2011

Meta-Analysis of Single-Case Data: A Monte Carlo Investigation of a Three Level Model

Corina M. Owens
University of South Florida, cmowens@mail.usf.edu

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the American Studies Commons, Educational Assessment, Evaluation, and Research Commons, and the Statistics and Probability Commons

Meta-Analysis of Single-Case Data: A Monte Carlo Investigation of a Three Level Model

by

Corina M. Owens

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Educational Measurement and Research
College of Education
University of South Florida

Major Professor: John M. Ferron, Ph.D.
Robert F. Dedrick, Ph.D.
Liliana Rodriguez-Campos, Ph.D.
Julia Ogg, Ph.D.

Date of Approval: March 28, 2011

Keywords: single-subject, research synthesis, multilevel modeling, hierarchical linear modeling, simulation

DEDICATION

I dedicate this dissertation to my grandparents whose belief in me could quite possibly have exceeded my belief in myself. Your unwavering love and support motivated me to work hard and succeed. To my Oma and Opa, may you rest in peace and as you look down on me I hope I have made you proud. To my Grandma and Grandfather, your constant encouragement has helped propel me through my studies.

I want to thank my parents, Karen and John, for their unconditional love and support. I am forever grateful to you both for pushing me beyond what I thought was possible. I would also like to thank my partner, Ashley, who came in during that final stretch to stand beside me and remind me anything is possible. Thank you all for always believing in me!

ACKNOWDLEGEMENTS

TABLE OF CONTENTS

iii

LIST OF TABLES

LIST OF FIGURES

x

ABSTRACT

Numerous ways to meta-analyze single-case data have been proposed in the literature, however, consensus on the most appropriate method has not been reached. One method that has been proposed involves multilevel modeling. This study used Monte Carlo methods to examine the appropriateness of Van den Noortgate and Onghena's (2008) raw data multilevel modeling approach to the meta-analysis of single-case data. Specifically, the study examined the fixed effects (i.e., the overall average baseline level and the overall average treatment effect) and the variance components (e.g., the between person within study variance in the average baseline level, the between study variance in the overall average baseline level, the between person within study variance in the average treatment effect) in a three level multilevel model (repeated observations nested within individuals nested within studies). More specifically, bias of point estimates, confidence interval coverage rates, and interval widths were examined as a function of specific design and data factors. Factors investigated included (a) number of primary studies per meta-analysis, (b) modal number of participants per primary study, (c) modal series length per primary study, (d) level of autocorrelation, and (3) variances of the error terms. The results of this study suggest that the degree to which the findings of this study are supportive of using Van den Noortgate and Onghena's (2008) raw data multilevel modeling approach to meta-analyzing single-case data depends on the particular effect of interest. Estimates of the fixed effects tended to be unbiased and produced confidence

intervals that tended to overcover but came close to the nominal level as level-3 sample size increased. Conversely, estimates of the variance components tended to be biased and the confidence intervals for those estimates were inaccurate.

CHAPTER ONE: INTRODUCTION

Single-case research has grown in popularity over the past decade and is being

conducted in a variety of settings such as school psychology (Skinner, 2004), special

education (Algozzine, Browder, & Karvonen, 2001), teacher education (Hsieh,

Hemmeter, McCollum, & Ostrosky, 2009), and behavioral intervention research (Filter &

Horner, 2009). This type of research allows for the repeated measurement of one case

over a certain period of time to assess a treatment's effect on an individual case.

Typically, data are collected during a baseline phase (prior to treatment) and then during

or after the implementation of the treatment or intervention. This is the most basic design;

additional design types include the removal of the intervention, reintroduction of the

intervention, and maintenance of the intervention. In addition, several cases or settings

can be studied at the same time in a multiple baseline design.

Across single-case studies there have been numerous ways to analyze this type of

data, such as visual analysis, computing descriptive summaries, randomization tests,

regression analysis, and multilevel modeling. In addition to a variety of analysis options,

a variety of effect size estimates have been proposed for use in single-case research, such

as percentage of non-overlapping data (PND, Scruggs, Mastropieri, & Castro, 1987), a

form of standardized mean difference (Busk & Serlin, 1992), change in R^2 values

(Center, Skiba, & Casey, 1985-1986; Kromrey & Foster-Johnson, 1996; Beretvas &

Chung, 2008a), and the use of standardized regression coefficients in multilevel models (Van den Noortgate & Onghena, 2003a, 2007, 2008).

Quantitative integration of study results, termed meta-analysis, involves the combining of data across multiple studies to evaluate and summarize research findings. The term meta-analysis was first coined by Glass (1976) and was defined as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p.3). Meta-analysis has been used to synthesize results from a wide variety of studies, both non-experimental (e.g., gender differences) and experimental (e.g., intervention effectiveness). This type of research is necessary to determine relationships among variables and the effectiveness of interventions across studies. It also allows researchers to integrate study findings with the goal of generalization. Quantitative integration of study findings should cross research domains and include all types of quantitative research, including single-case research.

Meta-analysis of single-case research has resulted in much disagreement in the field. In a study synthesizing single-case meta-analyses conducted between 1985 and 2005, the majority of meta-analyses were simply reporting mean effect sizes across studies (Beretvas & Chung, 2008b). However, another possible option for combining effect sizes across studies is the use of multilevel modeling. Multilevel modeling has been proposed for use with single-case data by many researchers because of its flexibility in handling nesting of observations within people (Nugent, 1996; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003b). One specific example is Van den Noortgate and Onghena's (2008) application of multilevel modeling to the meta-analysis of single-case data. Their study proposed the use of a multilevel model to meta-analyze

single-case data. Equations 1-5 represent their proposed individual level raw data model. Equation 1 represents an outcome (*y*) that is modeled on measurement occasion *i* for participant *j* of study *k* (y_{ijk}) as a linear function of a single-predictor, *phase:*

$$y_{ijk} = \pi_{0jk} + \pi_{1jk} phase + e_{ijk} \tag{1}$$

where *phase* is a dichotomous variable indicating whether a measurement occasion or observation occurred during baseline or treatment phase. π_{0jk} is the level of the outcome during baseline for participant *j* from study *k*, π_{1jk} is the treatment effect for participant *j* from study *k*, and e_{ijk} is within-phase error (σ^2 represents the variance of e_{ijk}).

At the second level, the variation across participants is modeled in the following equations:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \tag{2}$$

and

$$\pi_{1jk} = \beta_{10k} + r_{1jk} \tag{3}$$

where the fixed effects are β_{00k}, the average baseline level for study *k*, and β_{10k}, the average treatment effect for study *k*, and the error terms are r_{0jk} and r_{1jk}, allowing variation in both baseline levels and treatment effects among participants ($\tau_{\pi00}$ represents the variance of r_{0jk} and $\tau_{\pi10}$ represents the variance of r_{1jk}).

At the third level, the variation across studies is modeled in the following equations:

$$\beta_{00k} = \gamma_{000} + u_{00k} \tag{4}$$

and

$$\beta_{10k} = \gamma_{100} + u_{10k} \tag{5}$$

3

where the fixed effects are γ_{000}, the overall average baseline level, and γ_{100}, the overall average treatment effect, and the error terms are u_{00k} and u_{10k}, which allow variation in both the average baseline levels and average treatment effects among studies ($\tau_{\beta00}$ represents the variance of u_{00k} and $\tau_{\beta10}$ represents the variance of u_{10k}). It should be noted that in multilevel modeling analysis, errors on all levels are typically assumed to be normally distributed and have a mean of zero.

Problem Statement

Although the use of single-case designs has grown over the past decades, the majority of literature on meta-analysis focuses on group comparison studies and leaves out single-case research (Van den Noortgate & Onghena, 2008). This lack of literature related to single-case designs is often why these designs are excluded from meta-analyses. This exclusion of single-case designs is concerning when one considers the plethora of information single-case research can add to the literature. Single-case designs not only provide information related to average treatment effects but also offer information related to how that treatment effect is related to specific cases. Meta-analyses of single-case designs offer the ability to summarize and evaluate the overall effect without the loss of that specific case information. In addition, the meta-analysis of single-case data increases the generalizabilty of research findings.

Researchers have proposed a variety of methods to meta-analyze single-case data. Van den Noortgate and Onghena's (2008) proposed method of using multilevel modeling to meta-analyze single-case data offers many advantages. The use of multilevel modeling provides the flexibility of appropriately modeling the autocorrelational nature of single-case data, can take into consideration multiple effect sizes per study, and can

4

apply appropriate meta-analytic models, such as fixed or random effects models.

Although the use of multilevel modeling offers advantages in the analysis of single-case

data, there are still concerns as to whether the use of multilevel modeling is appropriate

for single-case data. Specifically, multilevel modeling is based on large sample theory,

which is not representative of single-case data. Therefore, it is necessary to further

investigate the utility of inferences made from multilevel modeling when applied to

single-case data.

Purpose of the Study

The purpose of this study was to examine the appropriateness of Van den

Noortgate and Onghena's (2008) raw data multilevel modeling approach to the meta-

analysis of single-case data. Specifically, the study examined the fixed effects (i.e., the

overall average baseline level and the overall average treatment effect) and the variance

components (e.g., the between person within study variance in the average baseline level,

the between study variance in the overall average baseline level, the between person

within study variance in the average treatment effect) in a three level multilevel model.

More specifically, bias of point estimates, confidence interval coverage rates, and interval

widths were examined as a function of specific design and data factors. The following

research questions are of interest:

Research Questions

1. To what extent are the fixed effect estimates from a three level meta-analytic

    single-case model biased as a function of specific design factors (number of

    primary studies per meta-analysis, modal number of participants per primary

study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

2. To what extent does the confidence interval coverage of the fixed effect estimates from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

3. To what extent does the confidence interval width of the fixed effect estimates from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

4. To what extent are the variance components from a three level meta-analytic single-case model biased as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

5. To what extent does the confidence interval coverage of the variance components from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

6. To what extent does the confidence interval width of the variance components
    from a three level meta-analytic single-case model vary as a function of specific
    design factors (number of primary studies per meta-analysis, modal number of
    participants per primary study, modal series length per primary study, level of
    autocorrelation, and variances of the error terms)?

Overview of Study

Monte Carlo simulation methods were used to examine the appropriateness of the
multilevel modeling inferences. The use of simulation methods allowed for the control
and manipulation of specific design and data factors. The Monte Carlo study included
five factors in the design (see Table 1). These factors were (a) number of primary studies
per meta-analysis (10, 30, and 80); (b) modal number of participants per primary study
(small [mode = 4] and large [mode = 8]); (c) modal series length per primary study (small
[mode = 10], medium [mode = 20], and large [mode = 30]); (d) level of autocorrelation
(0, .2, and .4); and (e) variances of the error terms (most of the variance at level-1 [$\sigma^2 =$
1, $\tau_{\pi00} = \tau_{\pi10} = .2$, and $\tau_{\beta00} = \tau_{\beta10} = .05$] and most of the variance at level-2 [$\sigma^2 = 1$,
$\tau_{\pi00} = \tau_{\pi10} = 2$, and $\tau_{\beta00} = \tau_{\beta10} = .5$]). The appropriateness of the inferences made from
the estimates was evaluated in terms of coverage and width of 95% confidence intervals
as well as bias of point estimates.

Table 1
Study Design

Number of Primary Studies per Meta-Analysis	Modal number of Participants per Primary Study	Modal Series Length per Primary Study	Error Variances					
			level-1 = 1; level-2 = .2; level-3 = .05			level-1 = 1; level-2 = 2; level-3 = .5		
			Level of Autocorrelation					
			0	.2	.4	0	.2	.4
10	Small (mode = 4)	Small (mode = 10)						
		Medium (mode = 20)						
		Large (mode = 30)						
	Large (mode = 8)	Small (mode = 10)						
		Medium (mode = 20)						
		Large (mode = 30)						
30	Small (mode = 4)	Small (mode = 10)						
		Medium (mode = 20)						
		Large (mode = 30)						
	Large (mode = 8)	Small (mode = 10)						
		Medium (mode = 20)						
		Large (mode = 30)						
80	Small (mode = 4)	Small (mode = 10)						
		Medium (mode = 20)						
		Large (mode = 30)						
	Large (mode = 8)	Small (mode = 10)						
		Medium (mode = 20)						
		Large (mode = 30)						

Limitations

The data in this study were simulated based on specific conditions. Those conditions were chosen based on a review of single-case literature and meta-analyses of single-case data. The specific conditions chosen for this study are only some of the possible options. Therefore, the results of this study can only be generalized to studies with similar conditions. Any conclusions beyond the observed conditions should be interpreted with caution.

