a_{32} - a_{13})}{a_{22} - \dfrac{a_{21}^2}{a_{11}}}}{\sqrt{a_{33} - \dfrac{a_{13}^2}{a_{11}} - \dfrac{\left(a_{32} - \dfrac{a_{13}a_{21}}{a_{11}}\right)^2}{a_{22} - \dfrac{a_{21}^2}{a_{11}}}}}\right)^2} \end{bmatrix}$$

APPENDIX B

EFFECT-SIZE CORRELATION

Gleser and Olkin (2009) note that the correlation between multiple endpoint effect sizes is a result of the correlation among the respective raw data. This notion is evident from the formula for the correlation between two standardized-mean-difference effect sizes:

$$\rho_{\delta_1,\delta_2} = \frac{cov(\delta_1, \delta_2)}{\sigma_{\delta_1}\sigma_{\delta_2}}, \tag{B1}$$

where ρ_{δ_1,δ_2} is the population correlation coefficient with respect to the δ_1 and δ_2 effect sizes, $cov(\delta_1, \delta_2)$ is the population covariance for the δ_1 and δ_2 effect sizes, and σ_{δ_1} and σ_{δ_2} are the population standard deviations for the δ_1 and δ_2 effect sizes, respectively. Expanding Equation B1, we are able to analyze exactly how the correlation between effect sizes is a function of the correlation among the raw data:

$$\rho_{\delta_1,\delta_2} = \frac{\left(\frac{1}{n^{\mathrm{T}}} + \frac{1}{n^{\mathrm{C}}}\right)\rho_{Y_1,Y_2} + \left(\frac{\delta_1\delta_2}{2(n^{\mathrm{T}} + n^{\mathrm{C}})}\right)\rho_{Y_1,Y_2}^2}{\sqrt{\left(\frac{n^{\mathrm{T}} + n^{\mathrm{C}}}{n^{\mathrm{T}}n^{\mathrm{C}}} - \frac{\delta_1^2}{2(n^{\mathrm{T}} + n^{\mathrm{C}})}\right)\left(\frac{n^{\mathrm{T}} + n^{\mathrm{C}}}{n^{\mathrm{T}}n^{\mathrm{C}}} - \frac{\delta_2^2}{2(n^{\mathrm{T}} + n^{\mathrm{C}})}\right)}}, \tag{B2}$$

where all terms are as previously defined. Equation B2 is representative of the correlation of a single pair of effect sizes from one study in a meta-analysis.

The purpose of this appendix is to briefly examine the precise nature of the effect-size correlation that was used in this thesis. More specifically, I examine the relationship between raw-data correlations and effect-size correlations from situations applicable to my research. Only conditions where treatment and control groups were equal in size ($n^{\mathrm{T}} = n^{\mathrm{C}} = n$) were considered, therefore the above formula is reduced to the following form:

$$\rho_{\delta_1,\delta_2} = \frac{\left(\frac{2}{n}\right)\rho_{Y_1,Y_2} + \left(\frac{\delta_1\delta_2}{4n}\right)\rho_{Y_1,Y_2}^2}{\sqrt{\left(\frac{2}{n} - \frac{\delta_1^2}{4n}\right)\left(\frac{2}{n} - \frac{\delta_2^2}{4n}\right)}}. \tag{B3}$$

This thesis also restricted the type situations such that effect-size estimates within a single study were generated to be equivalent ($\delta_1 = \delta_2 = \delta$), which produces the revised form:

$$\rho_{\delta_1,\delta_2} = \frac{\left(\frac{2}{n}\right)\rho_{Y_1,Y_2} + \left(\frac{\delta^2}{4n}\right)\rho_{Y_1,Y_2}^2}{\frac{2}{n} - \frac{\delta^2}{4n}}. \quad \text{(B4)}$$

The above assumptions are fairly strong and may not be entirely realistic compared to a typical meta-analytic setting. However, assuming effect-size estimates are equal is comparable to the assumption that effect-size estimates are approximately equal, which is a typical observation in meta-analysis. For example, it might be unexpected for two effect sizes that measure mathematical constructs (e.g., algebra and number sense) to be drastically different.