Definitions of Terms

Autocorrelation. The degree to which errors from repeated observations are correlated with each other.

Bias. The average difference between a known parameter estimate and an estimated parameter estimate.

Confidence interval coverage. The proportion of 95% confidence intervals that contain the estimated parameter.

Confidence interval width. The average difference between the upper and lower limits of the 95% confidence intervals for the estimated parameter.

Effect size. A measure of the magnitude of the relationship between two variables.

Fixed effects. Parameter estimates of the coefficients represented in the multilevel model (e.g., overall average baseline level, overall average treatment effect).

Kenward-Roger degrees of freedom method. A method for estimating degrees of freedom that approximates the degrees of freedom and was developed to be used with unbalanced designs and complex covariance structures. This method is an extension of

the Satterthwaite method; it adjusts for small-sample size bias in the estimation of variances.

Meta-analysis. The quantitative integration of study results that involves the combining of effects sizes across multiple studies to evaluate and summarize research findings.

Multilevel modeling. A statistical model used to account for hierarchical or nested data, also known as hierarchical linear modeling. "A hierarchical linear model consists of one or more regression equations at each level in which the characteristics of the units from that level are used as predictors in describing the coefficients of the equation(s) of the level just below" (Van den Noortgate & Onghena, 2003a, p. 329).

Primary studies. The original studies that comprise the sample for the meta-analysis.

Satterthwaite degrees of freedom method. A method that approximates the degrees of freedom, and was developed to be used with unbalanced designs and complex covariance structures.

Series length. The level-1 sample size in the multilevel model, or the number of times a participant is observed.

Single-case research. The study of a single participant or a group (e.g., a classroom), measured at multiple points in time to determine the effectiveness of one or more interventions or treatments.

Treatment effect. The change in a dependent variable that is attributable to a specific treatment.

Variance components. Parameters that estimate variation within person, between persons within studies, and between studies.

CHAPTER TWO: LITERATURE REVIEW

This literature review will be divided into three parts. First, single-case research is described. Second, a brief overview of meta-analysis is described and finally, the meta-analysis of single-case research will be discussed.

Single-Case Research

Single-case research, like case studies, can be defined as the study of a single participant or group (e.g., a classroom). However, unlike case study research which gathers in depth narrative or anecdotal information on a single case, single-case research systematically measures a single case at multiple points in time to determine the effectiveness of one or more interventions or treatments (Kazdin, 2011). Single case research designs have taken on a variety of different names, such as single-case, single-subject, N=1, and intra-subject. Regardless of the name identified by the researcher, the focus of this type of research is on the single case and its growth over time. This type of research allows the researcher to focus on individual variations in the treatment effect, which have a tendency to be lost in group comparison designs where the focus is the average treatment effect (Barlow, Nock, & Hersen, 2009). In addition to individual variation, this type of design also allows the individual to be measured at various points in time, thereby allowing the treatment effect to be evaluated with more than a single observation, which allows researchers to see how the treatment effect will change over time. Single-case research also allows practitioners to implement research in their own

setting, therefore reducing the gap between research and practice (Morgan & Morgan, 2001). Finally, due to the fact that only a small sample size is needed, researchers are able to study populations of people that have a low prevalence rate (e.g., children with autism, the homeless) (Van den Noortgate & Onghena, 2003a).

Methodological Issues

Single-case research offers many advantages to researchers. However, as with any type of research, with those advantages come certain methodological concerns. One such concern comes in the form of generality of findings. This concern stems from the fact that when studying a single case it is difficult to know if results from that particular case will be applicable to other cases (Barlow, Nock, & Hersen, 2009). Although generality of findings can be a concern in single-case research specific replication strategies can be implemented to improve generalizations.

Another important methodological concern centers on a key feature of single-case research, repeated measurement. Barlow, Nock, and Hersen (2009) suggest that repeated measurements need to be "specific, observable, and replicable" (p. 62). A repeated measurement is (a) specific when it is obvious that a behavior has or has not occurred; (b) observable when multiple observers can measure it without difficulty; and (c) replicable when the methods used to observe the behaviors can be duplicated on several occasions (Barlow, Nock, & Hersen, 2009). In addition, it is important to take into consideration the frequency of measurements. Specifically, one should balance the importance of having enough data with which to evaluate change with the importance of not causing fatigue on the part of the subject.

Another issue to consider when using repeated measurements is the use of self-report data. Often in research it is necessary to measure a participant's perceptions of a particular behavior (e.g. feelings of depression, anxiety, or happiness, level of control over life choices), however, attempting to measure self-report data is not without limitations. One possible limitation is the role social desirability (Crowne & Marlowe, 1960) can play on self-report data. It is possible that a true behavior change is not occurring and instead the participant is reporting what they think is socially desirable. Single-case researchers should be aware of these methodological issues and design their studies to minimize these concerns.

Design Types

All studies are based on specific types of research designs and within a single-case framework there are multiple research designs that can be implemented. Baseline logic is a set of guidelines that can be used to organize the experimental design process (Riley-Tillman & Burns, 2009). Baseline logic is comprised of four steps, (1) prediction, (2) affirmation of the consequent, (3) verification, and (4) replication by affirmation of the consequent (Riley- Tillman & Burns, 2009). The first step, prediction, is used to determine what the behavior looks like prior to the intervention and is typically termed baseline or A phase. This stage is necessary to illustrate what level the behavior is occurring at and how stable and/or variable the behavior is prior to the intervention (Riley-Tillman & Burns, 2009). By examining these things researchers are able to predict what the behavior would look like if no intervention were implemented. The second step of baseline logic, affirmation of the consequent, allows the researcher to first test whether the intervention had some impact on the participant's behavior and is

14

typically termed treatment or B phase. In this phase the intervention has been

implemented and the behavior is being measured to determine if there is a predictable

change in the data (Riley-Tillman & Burns, 2009). Thus far in the steps of baseline logic

the most basic single-case design type has been described, an AB, or interrupted time

series, design (see Figure 1). This type of design consists of observations of the

dependent variable both before and after an intervention. The observations that occur

before an intervention are considered part of the baseline (A) phase, and the observations

that occur after the intervention are considered part of the treatment (B) phase.
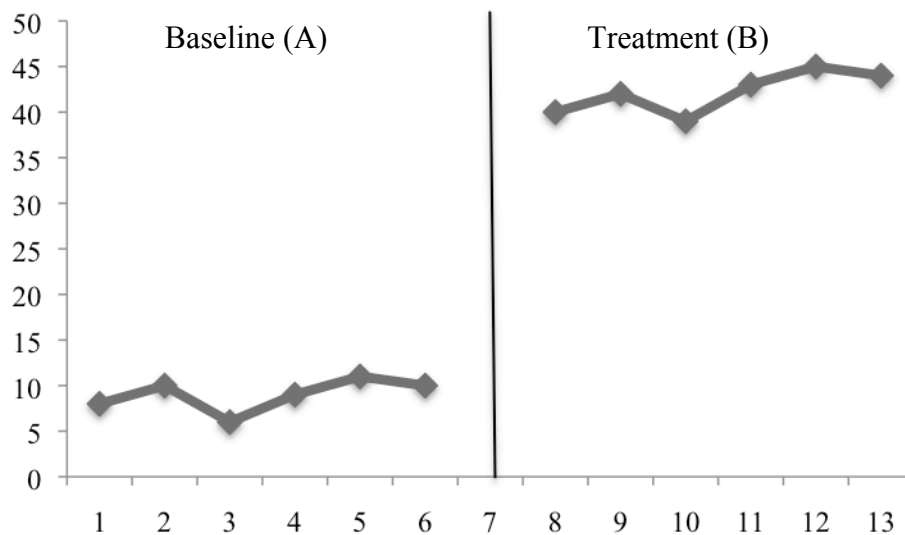
Figure 1. Graphical display of interrupted time series design

     This basic AB design type is not without criticism. For example, when using this

type of design it is difficult to attribute a change in the data to the treatment and not to

some other event which could have occurred at the same time. Another plausible

explanation for a shift in data could be developmental milestones or a change in

instrumentation (Ferron & Rendina-Gobioff, 2005). These limitations can be addressed

by utilizing more complex study designs, such as a reversal or multiple baseline design, which will also address the final two steps of baseline logic.

The third step of baseline logic is verification. This step is used to verify what was observed in the original baseline phase by removing the intervention and returning to a second baseline or A phase. This step allows one to gain increased confidence in what was originally seen in the first baseline as well as attributing the changes observed in the treatment phase to the introduction of the intervention rather than some extraneous variable (Riley-Tillman & Burns, 2009). This stage of baseline logic is clearly illustrated with the use of the most simplistic reversal design, an ABA design.

Reversal, or withdrawal, designs are extensions of the basic AB design. Although the terms reversal and withdrawal are often used interchangeably in the literature there is a slight distinction (Barlow, Nock, & Hersen, 2009). Reversal designs refer to situations when the intervention is reversed and applied to an incompatible behavior, whereas withdrawal designs refer to situations where the intervention is simply withdrawn and returned to the A phase (Barlow, Nock, & Hersen, 2009; Rusch& Kazdin, 1981). Nonetheless, the most simplistic reversal or withdrawal design is removal of the treatment from participants (ABA; see Figure 2). This design consists of observations during an initial baseline (A) phase, then observations during a treatment (B) phase, followed by observations in a second baseline (A) phase. The implementation of a second baseline phase allows the researchers to observe if the behavior reverts back to the original baseline levels. If this occurs then it is easier to attribute the changes observed to the treatment, and other alternative explanations become less plausible. One major

limitation of this design is that in certain settings it may not be legal or ethical to remove treatment from a participant.
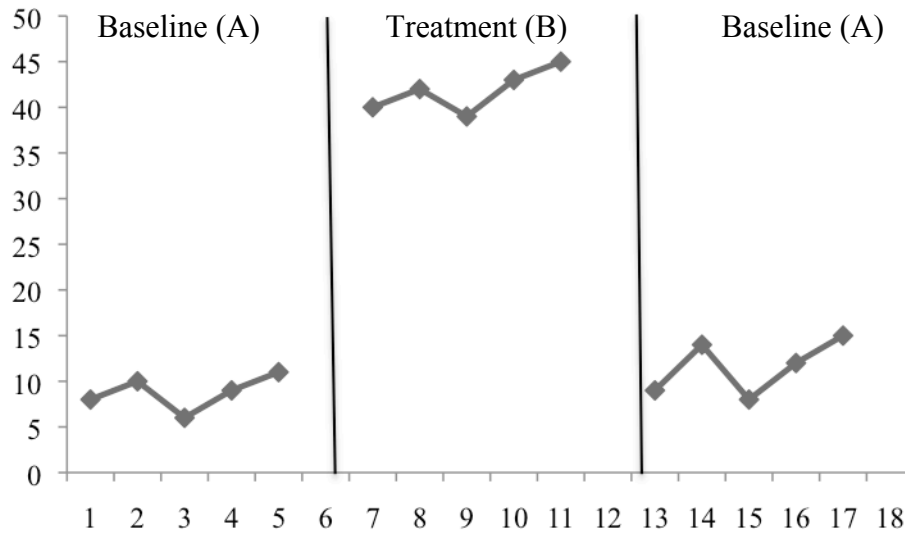
Figure 2. Graphical display of ABA reversal design

The final step of baseline logic, replication by affirmation of the consequent, is an attempt to strengthen what was observed in the initial treatment phase by reintroducing the intervention and creating an opportunity to observe the behavior change once again (Riley-Tillman & Burns, 2009).   This replication increases our confidence in the likelihood of a relationship existing between the participant's behavior and the implementation of the intervention. This replication can also be accomplished in other ways when the removal and reintroduction of the intervention is not feasible or is unethical.

An extension of the most simplistic reversal design (ABA) is the reintroduction of a treatment phase in an ABAB design (see Figure 3). This design consists of observations in an initial baseline (A) phase, then observations in an initial treatment (B) phase, followed by observations in a second baseline phase (A), and ending with

observations in a final treatment (B) phase. The inclusion of a final treatment phase

provides the opportunity for replication of the initial treatment phase in which the

observed behavior should revert back to the change seen in the initial treatment phase.
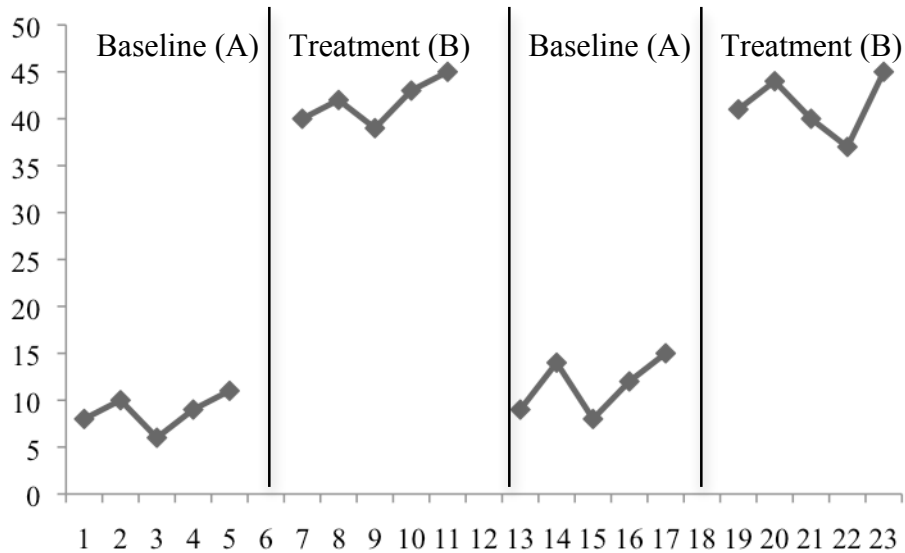
Figure 3. Graphical display of ABAB reversal design

One major limitation of single-case designs is their lack of generalization beyond

the one case that is being studied. The ability to generalize can be accomplished through

replication. Barlow, Nock, and Hersen (2009) state that there are at least three types of

generalization in behavior change research: (1) generality of findings across participants,

(2) generality of findings across behaviors, and (3) generality of findings across settings.

One natural way of achieving these various types of generalizations is through

replication. There are various ways to replicate single-case experiments, such as

replication of the baseline and treatment phase, as discussed previously in baseline logic,

or simultaneous replication built into the study design (Van den Noortgate & Onghena,

2007). A multiple baseline design allows for this simultaneous replication and can often

be used when the removal and reintroduction of the intervention is not feasible or is unethical.

A multiple baseline design is another type of extension of the traditional AB design (see Figure 4). This extension of the AB design simply establishes a baseline and treatment phase for multiple participants, behaviors, or settings. The initiation of the treatment phase is staggered across time, creating different baseline lengths for different participants, behaviors, or settings. By staggering the length of the baseline phases, it is more plausible to attribute a change in the data to the treatment, as we would not expect changes in history or maturation to stagger themselves across time (Ferron & Rendina-Gobioff, 2005). While this type of design does have many advantages, it does still have a few limitations. For example, when there is a lack of independence between baselines or when treatment effects vary across participants, behaviors, or settings it is more difficult to accurately attribute changes in the data to the treatment.
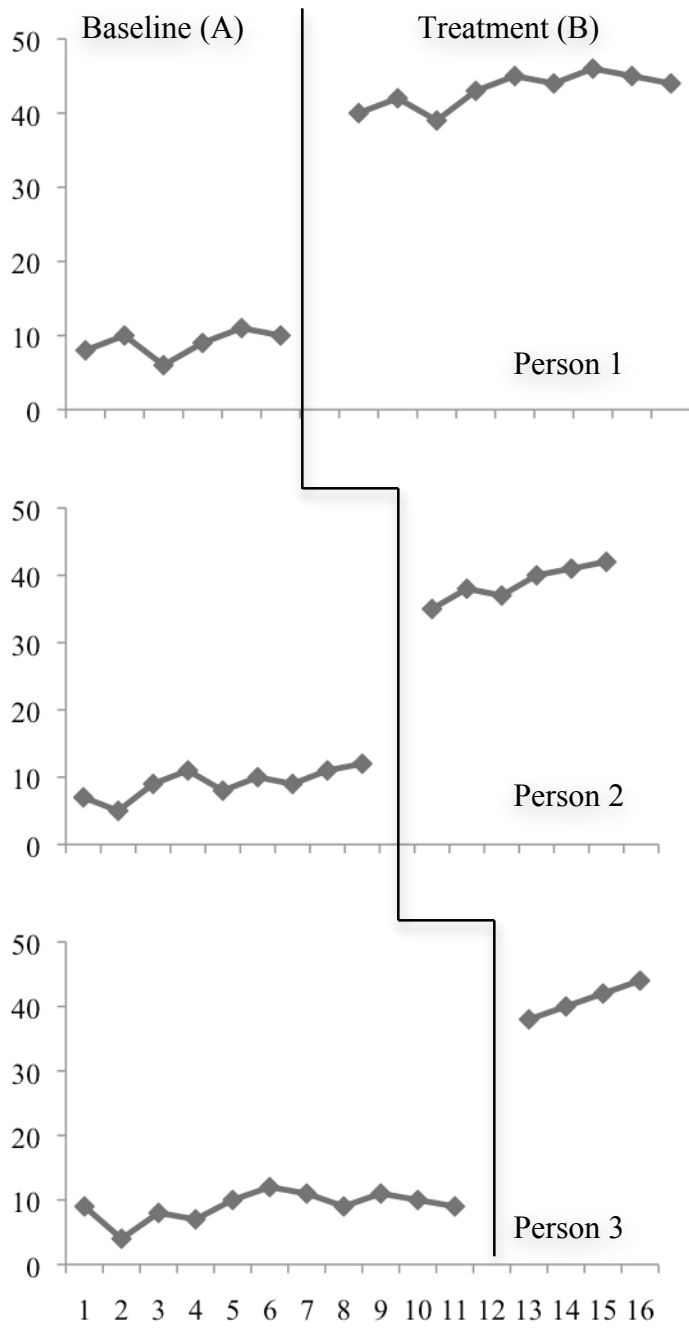
Figure 4. Graphical display of multiple baseline design

Analysis Options

Single-case research has been wrought with disagreement on the most appropriate method to analyze data. These analysis options can be grouped into three broad categories: (1) visual analysis, (2) descriptive statistics, and (3) inferential statistics.

Visual analysis. Historically, visual analysis of data has been the preferred analysis option (Kazdin, 2011; Parsonson & Baer, 1992). "The underlying rationale is to encourage investigators to focus on interventions that produce potent effects and effects that would be obvious from merely inspecting the data" (Kazdin, 2011, p.286). Proponents of visual analysis have argued that researchers who primarily rely on visual analyses of their graphed data are more likely to commit Type II (miss) errors than those who primarily rely on statistical analyses (Kazdin, 2011),in essence stating that visual analysts tend to be more conservative when evaluating the effectiveness of a particular treatment and therefore visual analysts commit fewer Type I (false alarm) errors (Parsonson & Baer, 1986). However, despite these claims, there have been several criticisms of visual analysis (DeProspero & Cohen, 1979; Jones, Weinrott, & Vaught, 1978; Matyas & Greenwood, 1990; Wampold & Furlong, 1981). Matyas and Greenwood (1990) argued that visual analysts were not as conservative as previously claimed, committing Type II errors 0% to 22% of the time and Type I errors 16% to 84% of the time. Additionally, Jones, Weinrott, and Vaught (1978) examined conclusions made from visual analysis as compared to statistical analysis and found that there was little agreement between the two. Also, Jones et al. (1978) and DeProspero and Cohen (1979) examined inter-rater agreement among judges and found that reliability was low. These conclusions support the assertion that visual analysis is not as consistently reliable and

conservative as once purported. Therefore, single-case researchers can supplement visual analysis with varying statistical analysis options.

Descriptive statistics. One such statistical analysis option is computing descriptive statistics or summary measures. These descriptive statistics include within phase measures (i.e., means, medians, standard deviations, root mean square error, and trend lines) and between phases measures (i.e., varying types of effect sizes). Just as there is contention in the literature as to how to analyze single-case data, there is also disagreement over how to summarize these effects. Effect sizes can be broken down into three overarching categories: (1) standardized mean difference, (2) regression based, and (3) non-regression based.

One approach is the standardized mean difference (Busk & Serlin, 1992), where the difference in baseline and intervention means is divided by the baseline standard deviation (d_B) or by the pooled standard deviation (d_P). More formally,

$$d_b = \left(\frac{\overline{x}_{Treatment} - \overline{x}_{Baseline}}{SD_B} \right) \tag{6}$$

and

$$d_p = \left(\frac{\overline{x}_{Treatment} - \overline{x}_{Baseline}}{SD_P} \right) \tag{7}$$

where $\overline{x}_{Treatment}$ is the mean of the treatment phase, $\overline{x}_{Baseline}$ is the mean of the baseline phase, SD_B is the standard deviation of baseline phase, and SD_P is the pooled standard deviation across baseline and treatment phases. Busk and Serlin (1992) recommend using the difference in baseline and treatment means divided by the baseline standard deviation (see Equation 6) when normality of the population distribution and equality of

the variances cannot be assumed. Otherwise, if the assumptions of normality and

equality of variances or at least the assumption of equality of the variances are met then it

is suggested to pool the variances and calculate a standardized mean difference based on

the formula in Equation 7 (Busk & Serlin, 1992).

Another category of approaches for effect size calculations is regression based.

These types of effect sizes are able to account for trends in data. One variation of this

approach includes the difference between the treatment trend line and the extension of the

baseline trend line at the first point in treatment or at the last point in the treatment

(Allison & Gorman, 1993). Other variations include computation of an f^2 value

representing a change in R^2 values corresponding to a change in level and a change in

slope (Kromrey & Foster-Johnson, 1996) and standardizing regression coefficients that

correspond to a shift in level and a shift in slope (Van den Noortgate & Onghena, 2003).

A final category is non-regression based effect sizes. There are several possible options

in this category. One possible option is the percentage of non-overlapping data (PND;

Scruggs, Mastropieri, & Castro, 1987). The PND is calculated by identifying the highest

or lowest point (depending upon which direction the data is expected to move) in the

baseline phase and then finding the percent of treatment phase data points that exceed it.

Other possible options have been created as alternatives to the PND, such as (a) the

percentage of data points exceeding the median (PEM; Ma, 2006), which is calculated by

finding the percentage of treatment data points above the median level of baseline data

points, (b) the mean baseline reduction (MBLR; Lundervold & Bourland, 1988), which is

the difference between baseline and treatment phase means divided by the baseline mean

and then multiplied by 100, and (c) the percentage of all non-overlapping data (PAND;

Parker, Hagan-Burke, & Vannest, 2007), which is the percentage of data points whose removal from either phase would eliminate all data overlap between phases.

Inferential statistics. Another possible statistical analysis option is inferential tests. There are a plethora of options available; however, one of the most often employed and well-researched options are randomization tests.

Randomization tests. Randomization tests make no assumptions about the distribution of the data and only use information obtained from the sample to evaluate the null hypothesis. However, these tests do assume random assignment of the data points or measurement occasions to either baseline or treatment phase. By randomly assigning measurement occasions to baseline or treatment, the study design can be classified as an experimental one (Onghena & Edgington, 2005). The use of an experimental design minimizes threats to internal validity by accounting for extraneous variables in both baseline and treatment phases.