While Gleser and Olkin (2009) discuss the origin of dependence among multiple endpoint effect sizes, they do not discuss the relationship between the raw-data correlation and the effect-size correlation. That is to say, there is a lack of literature regarding how the raw-data correlation that produces correlated effect sizes relates to the degree of correlation among said effect sizes. Using Equation B4, this relationship, as it relates to the specific meta-analytic setting in this thesis, is explored in Figure B.1.

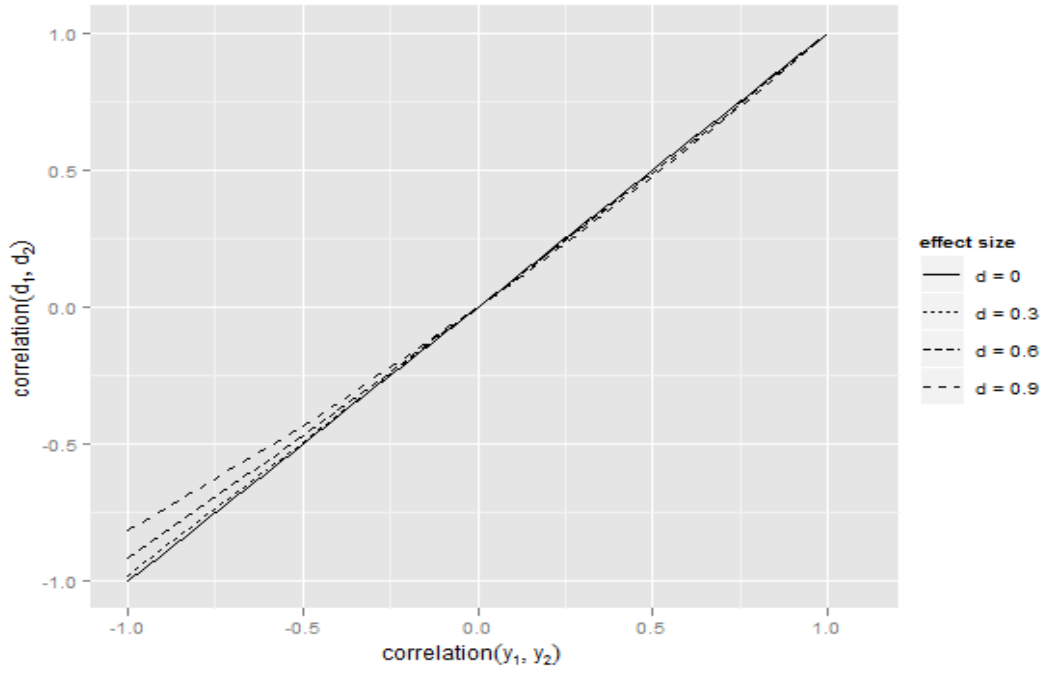

Figure B.1: Effect-Size Correlation Against Raw-Data Correlation.

Since ρ_{δ_1,δ_2} does not vary for values of n, study sample sizes were constrained to 100. Figure B.1 shows resulting values of effect-size correlations plotted against a spectrum of raw-data correlations while varying effect-size estimate magnitudes. It appears that the degree of correlation of raw data maps onto the degree of correlation between respective effect sizes almost exactly in most instances. That is to say, the degree of correlation among raw data that produce multiple endpoint type effect sizes will be the approximate degree of correlation between the associated effect sizes. The effect-size correlation appears to be slightly lower than the raw-data correlation with positive effect-sizes that are moderately large (e.g., raw-data correlation is .50), but the discrepancy is very small.

Conversely, when the raw-data correlation is extremely negative, results are more divergent. The lower left corner of Figure B.1 shows that for extreme cases of negatively correlated raw data, respective effect-size correlations for large effect-size estimates will be lower than the raw-data correlation. Simply assuming the degree of correlation between effect sizes is equal to the raw-data correlation will result in over-estimation. This scenario is far from what is typically seen in a meta-analysis. Situations where pairs of effect-size estimates have very strong negative correlations are not characteristic of the literature. However, one possible example could compare opposite emotional constructs: contentment and depression. This situation may possess a strong negative association.