The random assignment of measurement occasions can be thought of in two schemes. The first assumes the intervention can be alternated at any given measurement occasion. For example, let's assume a researcher wants to gather 12 observations on a single individual with 6 observations in baseline and 6 observations in treatment. The researcher could randomly assign the 12 observations to either baseline or treatment. The second is utilized when alternating the intervention at any given measurement occasion is not feasible. This randomization scheme assigns the timing of the phase change from baseline to intervention (Barlow, Nock, & Hersen, 2009). Specifically, the measurement occasions are assigned to a specific phase shift. For example, let's assume we have a basic AB design with 27 total measurement occasions, and each phase must have a

minimum of four observations each. The start of the intervention phase could occur on

one of 20 possible occasions (see Figure 5). The logic behind randomization tests is that

if the treatment has no impact on the dependent variable, then what is observed will not

be affected by the independent variable (treatment assignment); the order of the

assignment of the independent variable should not matter (Barlow, Nock, & Hersen,

2009). This null hypothesis is tested by comparing an obtained test statistic to a

randomization distribution that is formed by calculating a test statistic for all possible

permutations of the data.
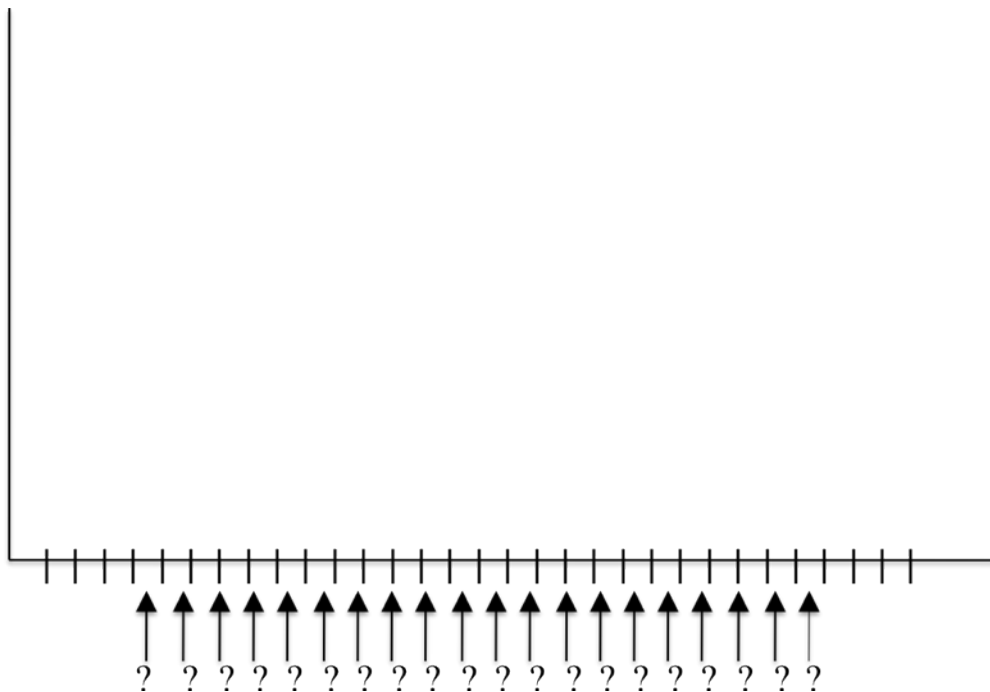
Figure 5. Example randomization scheme assigning phase shift from baseline to
treatment

A limitation of this method is that it only allows inference of the presence of a

treatment effect and not of the type of effect (i.e., change in level and change in slope) or

how big that effect is (Morgan & Morgan, 2001; Onghena & Edgington, 2005). In

addition, randomization tests are unable to provide interval estimates of the treatment effect (Ferron, Farmer, & Owens, 2010). It is also difficult to estimate power for this type of test since it depends on many factors, such as effect size, design, series length, and form of randomization, to name a few (Ferron & Onghena, 1996; Onghena & Edgington, 2005).

Regression analysis. Regression methods have also been proposed in the literature as a possible analysis option for single-case data (Huitema & McKean, 1998). A regression analysis can be performed to compare the treatment phase mean to the baseline phase mean for a specific individual using the following model:

$$y_i = \beta_0 + \beta_1 phase + e_i \tag{8}$$

where y_i is the observed value at i^{th} point in time, *phase* is a dummy coded variable (0 = baseline and 1= treatment), β_0 is the baseline mean, β_1 is the difference in means between the baseline and the treatment phases, and e_i is the error term at the ith point in time. A treatment effect can be determined by testing the regression coefficient to determine statistical significance. Equation 8 is the most basic model and can be further extended to include terms to evaluate trends in the phases (Center, Skiba, & Casey, 1985-1986; Huitema & McKean, 2000). Although the use of ordinary least squares (OLS) has been suggested for use with single-case data (Huitema & McKean, 1998; Shine & Bower, 1971), specifically multiple baseline designs, a major limitation of this model is that it does not take into consideration the dependency of the errors and it assumes the errors modeled are independent.

Autocorrelation. Due to the fact that single-case research is based on the premise that a single case is being measured repeatedly across time, many have argued that the

26

errors produced by these repeated measurements will be more similar when they are close

to each other in time and therefore positively autocorrelated (Kratochwill, Alden,

Demuth, Dawson, Panicucci, & Arnston, 1974; Matyas & Greenwood, 1997). Research

has shown that positive autocorrelation can impact statistical inferences by increasing

Type I error rates (finding a treatment effect when a treatment effect does not exist)

(Matyas & Greenwood, 1990; Toothaker, Banz, Noble, Camp, & Davis, 1983). Although

there is agreement on the negative effects of autocorrelation, there has been debate on the

extent to which single-case data are likely to illustrate autocorrelation (Busk &

Marascuilo, 1988; Huitema, 1985; Huitema & McKean, 1998; Matyas & Greenwood,

1997; Suen & Ary, 1987). According to Kazdin (2011), "The current verdict after

several studies is that serial dependence is likely to be present and ought to be taken into

account in evaluation of the data" (p.409).

  Multilevel modeling. As an alternative to the simple OLS regression model, the

use of multilevel models has been suggested for analyzing single-case data (Ferron, Bell,

Hess, Rendina-Gobioff, & Hibbard, 2009; Jenson, Clark, Kircher, & Kristjansson, 2007;

Nugent, 1996; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a,

2003b). Multilevel models allow for the analysis of hierarchical data that are organized

into two or more levels (Raudenbush & Bryk, 2002). For example, in educational

research when the focus is on the effectiveness of a new curriculum, students are

assigned to the treatment (receive the new curriculum) or control group by classroom.

The students (level one) in each of the classrooms are therefore nested within classrooms

(level two). Another example is when repeated measurements are gathered over time on

a set of participants. The measurements (level one) are therefore nested within the

participants (level two). This type of data structure is representative of single-case data and for the purposes of this study the focus was on this second example.

To examine single-case data within a study, a two level model can be used. The first level of the multilevel model is based on a simple linear regression model,

$$y_{ij} = \pi_{0j} + \pi_{1j} phase + e_{ij} \tag{9}$$

where y_{ij} is the observed score at measurement occasion i for participant j, *phase* is a dichotomous variable indicating whether a measurement occasion or observation occurred during baseline or treatment, π_{0j} is the baseline mean for participant j, π_{1j} is the treatment effect for participant j (i.e., the difference in means between baseline and treatment phases for participant j), and e_{ij} is the error at measurement occasion i for participant j, which accounts for within-phase error variance. The errors for participant j are typically assumed to be independent with a variance of (σ^2). However, this assumption of independence could be violated due to autocorrelation (Van den Noortgate & Onghena, 2003a). Therefore, it is possible with the use of multilevel modeling to assume a more complex covariance structure, such as a first-order autoregressive structure, which would account for possible autocorrelation (Ferron, Farmer, & Owens, 2010; Van den Noortgate & Onghena, 2003b). It should also be noted that just as in OLS regression, the first level of the multilevel model could be expanded to account for trends in the data (Van den Noortgate & Onghena, 2003b).

The second level of the multilevel model allows for variation across participants in both their baseline levels and their treatment effects.

$$\pi_{0j} = \beta_{00} + r_{0j} \tag{10}$$

and

$$\pi_{1j} = \beta_{10} + r_{1j} \tag{11}$$

where β_{00} is the average baseline, β_{10} is the average treatment effect, r_{0j} is an error term that indicates how far participant j's baseline mean is from the average baseline mean with a variance of τ_{00}, and r_{1j} is an error term that indicates the difference between participant j's treatment effect and the average treatment effect with a variance of τ_{11}. The error terms are assumed to be normally distributed and have a mean of zero.

Multilevel modeling provides three different types of parameter estimates: (1) variance components, (2) fixed effects, and (3) individual estimates. The variance components of a two-level model are the variance between participants' baseline means (i.e., τ_{00}) and the variance between participants' treatment effects (i.e., τ_{11}). The fixed effects are the average baseline means across participants (i.e., β_{00}) and the average treatment effect across participants (i.e., β_{10}). Finally, the individual estimates for each participant are the baseline mean for participant j (i.e., π_{0j}) and the treatment effect for participant j (i.e., π_{1j}).

A major advantage of multilevel modeling over other statistical analysis options, such as OLS regression, is its flexibility in handling serial dependency or autocorrelation. As discussed previously, the nature of single-case data lends itself to serial dependency. Some researchers have argued that autocorrelation does not exist in single-case data and therefore an OLS piecewise regression technique is an appropriate analysis option (Center, Skiba, & Casey, 1985-1986), and still others have debated the use of interrupted time-series analysis because the influence of autocorrelation is removed prior to analysis of the data (Crosbie, 1993). Although the interrupted time series method has the ability

to account for the influence of autocorrelation, it requires a large number of data points to adequately implement the procedure (Busse, Kratochwill, & Elliott, 1995; Crosbie, 1993). Multilevel modeling provides an alternative solution to handling serial dependency by having the flexibility to model a more complex covariance structure, such as a first-order autoregressive structure (Raudenbush & Bryk, 2002).

According to Van de Noortgate and Onghena (2003a), several other advantages of the use of multilevel modeling exist as well. One advantage of multilevel modeling is the flexibility of the model to handle heterogeneous variances and moderating variables. Another advantage of the use of multilevel models is that the individual parameter estimates are based on data from all the cases and therefore can still be relatively reliable, even with a small number of observations per case. Lastly, software for estimating the parameters has become readily accessible.

Although several advantages exist, some limitations or concerns also exist. One concern focuses on sample size. Multilevel models are typically estimated using restricted maximum likelihood methods. Those methods were developed under a large-sample theory and most recommendations specify the use of at least 30 units at the upper level (Hox, 1998). Previous research has indicated that regardless of sample size, fixed effect parameter estimates are unbiased but variance components may be biased (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Maas & Hox, 2004;Mok, 1995;Raudenbush & Bryk, 2002).

Ferron, Bell, Hess, Rendina-Gobioff, and Hibbard (2009) investigated the quality of inferences from multilevel modeling of multiple baseline data. Specifically, the authors examined, for the models in Equations 9 – 11, the interval estimates of the

average treatment effect. Ferron et al. (2009) used Monte Carlo simulation methods to examine multiple baseline studies having four, six, or eight participants (level-2 sample size) and series lengths of 10, 20, or 30 (level-1 sample sizes) for each participant. Their results indicated the fixed effect estimate of the average treatment effect was unbiased, regardless of sample size. In addition to relative bias, confidence interval estimates were also examined and as long as the Kenward-Roger or Satterthwaite degrees of freedom methods were used, accurate confidence interval estimates could be obtained. Specifically, the coverage estimates were close to the nominal .95 value, ranging from .965 to .935, when autocorrelation was modeled.

However, the results of Ferron et al. (2009) also indicated that estimates of the variance components tended to be biased. Although the average relative bias estimates of the variance of the treatment effect did decrease as sample size got larger, ranging from 34% when the sample size was four to 21% when the sample size was eight, a 21% upperward bias for a sample size of eight still represents substantial bias. These results were similar to previous research on two level organizational models, where Maas and Hox (2004) indicated a 25% upperward bias in the level-2 variance components with a level-2 sample size of 10 and a level-1 sample size of 5.

Ferron, Owens, and Bell (2010), in an extension of past research to include more complex treatment effects and a larger number of participants, found results similar to Ferron et al. (2009). Equations 12 - 16 contain the model that was under investigation. Equation 12 represents the first level of the multilevel model where the outcome (y_{ij}) was modeled as a function of *time* (centered so 0 represents the first point in treatment), a

dichotomous variable *phase* (0 = baseline, 1 = intervention), and the interaction between *time* and *phase*,

$$y_{ij} = \pi_{0j} + \pi_{1j}time + \pi_{2j}phase + \pi_{3j}time * phase + e_{ij} \tag{12}$$

where π_{0j} is the predicted value of the baseline trajectory for participant j when time = 0 or the first point in treatment, π_{1j} is the baseline slope for participant j, π_{2j} is the treatment effect for participant j (i.e., the difference in predicted values between baseline and treatment trajectories for participant j) at the first point in treatment, π_{3j} is the change in slope from baseline to treatment for participant j, and e_{ij} is the error at measurement occasion i for participant j, which accounts for the within-phase error variance. At level-2 each of the level-1 coefficients was allowed to vary across participants,

$$\pi_{0j} = \beta_{00} + r_{0j}, \tag{13}$$

$$\pi_{1j} = \beta_{10} + r_{1j}, \tag{14}$$

$$\pi_{2j} = \beta_{20} + r_{2j}, \tag{15}$$

and

$$\pi_{3j} = \beta_{30} + r_{3j}. \tag{16}$$

Results from Ferron, Owens et al. (2010) indicated that fixed effects coverage estimates for both the average treatment effect and average change in slope ranged from .917 to .962 and .908 to .963, respectively, when the Kenward-Roger degrees of freedom method was used. In addition, as participants increased from three to 32, the average fixed effects confidence interval coverage for the Kenward-Roger method increased.

Ferron, Farmer, and Owens (2010) continued to research the statistical functioning of multilevel modeling. Ferron, Farmer et al. (2010) moved beyond average treatment effects and their variance components to examining the accuracy of individual treatment effects and their confidence intervals (model given in Equations 9 – 11). The researchers conducted a Monte Carlo simulation study that examined multiple baselines of four, six, or eight; series lengths of 10,20, or 30 observations; and autocorrelation values of 0, .1, .2, .3, or .4. The confidence intervals of the empirical Bayes estimates of the individual treatment effects (i.e., π_{ij} from Equation 9), using the Kenward-Roger method, provided accurate confidence intervals across all design factors studied. The precision of the confidence interval width was widest utilizing the Kenward-Roger method but rapidly decreased as series length increased. In addition, the confidence interval coverage showed variation across study conditions when the OLS method of estimation was used. The confidence interval was accurate when no autocorrelation was simulated and tended to undercover when positive autocorrelation was simulated. This finding was not too surprising, given previous research that examined the utility of OLS methods in the presence of autocorrelation (Matyas & Greenwood, 1990; Toothaker, Banz, Noble, Camp, & Davis, 1983).

In conclusion, the research examining the use of multilevel modeling (specifically two-level models) to analyze single-case data has been promising. The degree to which multilevel modeling is functioning properly, under small sample sizes, depends on the type of parameter being estimated (Ferron et al., 2009). If the focus is on the variance components and sample sizes are small, then the estimates will not be very accurate. However, if the focus is on the fixed effects, the parameter estimates are often accurate,

as long as the error structure is correctly specified and the Kenward-Roger degrees of freedom method is used.

Meta-Analysis

The quantitative integration of findings is a necessary component of all types of research. The ability to integrate findings across studies allows researchers to make statements about the relationships between variables and the effectiveness of interventions across varying study characteristics. The idea of research synthesis and moving beyond statistical significance has been around since the early 1900s but was not termed meta-analysis until 1976, by Gene Glass (Cooper & Hedges, 2009). Glass (1976) defines meta-analysis as "…the analysis of analyses…the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p.3). Although Glass (1976) was technically the first to coin the term "meta-analysis", other researchers have been involved in expanding the analysis options available in terms of meta-analysis methods (Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Rosenthal & Rubin, 1986).

Meta-analysis provides many advantages when summarizing results across research studies. One advantage of meta-analysis is that it is a structured and systematic research technique that is open to replication. The steps involved in conducting a meta-analysis are required to be well documented and therefore open to replication. "By making the research summarizing process explicit and systematic, the consumer can assess the author's assumptions, procedures, evidence, and conclusions rather than take on faith that the conclusions are valid" (Lipsey & Wilson, 2001, p. 6). In addition, the analysis is more sophisticated than traditional review processes such as "vote-counting".

Vote-counting is simply the process of taking all the studies measuring the relationship of interest and counting the number of statistically significant results and the number of non-statistically significant results. The category with the most counts wins, so if numerous non-significant studies are found then the conclusion may be reached that there was not a relationship or effect. This method becomes problematic because statistical significance is dependent on sample size, so studies with small samples may find effects that are meaningful but may not find statistically significant results due to low power (Lipsey & Wilson, 2001; Schmidt, 1996). Another major advantage of meta-analysis is its ability to move beyond a qualitative review of study findings and into a more detailed analysis of the relationships between the study characteristics and the study findings. This analysis of the relationships between study characteristics and study findings is typically called a moderator analysis and provides a means of explaining possible variation in effect sizes (Lipsey & Wilson, 2001). A final advantage of meta-analysis is its ability to handle a large number of studies. The procedures involved in a meta-analysis allow researchers to systematically keep track of study details without losing information. Lipsey and Wilson (2001) did append this advantage by stating, "Meta-analysis does not require large numbers of studies and, in some circumstances, can be usefully applied to as few as two or three study findings" (p.7).

Individual Participant Data Versus Aggregate Data

There are two forms of meta-analysis: aggregate data meta-analysis and individual participant data meta-analysis. In an aggregate data (AD) meta-analysis the statistical synthesis is conducted by utilizing summary statistics from published and/or unpublished studies to calculate effect sizes and then statistically combining these effect

sizes in order to obtain an average effect size across studies as well as an associated confidence interval. In contrast, an individual participant data (IPD; Cooper & Patall, 2009) meta-analysis "…involves the central collection, checking, and re-analysis of the raw data from each study in order to obtain combined results" (p.166). After the raw data is collected from each study and if the outcomes across studies have been measured the same then the data is pooled and re-analyzed using traditional inferential statistics (Cooper & Patall, 2009). Although IPD meta-analyses are rare in large group social science literature, they have been extensively investigated in the medical literature.

Table 2 provides a listing of the relative benefits of both the IPD and AD meta-analysis. According to Cooper and Patall (2009), two major benefits of AD meta-analysis are that the meta-analysis can be done relatively quickly and with relatively low cost incurred to the meta-analyst, as compared to the IPD meta-analysis. Two benefits of the IPD meta-analysis are (a) the ability to perform subgroup analyses that were not performed by the original researchers; and (b) the ability to check the data for possible errors. While it is evident that both have benefits over the other, it is obvious that when availability of the data is not an issue the benefits of IPD meta-analysis outweigh those of AD meta-analysis (Cooper & Patall, 2009). However, obtaining individual data from large-group studies is highly unlikely and the use of AD meta-analysis will continue until such "data sharing" becomes available. It is, however, commonplace to include individual level data in studies utilizing a single-case design. This notion of having access to the individual participant data is certainly an advantage for single-case meta-analysts. Nevertheless, AD meta-analysis has historically been the focus of meta-analysis literature.

Table 2
Benefits of Individual Participant Data and Aggregate Data Meta-Analysis

Individual Participant Data (IPD) Meta-Analysis
Subgroup analyses that were not originally conducted can be performed
Data from the original studies can be checked
Ability to ensure that the original analyses were conducted properly, as well as standardization analyses across studies
Complex analyses can be performed more easily
New information can be added to the data sets
Moderator analyses can be conducted with greater power, assuming all individual participant data sets are available
Between-study and within-study moderator analyses can be performed

Aggregate Data (AD) Meta-Analysis
Cost, in both money and time, is less
Time to complete analysis is faster
Ability to include group-level statistics for which individual participant data are not available
Bias could be decreased if study results are associated with availability of individual participant data
Power could be increased to detect effects if many studies are available without individual participant data

Note: From "The Relative Benefits of Meta-Analysis Conducted With Individual Participant Data Versus Aggregated Data," by H. Cooper and E.A. Patall, 2009, *Psychological Methods*, *14*, p. 172. Copyright 2009 by the American Psychological Association. Adapted with permission.

Procedures

  A meta-analysis, or research synthesis, moves beyond the traditional literature search and combines data into a quantitative analysis. However, prior to and after the analysis stage of the meta-analysis process there are several steps that also need to be considered. Cooper (2007) outlines six stages of research synthesis. The first stage is to define the problem. This stage consists of identifying and defining variables and the relationships among those variables in order to identify the research studies that will be relevant to the problem of interest. One characteristic of a good meta-analysis is an explicit statement about inclusion and exclusion criteria (Lipsey & Wilson, 2001). This allows the readers of the meta-analysis to determine the specific research domain and the

criteria for why a study was included or not. The second stage is to collect the research evidence. Specifically, identify sources (e.g., databases, journals, conference proceedings) and key terms needed to identify relevant studies. During this stage meta-analysts attempt to identify and locate every study defined within the specified research domain that meets the eligibility requirements. The third stage is to evaluate the data. Once the relevant studies have been collected, specific information or data must be extracted from the studies in order to best synthesize the information to address the problem of interest. Specifically, this stage involves the coding of data. "The coding procedures for meta-analysis revolve around a coding protocol that specifies the information to be extracted from each eligible study"(Lipsey & Wilson, 2001, p.73). The fourth stage consists of data analysis. This stage involves the identification and application of specific statistical procedures to quantitatively integrate the data from each individual study. In an AD meta-analysis, the distribution of effect sizes are analyzed to examine the variability and obtain an estimate of the average effect size and its corresponding confidence interval, as well as testing for differences among effect sizes. In an IPD meta-analysis, raw data are obtained from all studies and if the outcomes were measured the same across the studies then the data are pooled together and re-analyzed using "traditional inferential statistics or more sophisticated techniques" (Cooper & Patall, 2009, p. 166). The fifth stage involves the interpretation of the analysis results. Meta-analysis methods allow researchers to make inferences about specific relationships and the average magnitude of effects sizes across studies. The sixth and final stage is presenting the meta-analysis results. This stage involves making judgments about what to report and how to disseminate findings to a broader audience.

Analysis Considerations and Methods

Although there are many steps involved in the meta-analysis process, a major component and arguably the most defining feature is the analysis or quantitative integration of data across studies.

IPD meta-analysis. In an IPD meta-analysis raw data or individual participant data are obtained from each study and then each participant's data are incorporated into an analysis option that is appropriate for the research questions. The use of IPD meta-analysis allows for many possible analysis options including MANOVA, multiple regression, structural equation modeling (SEM), or multilevel modeling. However, a recent review of IPD meta-analyses indicated that the most common analysis option used was a two-stage process that consisted of obtaining the raw data in each study converting to a standardized effect size and then combining the effect sizes across studies. This process parallels the processes involved in an AD meta-analysis (Simmonds et al., 2005).

AD meta-analysis. In order to quantitatively integrate findings, an AD meta-analyst needs an effect size from each included study. "An effect size is a number that reflects the magnitude of the relationship between two variables" (Borenstein, 2009, p. 220). Specifically, an effect size represents the strength and direction of an effect. There are various types of effect sizes, and their applicability is specific to the research problem. For example, an effect size could represent how much a treatment (independent variable) impacted social skills (dependent variable) as compared to no treatment, or an effect size could represent an index of the relationship between two variables such as depression and alcoholism. These effect sizes are then combined and compared in a meta-analysis.

Effect sizes provide standardized estimates, which allow us to combine them across studies. If all studies investigated exactly the same constructs and used the same sample sizes and instruments, then combining effect sizes would be easy; all studies would be exact replicates of each other. However, this is rarely--if ever--the case and meta-analysts must make certain decisions to determine how to combine studies that differ in many methodological and substantive ways (Shadish & Haddock, 2009). Therefore, Hershberger, Wallace, Green, and Marquis (1999) suggested that the method chosen for combining effect sizes across studies "must be able to provide overall estimates of treatment effectiveness and the precision of those estimates as well as assessments of the magnitude and direction of effects of other variables or factors on treatment effectiveness" (p.119).

Weights. As illustrated earlier, not all studies are exact replicates of each other and therefore it has been suggested in the literature to account for varying study characteristics by weighting each effect size. Shadish and Haddock (2009) suggest that weighting schemes rest on three assumptions. First, studies with certain characteristics are less biased, with regard to inferences, than studies with other characteristics. Second, prior to combing effect sizes, the bias of those characteristics can be estimated. Third, in order to compensate for the bias, suitable weights can be calculated and are defensible. Several weighting schemes have been proposed in the literature and adequately address all three of the assumptions outlined previously (Hedges & Olkin, 1985; Hunter & Schmidt, 2004). However, most of the literature relating to the weighting of effect sizes focuses on large-group studies. The types of weights proposed for large-group studies, such as the inverse of the variance or the within study sample size, would not be

appropriate for use with single-case data due to the relatively small sample sizes used in these types of designs. Therefore, some singe-case researchers have suggested weighting each effect size by the number of observations in the series (Shadish& Rindskopf, 2007; Faith, Allison, & Gorman, 1996).