Equation B4 restricts many parameters; a more complete version of this relationship remains unexamined. Positive raw-data correlation values (right half of graph) where $d = 0$ are applicable to this thesis, in which case the raw-data correlation and respective effect-size correlation form a near perfect linear relationship. This provides support for the use of ρ to represent both the between-outcomes and raw-data correlations.

APPENDIX C

SAS CODE FOR SIMULATION

```
options nodate nonumber;
proc iml;
do corrid=1 to 3;
      if corrid=1 then
            r={1 .5 0 0,
               .5 1 0 0,
               0 0 1 .5,
         0 0 .5 1};
      if corrid=2 then
            r={1 .75 0 0,
               .75 1 0 0,
               0 0 1 .75,
         0 0 .75 1};
      if corrid=3 then
            r={1 .99 0 0,
               .99 1 0 0,
               0 0 1 .99,
               0 0 .99 1};

do studyid=1 to 3;
      if studyid=1 then k=24;
      if studyid=2 then k=48;
      if studyid=3 then k=96;

*Cholesky decomposition;
chol=root(r);
trans=t(chol);

*Number of iterations per loop;
k2=k/2;
k3=(3*k)/4;
k4=(k/4);
nreps=3000;

*Initialize data matrix;
z=j(4,1,0);

do N=20 to 260 by 80;
      if N=20  then nseed=126;
      if N=100 then nseed=226;
      if N=180 then nseed=326;
      if N=260 then nseed=426;
      do irep=1 to nreps;
            do ik=1 to k;
                  do in=1 to N;
                        do irow=1 to 4;
                              z[irow,1]=rannor(nseed);
                        end; *End of (irow) loop;
                        y=trans*z;
                        if in=1 then yset=y`;
```

```
                if in>1 then yset=yset//y`;
            end; *End of (in) loop;

        *Loop for sample means and standard deviations;
        m=j(4,1,0);
        SD=j(4,1,0);
        do i= 1 to 4;
            m[i,1]=sum(yset[,i])/N;
            SD[i,1]=sqrt((ssq(yset[,i])-
            2*sum(yset[,i])*m[i,1]+N*m[i,1]**2)/(N-1));
            if i=1 then means=m`;
            if i>1 then means=means//m`;
            if i=1 then standevs=SD`;
            if i>1 then standevs=standevs//SD`;
        end; *End of means and standard deviations loop;

        means=means[4,];standevs=standevs[4,];

        *S1 pooled calculation;
        S1=sqrt(((standevs[1,1])**2 +(standevs[1,3])**2)/2);
        *S2 pooled calculation;
        S2=sqrt(((standevs[1,2])**2 +(standevs[1,4])**2)/2);
        *c calculation;
        c=1-(3/(8*N-9));
        *d1 calculation;
        d1=c*((means[1,1]-means[1,3])/S1);
        *d2 calculation;
        d2=c*((means[1,2]-means[1,4])/S2);

        *Treatment & control columns;
        d=d1||d2;
        if ik=1 then dstack=d;
        if ik>1 then dstack=dstack//d;