Calculating the effect size mean and distribution. After gathering effect sizes from each study and choosing an appropriate weighting scheme, the effect sizes are statistically combined to describe the distribution of the effect sizes. Specifically, means and confidence intervals are calculated. The mean effect size represents a point estimate of the population effect size, and the confidence interval indicates a range of possible values in which the population effect size is likely to be. The confidence interval provides a degree of precision around the mean effect size and can also be used to determine statistical significance in relation to the null hypothesis that there is no effect in the population (Lipsey & Wilson, 2001).

Meta-analysis models. Beyond calculating mean effect sizes and confidence intervals lies another important component of meta-analysis: homogeneity of the effect size distribution. Meta-analysts must decide if the effect sizes included in their estimate of the mean effect size are all estimating a single population effect size or are from a distribution of population effect sizes. This decision leads meta-analysts to choose between two types of statistical models, fixed or random effects.

A fixed effects model assumes a common effect size across all studies ($\theta_1 = \theta_2 = \theta_3 = ... = \theta_k$) (Shadish & Haddock, 2009). In other words, in a fixed effects model it is assumed that one true effect size exists in the population, with variability being only due to sampling error. In contrast, under a random effects model one would

not assume that one population effect size exists but rather a distribution of population effect sizes exists. Therefore differences in effect sizes are based on underlying population differences and are not just due to sampling error.

The decision whether to use a fixed or random effects model does not have one single correct answer. Some would argue that conceptually the random effects model makes the most sense due to the fact that it reduces to the fixed effects model when the variance component is zero or when no random variation exists (Shadish & Haddock, 2009). Others would encourage the use of a homogeneity test statistic, such as the Q statistic (Hedges & Olkin, 1985). This test allows the homogeneity of variance to be tested statistically, indicating that rejection of homogeneity implies that it is tenable to assume that the variability among effects sizes is greater than what could have occurred due to sampling error alone (Lipsey & Wilson, 2001). However, the Q statistic has low power with small sample sizes and therefore may fail to reject homogeneity when in fact there is variability among the effect sizes that is due to more than just sampling error. Still others would argue that the choice of models depends on the inferences the researcher hopes to make (Hedges & Vevea, 1998).

Threats to Validity

Researchers have discussed several potential threats to the validity of inferences made from meta-analysis (Matt & Cook, 2009). Some threats relate specifically to inferences about the association between an independent and a dependent variable, such as an intervention effect on an outcome variable. These possible threats are (a) unreliability in primary studies, (b) restriction of range, (c) missing effect sizes, (d) unreliability of meta-analytic codings, (e) increased Type I error rates, (f) sampling bias,

(g) dependent effect sizes, (h) failure to use weighted effect sizes, (i) inappropriate meta-analysis model selection, and (j) lack of statistical power (Matt & Cook, 2009). Although a single study's deficiencies will not likely threaten the inferences made from a meta-analysis, the occurrence of a deficiency across multiple, included studies can lead to increased Type I or Type II errors (Matt & Cook, 2009).

Another often and most persistent criticism of meta-analysis is the notion of apples and oranges (Lipsey & Wilson, 2001). The apples and oranges issue deals with the inclusion of studies that deal with a wide variety of different constructs and/or utilize different instruments to measure variables. This becomes an issue when combining effect sizes across studies and calculating a grand mean effect size. However, at the heart of meta-analysis is the examination of the distribution of effect sizes, and often of primary interest to the meta-analyst is the identification of sources of variability that are due to study differences (Lipsey & Wilson, 2001).

Single-Case Meta-Analysis

Although the use of single-case designs to evaluate interventions has grown in popularity over the last decade, their inclusion in meta-analyses and the methodological research encouraging their inclusion has been limited (Busk & Serlin, 1992; Busse, Kratochwill, & Elliott, 1995; Jenson, Clark, Kircher, & Kristjansson, 2007; Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008; Van den Noortgate & Onghena, 2003b). Most research involving meta-analysis has focused on large-group studies (Glass, 1976; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Rosenthal& Rubin, 1986) and while these methods have worked well with results from large-group comparison studies, there is still disagreement over the best way to meta-analyze results from single-

case studies (Beretvas & Chung, 2008). Nevertheless, the inclusion of results from single-case studies in meta-analyses is necessary for many reasons.

The inclusion of single-case studies in meta-analysis allows for information about the overall treatment effect without losing information about the individual cases. A single-case study involves the repeated measurement of one or a few cases over time, offering information on the variability in the treatment effect of individual cases. When several single-case studies are aggregated together, the overall treatment effect can be estimated as well as the effects for individual cases (Van den Noortgate & Onghena, 2003a). In addition, the aggregation of several single-case studies increases the generalizabilty of the findings. A major criticism of single-case designs is their lack of generalizabilty, and by combining several single-case studies together, it becomes tenable to assume greater generalizabilty of the results.

Analysis Options

The earliest proposal for meta-analyzing single-case data dates back to 1984, where Gingerich proposed the use of meta-analysis methods developed by Smith, Glass, and Miller (1980). Specifically, Gingerich (1984) proposed calculating a standardized mean difference between post-test and pre-test scores with the standard deviation of the pre-test scores serving as the standardizing unit. Although his idea for synthesizing single-case data across studies was well intended, his suggestion for using Glassian meta-analytic methods does not take into account the serial dependence among single-case data. His argument in favor of this method is actually more of an argument against the notion that serial dependency or autocorrelation is a characteristic of single-case data, a questionable argument given the nature of the data.

Summary measures. Following Gingerich's (1984) proposal, a non-parametric or non-regression based method was proposed. Scruggs, Mastropieri, and Castro (1987) suggested an approach to calculate the percentage of non-overlapping data (PND) between treatment and baseline phases. The PND is calculated by identifying the number of data points in treatment that exceed the highest data point in baseline divided by the total number of data points in treatment and then multiplied by 100. "When computation is completed, these outcome measures can be combined across studies to determine relative effectiveness of particular treatments" (Scruggs, Mastropieri, & Castro, 1987, p.27). Although relatively easy to compute, the use of PND as a meta-analytic approach has several limitations (Allison & Gorman, 1993). Allison and Gorman (1993) point out that the PND has the potential to misrepresent treatment effects when there is a trend in the data, outliers are present in the treatment phase, and the treatment has had a negative effect on the outcome. In addition, this proposed single-case meta-analytic approach does not take into consideration specific meta-analysis considerations such as the weighting of effect sizes or the use of appropriate meta-analytic models (i.e., fixed or random effects).

Busk and Serlin (1992) suggest that the most appropriate effect size measure for both between- and within- subject experiments, given the assumptions of equality of variance and compound symmetry, is the standardized mean difference effect size, where the denominator is the square root of the mean square error in the design. The authors describe four advantages for this type of effect size.

First, one single definition holds for all experimental designs. Second, because the distribution of the effect-size measure is known, one can test the effect size

directly and build a confidence interval for it. Third, Hedges and Olkin's meta-

analytic techniques can be used, because they are based on large-sample, normal

approximations to the noncentral *t* distribution. And fourth, it is straightforward to

convert individual *t*s to effect sizes. (Busk & Serlin, 1992, p. 195)

However, if the assumptions needed to pool the within-phase variances are not met, then

other methods are needed to calculate and test the effect size measure. Busk and Serlin

(1992) present three approaches to obtain the effect size estimate. The three approaches

differ in the assumptions concerning the population distribution form and equality of

variances. The first approach, the Glassian original effect size estimate, makes no

assumptions and the standardized mean difference score is calculated by taking the

difference between the baseline and treatment phase means and dividing by the baseline

standard deviation.

The second approach assumes equality of variances across the baseline and

treatment phases but still makes no assumption about the population distribution form. In

this approach the within-phase variances are pooled to obtain better estimates of the

effect size. In the third approach, assumptions are made about the population distribution

as well as about equality of variances across baseline and treatment phase. Calculation of

the effect size measure doesn't change; however, by making the assumption that the

phase scores are from a normal distribution and that the within-phase variances are equal,

the distribution of the effect size is considered to follow a noncentral *t* distribution and

confidence intervals can be constructed for the individual effect estimates. In addition,

with the assumption of a normal distribution large-group meta-analytic methods can be

used to synthesize effects across studies.

The third approach is most in line with what meta-analysts are hoping to do; however, most single-case data do not adhere to these strict assumptions of equality of variances, compound symmetry, and normality of the distribution across baseline and treatment phases. Therefore, the use of confidence intervals and meta-analytic procedures that allow for the testing of specific hypotheses becomes inappropriate and limits the amount of information available to the meta-analysts. In addition, the formula suggested by Busk and Serlin (1992) when no assumptions can be made yields a numerator and denominator that are not independent of each other and can no longer be used in large-group meta-analytic methods for combining effect sizes across studies.

Inferential statistics. Beyond the proposal of specific summary measures, other researchers have suggested the use of various inferential tests to meta-analyze single-case data (Allison & Gorman, 1993; Center, Skiba, & Casey, 1985-1986; Onghena & Edgington, 2005; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). Onghena and Edgington (2005) propose the use of p value combining, based on the use of randomization tests, as a method to meta-analyze single-case data. The authors demonstrate that if the single-case experiments used in a meta-analysis provide independent tests of the same null hypothesis, then the p values can be combined by summing the p values across studies and comparing the sum to all other possible sums that could have occurred. The proportion of summed p values that is as small or smaller than the observed summed p value is then calculated to determine if the overall treatment effect is significant. Although the use of randomization tests does provide a meta-analyst with information related to whether there was or was not a treatment effect, it does not provide an estimate of the size of that treatment effect or the ability to test the impact of

47

other variables on the treatment effect. In addition, it has been well documented that p values are influenced by the size of the treatment effect as well as the number of observations included in the analysis (Onghena & Edgington, 2005).

A series of regression methods have also been suggested for use in meta-analyzing single-case data. Center, Skiba, and Casey (1985-1986) proposed the use of a piecewise regression technique that utilized raw data from individual single case studies. The technique used the following model:

$$y_i = b_0 + b_1 x + b_2 t + b_3 x(t - n_a) + e_i \tag{17}$$

where n_a represents the number of points in baseline, $b_1 x$ is a term for change in level, $b_2 t$ is a term for the baseline trend, and $b_3 x(t - n_a)$ is an interaction term to measure the change in slope due to the treatment. This technique produces two separate effect sizes (b_1, b_3), which can make interpretation more complicated. However, based on what we know about single-case data, attempting to represent treatment effectiveness with an effect size that only illustrates a change in level would not adequately account for changes in slope or the combined effects of level and slope changes.

Center, Skiba, and Casey (1985-1986) also proposed computing one effect size by calculating a difference in R^2 values between the full model (given in Equation 17) and a model without each of the parameters (b_1, b_3), and then converting that difference in R^2 values into an F-statistic, which can be converted to an often recognized and easily interpretable d effect size. By allowing investigation of changes in both slope and level this model proved to be a significant improvement over what was available at the time. Nevertheless, this technique did not take into account the autocorrelational nature of single case data and assumed that errors of successive observations were independent. In

48

addition, none of the methods up to this point acknowledged crucial meta-analytical issues such as the use of fixed or random effects models, the weighting of effect sizes, or the use of multiple effect sizes per study.

Allison and Gorman (1993) modified the method proposed by Center, Skiba, and Casey (1985-1986) to address concerns inherent in the model. Three specific problems were discussed and the model was improved upon to rectify these problems. The first problem was that under certain conditions the model could overestimate the effects of trend and thereby underestimate the overall effect size. Allison and Gorman (1993) corrected for this by computing trend on the baseline data only instead of across both phases. The second problem was that due to the nature of how the effect size is calculated, the effect can never go below zero. This is problematic because it is not consistent with the notion that sometimes treatments can have a negative impact and worse results can be produced. This problem was corrected for by recommending the application of the appropriate sign as indicated by the regression coefficient. The third problem was that the effect could be overestimated due to an increase in predictability of the dependent variable, regardless of whether or not the change was in the intended direction or not (Alison & Gorman, 1993). In order to address the third problem, the authors recommend that if the zero-order correlations have different signs, simply estimate the change in level because the change in slope will automatically attenuate its effect (Allison & Gorman, 1993). This model was again a significant improvement over the previous models but still did not take into account autocorrelation or key meta-analytic issues, such as weighting of effect sizes, independence of effect sizes, and meta-analytic model selection.