        *dvar calculations;
        dvart=(2/N)+((dstack[ik,1]*dstack[ik,1]`)/(4*N));
        dvarc=(2/N)+((dstack[ik,2]*dstack[ik,2]`)/(4*N));
        dvar=dvart||dvarc;
        if ik=1 then dvarstack=dvar;
        if ik>1 then dvarstack=dvarstack//dvar;
        end; *End of (ik) loop;

*Rearranging dependency structure (INDEPENDENT);
dA=dstack[1:k2,1];
dB=dstack[k2+1:k,1];
dC=dstack[1:k2,2];
dD=dstack[k2+1:k,2];
/*Rearranging dependency structure (MODERATE);
dA=dstack[1:k3,1];
dB=dstack[1:k4,2];
dC=dstack[k3+1:k,1];
dD=dstack[k4+1:k,2];*/
/*Rearranging dependency structure (VERY);
dA=dstack[1:k2,1];
dB=dstack[1:k2,2];
dC=dstack[k2+1:k,1];
dD=dstack[k2+1:k,2];*/
```

```
dnew1=dA//dB;
dnew2=dC//dD;
dfinal=dnew1||dnew2;

*Rearranging dependency structure (INDEPENDENT);
dAvar=dvarstack[1:k2,1];
dBvar=dvarstack[k2+1:k,1];
dCvar=dvarstack[1:k2,2];
dDvar=dvarstack[k2+1:k,2];
/*Rearranging dependency structure (MODERATE);
dAvar=dvarstack[1:k3,1];
dBvar=dvarstack[1:k4,2];
dCvar=dvarstack[k3+1:k,1];
dDvar=dvarstack[k4+1:k,2];*/
/*Rearranging dependency structure (VERY);
dAvar=dvarstack[1:k2,1];
dBvar=dvarstack[1:k2,2];
dCvar=dvarstack[k2+1:k,1];
dDvar=dvarstack[k2+1:k,2];*/

dvarnew1=dAvar//dBvar;
dvarnew2=dCvar//dDvar;
dvarfinal=dvarnew1||dvarnew2;

*Variance matrix transposition;
dvartrans=dvarfinal`;

*Weight calculation;
w=1/dvartrans;

*d-bar numerator matrix (2X2);
dbarnum=(w*dfinal);

*Weighted d-bars;
dbar1=dbarnum[1,1]/sum(w[1,]);
dbar2=dbarnum[2,2]/sum(w[2,]);

*Weighted d-bar matrix;
dbar1stack=j(k,1,dbar1);
dbar2stack=j(k,1,dbar2);

*Weight matrix diagonalization;
WSQ1=DIAG(w[1,]);
WSQ2=DIAG(w[2,]);

*Q-statistic;
Q1=(dfinal[,1]-dbar1stack)`*WSQ1*(dfinal[,1]-dbar1stack);
Q2=(dfinal[,2]-dbar2stack)`*WSQ2*(dfinal[,2]-dbar2stack);

*Birge's ratio;
B1=Q1/(k-1);
B2=Q2/(k-1);

*I^2;
I1=(Q1-(k-1))/Q1;
if I1<0 then I1=0;
```

```
      I2=(Q2-(k-1))/Q2;
      if I2<0 then I2=0;

      *Variance of weighted d-bars;
      VarMean1=1/sum(w[1,]);
      VarMean2=1/sum(w[2,]);

      *Stack Data;
      if irep=1 then dbar1all=dbar1;
      if irep>1 then dbar1all=dbar1all//dbar1;
      if irep=1 then dbar2all=dbar2;
      if irep>1 then dbar2all=dbar2all//dbar2;

      if irep=1 then varmean1all=varmean1;
      if irep>1 then varmean1all=varmean1all//varmean1;
      if irep=1 then varmean2all=varmean2;
      if irep>1 then varmean2all=varmean2all//varmean2;

      if irep=1 then Q1all=Q1;
      if irep>1 then Q1all=Q1all//Q1;
      if irep=1 then Q2all=Q2;
      if irep>1 then Q2all=Q2all//Q2;

      if irep=1 then Birge1all=B1;
      if irep>1 then Birge1all=Birge1all//B1;
      if irep=1 then Birge2all=B2;
      if irep>1 then Birge2all=Birge2all//B2;

      if irep=1 then I1all=I1;
      if irep>1 then I1all=I1all//I1;
      if irep=1 then I2all=I2;
      if irep>1 then I2all=I2all//I2;
      end; *End of (irep) loop;

*Correlation and N indicator vectors;
kvec=j(nreps,1,k);
cholvec=j(nreps,1,corrid);
Nvec=j(nreps,1,N);

*Merged statistics;
stat=kvec||cholvec||Nvec||Birge1all||I1all||dbar1all||varmean1all;
if N=20 then allstat=stat;
if N>20 then allstat=allstat//stat;

*Merged Qs;
Qvec=kvec||cholvec||Nvec||Q1all;
if N=20 then Qvecs=Qvec;
if N>20 then Qvecs=Qvecs//Qvec;
end; *End of (N) loop;

if studyid=1 then Qvecss=Qvecs;
if studyid>1 then Qvecss=Qvecss//Qvecs;
if studyid=1 then allstats=allstat;
if studyid>1 then allstats=allstats//allstat;
end; *End of (studyid) loop;

*Add 'corrid' to mergerd data;
```

```
if corrid=1 then TOTALstat=allstats;
if corrid>1 then TOTALstat=TOTALstat//allstats;
if corrid=1 then TOTALQ=Qvecss;
if corrid>1 then TOTALQ=TOTALQ//Qvecss;
end; *End of (corrid) loop;

*Create PROC-ready datasets;
create TOTALstat from TOTALstat;append from TOTALstat;
create TOTALQ from TOTALQ;append from TOTALQ;
quit; *End of PROC IML;

*Label columns of the data appropriately;
data TOTALstat;set TOTALstat;
rename col1=K col2=Correlation col3=N col4=BirgesRatio col5=I_squared
col6=dbar col7=dbarvar;
run;quit;
data TOTALQ;set TOTALQ;
rename col1=K col2=Correlation col3=N col4=Q;
run;quit;
%macro changecorr(dataset);
data &dataset;set &dataset;
if Correlation=1 then Correlation=.50;
if Correlation=2 then Correlation=.75;
if Correlation=3 then Correlation=.99;
run;
%mend changecorr;
%changecorr(TOTALstat);
%changecorr(TOTALQ);

*Descriptive Statistics;
ods pdf file="[Dependency Structure Name] (Descriptives).pdf" style=journal;
proc means data=TOTALstat mean std min max;
  class Correlation K N;
  var BirgesRatio I_squared dbar dbarvar;
run; ods pdf close;
ods pdf file="[Dependency Structure Name] (Qs).pdf" style=journal;
proc means data=TOTALQ mean std min max;
  class Correlation K N;
  var Q;
run; ods pdf close;

*Format Q ranges;
proc format;
value Qrangea
0-35.17246 = '0-35.17'
35.17247-9999 = '> 35.17'; *35.17 is critical value for alpha=.05 of chi-
square df=23;
value Qrangeb
0-64.00111 = '0-64.00'
64.00112-9999 = '> 64.00'; *64.00 is critical value for alpha=.05 of chi-
square df=47;
value Qrangec
0-118.75161 = '0-118.75'
118.75162-9999 = '> 118.75'; *118.75 is critical value for alpha=.05 of chi-
square df=95;
run;
```

```
*Output Q statistic frequencies;
%macro freq(Qform, k);
ods pdf file="[Dependency Structure Name] (Q Rejection Rates).pdf"
style=journal;
proc freq data=TOTALQ order=freq formchar(1,2,7)='|-+';
  tables Q;format Q &Qform;
  by Correlation K N;where K=&k;
 run;
%mend freq;
%freq(Qrangea.,24);
%freq(Qrangeb.,48);
%freq(Qrangec.,96);
ods pdf close;
```