Although each of these models provided advances on their predecessors, they still leave much to be desired in the form of meta-analyzing single case data. Specifically, a method needs to be able to address the issue of autocorrelation, the standardization of effect sizes for combination across studies, and the use of a meta-analytical method that allows the further investigation of variability in effect sizes. The use of multilevel modeling provides the tools to be able to accomplish all of these goals.

Van den Noortgate and Onghena (2003a, 2003b, 2007, 2008) proposed the use of multilevel modeling to aggregate single-case data for the purposes of meta-analysis. The authors have suggested aggregating single-case data in three different ways. The first option includes individual level raw data from each primary study in the meta-analysis and makes the assumption that all dependent variables across studies are measured the same way. Van den Noortgate and Onghena (2008) illustrated this first option in a series of models provided in Equations 18 through 22.

Equation 18 represents within person variation, which can be modeled with a basic regression equation. Specifically, an outcome (y) is modeled on measurement occasion i for participant j in study k (y_{ijk}) as a linear function of a single predictor, *phase*:

$$y_{ijk} = \pi_{0jk} + \pi_{1jk} phase + e_{ijk} \tag{18}$$

where *phase* represents a dummy coded variable indicating whether measurement occasion i took place during the baseline (0) or treatment (1) phase. π_{0jk} is the level of the outcome during baseline for participant j from study k, π_{1jk} is the treatment effect for participant j from study k , and e_{ijk} is within-phase error variance.

At the second level, the variation across participants is modeled in the following equations:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad (19)$$

and

$$\pi_{1jk} = \beta_{10k} + r_{1jk} \qquad (20)$$

where the fixed effects are β_{00k}, the average baseline level for study k, and β_{10k}, the average treatment effect for study k, and the error terms are r_{0jk} and r_{1jk} that allow variation in both baseline levels and treatment effects among participants within study k.

At the third level, the variation across studies is modeled in the following equations:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad (21)$$

and

$$\beta_{10k} = \gamma_{100} + u_{10k} \qquad (22)$$

where the fixed effects are γ_{000}, the overall average baseline level, and γ_{100}, the overall average treatment effect, and the error terms are u_{00k} and u_{10k}, which allow variation in both the average baseline levels and average treatment effects among studies. It should be noted that errors on all levels were assumed to be independently normally distributed and have a mean of zero. However, multilevel models are quite flexible and the use of a complex covariance structure, such as a first order auto regressive structure, is possible to account for dependent errors.

Van de Noortgate and Onghena's (2008) second option assumes the dependent variable is measured differently across studies and therefore scores from individuals need to be standardized before combining them into one analysis. First, the individual level

raw data are standardized by performing an OLS regression for each participant

separately and dividing their scores by each resulting root mean squared error and then

combining the data into the models defined in Equations 18 through 22 (Van den

Noortgate & Onghena, 2008).

The third option proposed by Van den Noortgate and Onghena (2008) does not

include individual level data from each study in the meta-analysis. Instead, standardized

regression coefficients are calculated for each study and included in the meta-analysis as

effect sizes representing a standardized change in level and change in slope. In this

option, Equation 18 needs slight modifications to appropriately meta-analyze single-case

data. The first level of the model is adapted to model the effect sizes or standardized

regression coefficients from each study rather than the individual level data:

$$\hat{\pi}_{0jk} = \pi_{0jk} + e_{jk} \tag{23}$$

with $\hat{\pi}_{0jk}$ representing the observed effect size for participant j in study k modeled as the

true effects size (π_{0jk}) for participant j in study k plus some random variation or error

(e_{jk}), where the level-1 error variance matrix is assumed known. The second and third

level equations (see Equations 19 – 22) describing variation across participants and

between studies remain the same.

Multilevel modeling estimates (co) variance at each level but typically only

estimates fixed effect parameters at the highest level. Therefore, variance and covariance

estimates across all levels and fixed effects at the third level, the average baseline across

studies and the average treatment effect across studies can be reported. These types of

parameter estimates offer the ability not only to provide information on the overall

treatment effect but also information related to the variability of that overall average

treatment effect. In addition, predictors can be added to the model to account for that variability.

Van den Noortgate and Onghena (2008) argue that single-case study conclusions are restricted to the participants which were investigated, but multilevel modeling provides the ability to combine results from multiple participants and studies to gain information about not only the average treatment effect but also if and how the treatment effect varies across participants and studies. Another advantage of multilevel modeling is that it can be used to aggregate data from single-case studies that include multiple participants. This use of multiple data sources or effect sizes from the same study is typically problematic and has not been addressed by other proposed single-case meta-analytic methods. Multilevel modeling is structured to account for that "nesting" of data within studies by allowing variation within participants, between participants of the same study, and between studies (Van den Noortgate & Onghena, 2008).

Although all of the previous simulation research on multilevel modeling of single-case data (Ferron et al., 2009; Ferron, Farmer et al., 2010; Ferron, Owens et al., 2010) has focused on two-level models and the use of a three-level model has only been applied to a real world data set (Van den Noortgate & Onghena, 2008), the results have been encouraging. These findings provide motivation in the pursuit of empirically evaluating the utility of inferences made from a three-level model to meta-analyze single-case data.

Applications of Single-Case Meta-Analysis

Beretvas and Chung (2008a) conducted a narrative review of single-case meta-analyses that took place between 1985 and 2005; 24 articles were identified. Their results indicated that the most commonly used metric to summarize study results was the

PND, and it was most commonly used in combination with percent zero data (PZD). The next most popular effect size utilized was the standardized mean difference in various forms. Also, a form of time series analysis was used by a small percentage of studies, as well as the use of piecewise regression, which was incorrectly specified both times it was reported.

Although most of the meta-analyses reviewed by Beretvas and Chung (2008a) focused on studies using more complex designs (e.g., multiple baseline, reversal, alternating treatment) than a simple AB design, the most common metric used to summarize results only focused on the comparison of an intervention phase to a baseline phase. This focus can lead to a dependence of outcomes yielded by the same metric (Beretvas & Chung, 2008a). The results of Beretvas and Chung's (2008a) review indicated that the majority of meta-analyses reviewed did not clearly state how this dependence was handled. When analyzing multiple treatments per study, the most common method reported was to average the indices together. Further, when addressing the use of multiple measures per study, the majority of studies analyzed results separately for each measure and when multiple participants per study were involved, most of the meta-analyses ignored the dependence and treated each effect size as independent (Beretvas & Chung, 2008a). In terms of analyses conducted, the majority of meta-analyses simply averaged the effect sizes together. In addition, a few studies performed moderator analyses to explore variability in the effect sizes.

Farmer, Owens, Ferron and Allsopp (2010a) also conducted a review of single-case meta-analyses. Farmer et al. (2010a) searched for single-case meta-analyses that were conducted from 1999-2009. Their search yielded 39 articles for inclusion. Most of

the meta-analyses were related to education, with the majority in special education. The

majority of meta-analyses provided clear search procedures but did not tend to include

detailed information about the primary study characteristics (Farmer et al., 2010a).

However, when primary study information was provided, the meta-analyses reported the

use of studies that included more complex designs and tended to exclude simple AB

designs and those studies with less than three points per phase. Similar to Beretvas and

Chung (2008), the most common metric reported was the PND, and the majority of

studies computed averages of the effect sizes. The meta-analytic review also noted that

limited information was provided regarding effect size calculation, meta-analytic method

used, and any further analyses (e.g., moderator analyses) that were conducted. Farmer et

al. (2010a) cautioned single-case meta-analysts on the dangers of not providing enough

information to their readers and concluded with a suggestion that a table be included in

future single-case meta-analyses identifying the types of single-case designs used, the

phases used in the calculation of the effect sizes, and the number of effect sizes used from

each study.

Summary

Single-case designs provide the ability to intensively study the effect of a

treatment on a single case over time. The popularity of these designs has grown rapidly

over the past decades to include research in school psychology (Skinner, 2004), special

education (Algozzine, Browder, Karvonen, Test, & Wood, 2001), teacher education

(Hsieh, Hemmeter, McCollum, & Ostrosky, 2009), and behavioral intervention research

(Filter & Horner, 2009). However, the integration of single-case designs in meta-analytic

research has been far less frequent (Busk & Serlin, 1992; Busse, Kratochwill, & Elliott,

1995; Jenson, Clark, Kircher, & Kristjansson, 2007; Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008; Van den Noortgate & Onghena, 2003b). This infrequency may be due to the lack of methodological consensus on how to best synthesize single-case results across studies. Several methods have been proposed, such as the combining of the PND across studies (Scruggs, Mastropieri, & Castro 1987), the calculation of a standardized mean difference and use of traditional large-group meta-analytic methods (Busk & Serlin, 1992), the combining of p values through the use of randomization tests (Onghena& Edgington, 2005), several regression based methods that account for changes in level and slope (Casey, Center, & Skiba, 1985-1986; Allison & Gorman, 1993), and the use of multilevel modeling (Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). Among these methods, multilevel modeling has been recommended for use with single-case meta-analytic data due to features of the model that can handle characteristics of the data that are often problematic for other analysis options. However, further investigation into the utility of the inferences made from multilevel modeling is necessary to provide guidance to future single-case meta-analysts. Furthermore, the empirical evaluation of a three level single-case meta-analytic model under conditions that are similar to the field of social science is needed. Therefore, this study examined the utility of Van den Noortgate and Onghena's (2008) raw data multilevel modeling approach to the meta-analysis of single-case data. Specifically, the quality of the fixed effects (i.e., the overall average baseline level and the overall average treatment effect) and the variance components (e.g., the between person within study variance in the average baseline level, the between study variance in the overall average baseline level, the between person within study variance in the average treatment effect)

in a three level multilevel model were examined. More specifically, it investigated

confidence interval coverage rates, confidence interval widths, and bias of the point

estimates as a function of specific design and data factors. The raw data option was the

most fitting method to first evaluate, as it is the most basic model of Van den Noortgate

and Onghena's (2008) three proposed options.

CHAPTER THREE: METHOD

This chapter outlines the methods for this study, including the purpose, research questions, sample, and design.

Purpose

The purpose of this study was to examine the appropriateness of Van den Noortgate and Onghena's (2008) raw data multilevel modeling approach to the meta-analysis of single-case data. Specifically, the study examined the fixed effects (i.e., the overall average baseline level and the overall average treatment effect) and the variance components (e.g., the between person within study variance in the average baseline level, the between study variance in the overall average baseline level, the between person within study variance in the average treatment effect) in a three level multilevel model. More specifically, it investigated bias of the point estimates, confidence interval coverage, and confidence interval width as a function of specific design and data factors, such as the number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms.

Research Questions

1. To what extent are the fixed effect estimates from a three level meta-analytic single-case model biased as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary

study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

2. To what extent does the confidence interval coverage of the fixed effect estimates from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

3. To what extent does the confidence interval width of the fixed effect estimates from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

4. To what extent are the variance components from a three level meta-analytic single-case model biased as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

5. To what extent does the confidence interval coverage of the variance components from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

6. To what extent does the confidence interval width of the variance components from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

Design

This study utilized a 3 X 2 X 3 X 3 X 2 factorial design. The factorial design included five independent variables: (1) number of primary studies per meta-analysis (10, 30, and 80); (2) modal number of participants per primary study (small [mode = 4] and large [mode = 8]); (3) modal series length per primary study (small [mode = 10], medium [mode = 20], and large [mode = 30]); (4) level of autocorrelation (0, .2, and .4); and (5) variances of the error terms (most of the variance at level-1 [$\sigma^2 = 1$, $\tau_{\pi00} = \tau_{\pi10} = .2$, and

$\tau_{\beta00} = \tau_{\beta10} = .05$] and most of the variance at level-2 [$\sigma^2 = 1$, $\tau_{\pi00} = \tau_{\pi10} = 2$, and

$\tau_{\beta00} = \tau_{\beta10} = .5$]). For each of the 108 conditions, 5,000 data sets were simulated using SAS IML (SAS Institute Inc., 2008).