APPENDIX D

SAS CODE FOR MODEL ESTIMATION

```
options nodate;
proc iml;
*Number of meta-analyses;
l=3000;

do metaid=1 to l;
      do corrid=1 to 3;
            if corrid=1 then
                  r={1 .5 0 0,
                     .5 1 0 0,
                     0 0 1 .5,
                     0 0 .5 1};
            if corrid=2 then
                  r={1 .75 0 0,
                     .75 1 0 0,
                  0 0 1 .75,
                  0 0 .75 1};
            if corrid=3 then
                  r={1 .99 0 0,
                  .99 1 0 0,
                  0 0 1 .99,
                  0 0 .99 1};
            do corrtype=1 to 2;
                  do studyid=1 to 3;
                        if studyid=1 then k=24;
                        if studyid=2 then k=48;
                        if studyid=3 then k=96;

*Initialize matrices;
z=j(4,1,0);
x={1 0,0 1};
df=2*(k-1);

*Cholesky decomposition;
chol=root(r);
trans=t(chol);

do N=20 to 260 by 80;
      if N=20  then nseed=13;
      if N=100 then nseed=23;
      if N=180 then nseed=33;
      if N=260 then nseed=43;
            do ik=1 to k;
                  do in=1 to N;
                        do irow=1 to 4;
                              z[irow,1]=rannor(nseed);
                        end; *End of (irow) loop;
                        y=trans*z;
                        if in=1 then yset=y`;
                        if in>1 then yset=yset//y`;
```

```
      end; *End of (in) loop;

*Stack observations within outcome group (T over C);
y1=yset[,1]//yset[,3];
y2=yset[,2]//yset[,4];
yall=y1||y2;

*Loop for calculating mean and standard deviation;
m=j(4,1,0);
SD=j(4,1,0);
do i= 1 to 4;
      m[i,1]=sum(yset[,i])/N;
      SD[i,1]=sqrt((ssq(yset[,i])-
      2*sum(yset[,i])*m[i,1]+N*m[i,1]**2)/(N-1));
      if i=1 then means=m`;
      if i>1 then means=means//m`;
      if i=1 then standevs=SD`;
      if i>1 then standevs=standevs//SD`;
end; *End of means and standard deviations loop;

means=means[4,];standevs=standevs[4,];

*S1 pooled calculation;
S1=sqrt(((standevs[1,1])**2 +(standevs[1,3])**2)/2);
*S2 pooled calculation;
S2=sqrt(((standevs[1,2])**2 +(standevs[1,4])**2)/2);
*c calculation;
c=1-(3/(8*N-9));
*d1 calculation;
d1=c*((means[1,1]-means[1,3])/S1);
*d2 calculation;
d2=c*((means[1,2]-means[1,4])/S2);

*Treatment column & control column;
d=d1||d2;
if ik=1 then dstack=d;
if ik>1 then dstack=dstack//d;

*Variance calculations;
dvart=(2/N)+((dstack[ik,1]*dstack[ik,1]`)/(4*N));
dvarc=(2/N)+((dstack[ik,2]*dstack[ik,2]`)/(4*N));
dvar=dvart||dvarc;

if ik=1 then dvarstack=dvar;
if ik>1 then dvarstack=dvarstack//dvar;

*Design Matrix;
if ik=1 then design=x;
if ik>1 then design=design//x;

*Alternating treatment/control in one column;
f=d1//d2;
if ik=1 then dall=f;
if ik>1 then dall=dall//f;

*Correlation for Y's in study loop;
nobs=nrow(yall);
```

```
            sum=yall[+,];
            xpx=t(yall)*yall-t(sum)*sum/nobs;
            s=diag(1/sqrt(vecdiag(xpx)));
            corrmatrix=s*xpx*s;

            if corrtype=1 then corr=corrmatrix[1,2]; *Empirical;
            if corrtype=2 then corr=r[1,2]; *Hypothetical;

            *Covariance of d's per study;
            cov=(2/N)*corr+((dstack[ik,1]*dstack[ik,2])/(4*N))*corr**2;

            *Covariance 'matrix' for study;
            sigma1=dvart||cov;
            sigma2=cov||dvarc;
            sigma=sigma1//sigma2;

            *Create looped covariance matrix;
            if ik=1 then COVALL=sigma;
            else COVALL=block(covall,sigma);
      end; *End of study (ik) loop;

*Inverse covariance and design transpose matrices;
invcovall=inv(covall);
designtrans=design`;

*Design transpose x inverse covariance x design;
leftWLSd=designtrans*invcovall*design;

*Estimated covariance matrix;
invleftWLSd=inv(leftWLSd);

*Weighted Least Squares Estimates of Effect Sizes;
WLSd=invleftWLSd*designtrans*invcovall*dall;

*ES matrix - product of design matrix and WLS effect sizes;
sideWLSQ=dall-(design*WLSd);

*Weighted Least Squares Estimates (Q);
WLSQ=sideWLSQ`*invcovall*sideWLSQ;

*Birge's ratio;
B=WLSQ/(df);

*I squared;
Isq=100*((WLSQ-(df))/WLSQ);
if Isq<0 then Isq=0;

*Merge dataset;
allstat=corrtype||corrid||k||N||WLSQ||df||B||Isq;
if N=20 then allstats=allstat;
if N>20 then allstats=allstats//allstat;
end; *End of (N) loop;

if studyid=1 then STUDYstat=allstats;
if studyid>1 then STUDYstat=STUDYstat//allstats;
end; *End of (studyid) loop;
```

```
if corrtype=1 then CORRTYPEstat=STUDYstat;
if corrtype>1 then CORRTYPEstat=CORRTYPEstat//STUDYstat;
end; *End of (corrtype);

if corrid=1 then CORRIDstat=CORRTYPEstat;
if corrid>1 then CORRIDstat=CORRIDstat//CORRTYPEstat;
end; *End of (corrid) loop;

if metaid=1 then FINALstat=CORRIDstat;
if metaid>1 then FINALstat=FINALstat//CORRIDstat;
end; *End of (metaid) loop;

*Create PROC-ready datasets;
create alldata from FINALSTAT;
append from FINALSTAT;
quit; *End of PROC IML;

*Data Management and Output;
data alldata;set alldata;
rename col1=CorrelationType col2=Correlation col3=K col4=N
col5=Q col6=df col7=BirgeRatio col8=I_squared;
run;quit;
data alldata;set alldata;
if Correlation=1 then Correlation=.50;
if Correlation=2 then Correlation=.75;
if Correlation=3 then Correlation=.99;
run;quit;
proc sort data=alldata;
by CorrelationType Correlation K N;
run;quit;

*Homogeneity Descriptive Statistics;
ods pdf file="Q Descriptives.pdf" style=journal;
proc means data=alldata mean std min max;
  class CorrelationType Correlation K N;
  var Q I_squared BirgeRatio;
run; ods pdf close;

*Format Q ranges;
proc format;
value Qrangea
0-62.82962 = '0-62.83'
62.82963-9999 = '> 62.83'; *62.83 is critical value for alpha=.05 of chi-
square df=2(24-1)=46;
value Qrangeb
0-117.63165 = '0-117.63'
117.63166-9999 = '> 117.63'; *117.63 is critical value for alpha=.05 of chi-
square df=2(48-1)=94;
value Qrangec
0-223.16025 = '0-223.16'
223.16026-9999 = '> 223.16'; *223.16 is critical value for alpha=.05 of chi-
square df=2(96-1)=190;
run;

*Output Q statistic frequencies;
%macro freq(Qform, k);
ods pdf file="Q Rejection Rates.pdf" style=journal;
```

```
proc freq data=alldata order=freq formchar(1,2,7)='|-+';
  tables Q;format Q &Qform;
  by CorrelationType Correlation K N;where K=&k;
 run;
%mend freq;
%freq(Qrangea.,24);
%freq(Qrangeb.,48);
%freq(Qrangec.,96);
ods pdf close;
```

REFERENCES

Barrett-Connor, E., Grady, D., Sashegyi, A., Anderson, P. W., Cox, D. A., Hoszowski, K.,…MORE, I. (2002). Raloxifene and cardiovascular events in osteoporotic postmenopausal women: Four-year results from the MORE (Multiple Outcomes of Raloxifene Evaluation) randomized trial. *Journal of the American Medical Association*, *287*(7), 847-857.

Becker, B. J. (1989). Gender and science achievement. A reanalysis of studies from two meta-analyses. *Journal of Research in Science Teaching*, *26*(2), 141-169.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research*, *60*, 373-417.

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

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.

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review*, *40*(2), 207-227.

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*(4), 406-418.

Borenstein, M. (2009). Effect sizes for continuous data. In H. M. Cooper, L.V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. (pp. 221-235). New York: Russell Sage Foundation.

Caird, J. K., Willness, C. R., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effect of cell phones on driver performance. *Accident Analysis and Prevention*, *40*, 1282-1293.

Chae, J., Yu, D. T., Walker, M. E.,…Fang, Z. P. (2005). Intramuscular electrical stimulation for hemiplegic shoulder pain: A 12-month follow-up of a multiple-center, randomized clinical trial. *American Journal of Physical Medicine & Rehabilitation*, *84*(11), 832-842.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101-129.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145-153.

Galbraith, R. F. (1988). Graphical display of estimates having differing standard errors. *Technometrics*, *30*(3), 271-281.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Gleser, J. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. M. Cooper, L.V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. (pp. 357-376). New York: Russell Sage Foundation.

Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. M. Cooper, L.V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. (pp. 417-433). New York: Russell Sage Foundation.

Gross, J. (2003). *Linear regression*. Berlin, Germany: Springer.

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

Hartung, H.-P., Gonsette, R., König, N., Kwiecinski, H., Gueso, A., Morrissey, S. P., ...Mitoxantrone in Multiple Sclerosis Study Group (MIMS). (2002). Mitoxantrone in progressive multiple sclerosis: A placebo, controlled, double-blind, randomised, multicentre trial. *Lancet*, *360*(Dec. 21/28), 2018-2025.

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107-128.

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

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

Hedges, L. V. (2007). Meta-analysis. In Rao, C. R. & Sinharay, S. (Eds.), *Handbook of statistics*. (Vol. 26, pp. 919-953). The Netherlands: Elsevier.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39-65.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine*, *21*, 1539-1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal*, *327*, 557-560.

Huberty, C. J. (2002). A history of effect size indices. *Educational Psychology and Measurement*, *62*(2), 227-240.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistics or I^2 index? *Psychological Methods*, *11*(2), 193-206.

Huntley, M. A., Rasmussen, C. L., Villarubi, R. S., Sangtong, J., & Fey, J. T. (2000). Effects of standards-based mathematics education: A study of the Core-Plus Mathematics Project algebra and functions strand. *Journal for Research in Mathematics Education*, *31*(3), 328-361.

Ishak, K., Platt, R. W., Joseph, L., & Hanley, J. A. (2008). Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine*, *27*, 670-686.

Kim, R.-S., & Becker, B. J. (2010). The degree of dependence between multiple-treatment effect sizes. *Multivariate Behavioral Research*, *45*, 213-238.

Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. In H. M. Cooper, L.V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. (pp. 279-293). New York: Russell Sage Foundation.

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

Marin-Martinez, F., & Sanchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology*, *2*(1), 32-38.

Montori, V. M., Leung, T. W., Walter, S. D., & Guyatt, G. H. (2005). Procedures that assess inconsistency in meta-analysis can assess the likelihood of response bias in multiwave surveys. *Journal of Clinical Epidemiology*, *58*, 856-858.

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

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*, 111-120.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, *99*, 400-406.

Schott, J. R. (2005). *Matrix analysis for statistics*. Hoboken, N.J.: Wiley.

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

Timm, N. H. (1999). A note on testing for multivariate effect sizes. *Journal of Educational and Behavioral Statistics*, *24*(2), 132-145.

Walsh, J. E. (1947). Concerning the effect of the intraclass correlation on certain significance tests. *Annals of Mathematical Statistics*, *18*, 88-96.

BIOGRAPHICAL SKETCH

Christopher Thompson

Christopher Thompson is originally from Davie, Florida and began his academic studies at Florida State University in 2005. He completed his Bachelor's degree in Secondary Mathematics Education with minors in Statistics and Mathematics in 2009. He enrolled in the Measurement and Statistics Master's program in 2010. His current research interests include meta-analysis methodology and applications, structural equation modeling, and general statistical applications in educational policy and testing.

Christopher is a classical music enthusiast and enjoys foreign and experimental cinema.