The dependent variables were bias, the average difference between the known parameter value and the parameter estimate for both the fixed effects (γ_{000} and γ_{100}) and the variance components ($\tau_{\pi00}$, $\tau_{\pi11}$, $\tau_{\beta00}$, $\tau_{\beta11}$, σ^2, and $\hat{\rho}$), confidence interval coverage, the proportion of 95% confidence intervals that contain both the fixed effects estimates and the variance components, and confidence interval width, the average difference between the upper and lower limits of the 95% confidence intervals for both the fixed effects and the variance components.

Five experimental variables were examined: (1) number of primary studies per meta-analysis, (2) modal number of participants per primary study, (3) modal series length per primary study, (4) level of autocorrelation, and (5) variances of the error terms. Of these variables, (1), (2), and (3) represent aspects of the meta-analysis, (4) represents aspects of the primary studies within the meta-analysis, and (5) represents aspects of both the meta-analysis and the primary study data.

Sample

Crossing the two variance levels of the error terms with the three levels of autocorrelation, a total of six data conditions were examined for each of 18 combinations of number of primary studies per meta-analysis, modal number of participants per primary study, and modal series length per primary study. For each of the 108 conditions (6*18), 5,000 data sets will be simulated using SAS IML (SAS Institute Inc., 2008). The use of 5,000 replications leads to a standard error of .003 when coverage is .95, which is an adequate level of precision when estimating coverage.

The sample for this study was generated through Monte Carlo simulation methods. The sample generation consisted of two aspects: (1) primary study characteristics and (2) meta-analytic characteristics. The primary study characteristics were based on specific values of the following factors: level of autocorrelation, and variances of the level-2 error terms. The number of primary studies included in each meta-analysis, the modal number of participants per primary study, the modal series length per primary study, and the variances of the level-3 error terms address characteristics of the meta-analysis. The factors used to define the simulated data are further defined below.

Conditions Sampled

Number of primary studies per meta-analysis. The number of primary studies in each meta-analysis had three levels (10, 30, or 80). These values were chosen based on a review that was conducted by Farmer, Owens, Ferron, and Allsopp (2010b) on 39 single-case meta-analyses in social science between the years of 1999 and 2009. Farmer et al. (2010b) found that the number of primary studies included in the meta-analyses ranged from 3 to 117, with 60% of the meta-analyses including less than 30 primary studies.

Modal number of participants per primary study. The modal number of participants per primary study had two levels (small and large). The small category contained 70% of primary studies with four participants, 20% of primary studies with six participants, and 10% of primary studies with eight participants in each meta-analysis, indicating a mode of 4 and an average of 4.7 participants per primary study. The large category contained 70% of primary studies with eight participants, 20% with six participants, and 10% with four participants in each meta-analysis, indicating a mode of 8 and an average of 7.2 participants per primary study. These categories were defined based on findings from Farmer et al. (2010b), where the average number of participants per study ranged from 1.4 to 30.67, with 93% of those values falling at or below seven and Ferron, Farmer et al. (2010), who found in multiple baselines designs the number of participants ranged from 3 to 10, with a median of 4. In addition, these levels were chosen based on recommendations of a minimum of four baselines and upwards of eight or nine to show treatment effects across behaviors, persons, or settings (Kazdin, 2011).

Modal series length per primary study. The modal series length per primary study had three levels (small, medium, and large). The small level contained 70% of primary studies with series lengths of 10, 20% of primary studies with series lengths of 20, and 10% of primary studies with series lengths of 30 in each meta-analysis, indicating a mode of 10 and an average series length of 14 per primary study. The medium level contained 70% of primary studies with series lengths of 20, 20% of primary studies with series lengths of 10, and 10% of primary studies with series lengths of 30 in each meta-analysis, indicating a mode of 20 and an average series length of 19 per primary study. The large level contained 70% of primary studies with series lengths of 30, 20% with series lengths of 10, and 10% with series lengths of 20 in each meta-analysis, indicating a mode of 30 and an average series length of 25 per primary study. These categories were chosen to represent a range of possible values in single-case meta-analyses. These levels were chosen based on the consistency with previous simulation studies investigating the use of multilevel modeling as a method of analyzing single-case data where series length of 10, 20, and 30 were modeled (Ferron et al., 2009; Ferron, Farmer et al., 2010b; Ferron, Owens et al., 2010a). In addition, Ferron, Farmer et al. (2010) conducted a survey of multiple baseline studies published in 2008 and found that average series lengths ranged from seven to 58 with a median of 24.

Level of autocorrelation. The level of autocorrelation in the primary studies was 0, .2, or .4. These values cover the range of possible autocorrelation values typically found in behavioral studies (Busk & Marascuilo, 1988; Huitema, 1985; Matyas & Greenwood, 1996). In addition, these values were consistent with past simulation studies

that investigated the utility of multilevel modeling (Ferron et al., 2009; Ferron, Farmer et al., 2010; Ferron, Owens et al., 2010).

Variances of the error terms. The variances of the error terms were comprised of two categories. The first category modeled the data to have most of the variance at level-1 or within person, with values of 1 for the level-1 error term, .2 for the level-2 error terms, and .05 for the level-3 error terms. The second category modeled the data to have most of the variance at level-2, with values of 1 for the level-1 error variance, 2 for the level-2 error variances, and .5 for the level-3 error variances. These values covered a range of possible values, such as those presented in Van den Noortgate and Onghena (2008) and previous simulation research (Ferron et al., 2009). In addition, the variance in the average baseline levels equaled the variance in the average treatment effects. Constraining the level-2 variances to be equal was consistent with previous simulation research (Ferron et al., 2009; Ferron, Farmer et al. 2010; Ferron, Owens et al., 2010).

Data Generation

Data was generated based on Van den Noortgate and Onghena's (2008) raw data, three level, single-case meta-analytic model shown in Equations 24 through 28. The raw data method was chosen as it is the most basic model and therefore was the most logical model to first evaluate. At the first level, an outcome (*y*) was modeled on measurement occasion *i* for participant *j* of study *k* (y_{ijk}) as a linear function of a single-predictor, *phase:*

$$y_{ijk} = \pi_{0jk} + \pi_{1jk} phase + e_{ijk} \qquad (24)$$

where *phase* was a dichotomous variable indicating whether a measurement occasion or observation occurred during baseline or treatment phase, π_{0jk} was the level of the outcome

during baseline for participant j from study k, π_{1jk} was the treatment effect for participant j from study k, and e_{ijk} was within-phase error variance. This within-phase participant model was consistent with the multilevel modeling application presented by Van den Noortgate and Onghena (2008). In addition, it was the most basic interrupted time-series model (e.g., no trends or changes in trends); therefore it was the most logical model for an initial study into the three level meta-analytic modeling of single-case data. If estimation problems occurred in the simplest model one would suspect that those same problems would likely occur in any further complex model. Errors for the within participant model (e_{ijk}) were generated using the ARMASIM function in SAS version 9.2 (SAS Institute, 2008) with a variance of (σ^2) of 1.0 and autocorrelation values of 0, .2, or .4, as previously discussed.

At the second level, the variation across participants was modeled using the following equations:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \tag{25}$$

and

$$\pi_{1jk} = \beta_{10k} + r_{1jk} \tag{26}$$

where the fixed effects were β_{00k}, the average baseline level for study k, and β_{10k}, the average treatment effect for study k, and the error terms are r_{0jk} and r_{1jk} that allowed variation in both baseline levels and treatment effects among participants. Level-2 errors were generated from a normal distribution using the RANNOR random number generator in SAS version 9.2 (SAS Institute Inc., 2008). The variance of the level-2 errors were defined based on the previously discussed levels of .2 or 2 and the covariance between

65

r_{0jk} and r_{1jk} was set to 0. The covariance between the level-2 errors was set to zero which was consistent with past simulation research (Ferron et al., 2009; Ferron, Farmer et al. 2010; Ferron, Owens et al., 2010), as well as Van den Noortgate and Onghena's (2003a, 2007) application of multilevel modeling to single-case data.

At the third level, the variation across studies was modeled using the following equations:

$$\beta_{00k} = \gamma_{000} + u_{00k} \tag{27}$$

and

$$\beta_{10k} = \gamma_{100} + u_{10k} \tag{28}$$

where the fixed effects were γ_{000}, the overall average baseline level, and γ_{100}, the overall average treatment effect, and the error terms are u_{00k} and u_{10k}, which allowed variation in both the overall average baseline level and overall average treatment effect among studies. Level-3 errors were generated from a normal distribution using the RANNOR random number generator in SAS version 9.2 (SAS Institute Inc., 2008). The fixed effects (γ_{000} and γ_{100}) were set to 1.0. The variance of the level-3 errors were defined based on the previously discussed levels of .05 or .5 and the covariance between u_{00k} and u_{10k} was set to 0. The covariance between the level-3 errors was set to zero which was consistent with past simulation research (Ferron et al., 2009; Ferron, Farmer et al. 2010; Ferron, Owens et al., 2010), as well as Van den Noortgate and Onghena's (2003a, 2007) application of multilevel modeling to single-case data.

Analysis of Each Simulated Meta-Analytic Data Set

Each data set was analyzed using the same model that was used for data generation (see Equations 24 - 28). The three level model was estimated using restricted

maximum likelihood (REML) via PROC MIXED with the Kenward-Roger degrees of freedom method in SAS version 9.2 (SAS Institute Inc., 2008). In addition, a first order auto-regressive model for the level-1 errors was specified. Based on the current model, the treatment effect was modeled as a change in level, and estimates were obtained for autocorrelation, variance within participants, variance in baseline levels across participants and studies, and variance in treatment effects across participants and studies.

The estimated models were checked for consistency with data generation. Several checks were used to verify the accuracy of the simulation program by running the program for a small number of replications. The vectors created during data generation were examined for consistency with data specifications, output data sets from the PROC MIXED statements were created to ensure the intended models were being analyzed and the summary statistics from those data sets were compared to the output data sets for accuracy.

Analysis to Estimate Bias of the Point Estimates, Confidence Interval Coverage and Confidence Interval Width

For each of the 108 combinations of the five independent variables, bias of the fixed effects (γ_{000} and γ_{100}) and the variance components ($\tau_{\beta00}$, $\tau_{\beta11}$, $\tau_{\pi00}$, $\tau_{\pi11}$, σ^2, and $\hat{\rho}$) and confidence interval coverage and width of the fixed effects and variance components were the dependent variables. Bias was calculated as the average difference between the known parameter value and the estimated parameter value. More formally,

$$bias = \frac{\Sigma_{k=1}^{5000}\left(\hat{\gamma}_{1h} - \gamma_{1h}\right)}{5000} \tag{29}$$

where $\hat{\gamma}_{1h}$ was the estimated parameter from the h$^{\text{th}}$ simulated meta-analysis, and γ_{1h} was

the simulated parameter value from the h$^{\text{th}}$ simulated meta-analysis. Relative bias was

also calculated for those parameters whose known value was anything other than 1.0 so

that bias could be represented as a percentage of the known parameter value. More

formally,

$$relative\ bias = \frac{\Sigma_{k=1}^{5000} \left(\dfrac{\hat{\gamma}_{1h} - \gamma_{1h}}{\gamma_{1h}} \right)}{5000} \qquad (30)$$

where $\hat{\gamma}_{1h}$ was the estimated parameter from the h$^{\text{th}}$ simulated meta-analysis, and γ_{1h} was

the simulated parameter value from the h$^{\text{th}}$ simulated meta-analysis. Coverage was

calculated as the proportion of the 95% confidence interval that contained the parameter

value, and width was calculated as the average difference between the upper and lower

limits of the 95% confidence intervals. Bias, coverage, and width estimates were

calculated based on values that were summarized across all 5,000 replications.

Analyses to Examine Relationships Between Design Factors and Bias of the Point

Estimates, Confidence Interval Coverage, and Confidence Interval Width

Research Question One

Research Question One, evaluation of the bias of the fixed effect estimates from

the three level meta-analytic single-case model were addressed by examining box and

whisker plots to illustrate the distribution of the bias estimates of the fixed effects. In

addition, generalized linear modeling (GLM) was used to examine variability of each of

the bias estimates of the fixed effects as a function of the independent variables. Models

were built with the purpose of finding effects whose eta-squared values .06 or greater.

The effects size, eta-squared (η^2), was calculated to determine the proportion of variability associated with each effect. Those values were compared to Cohen's (1988) standards for interpreting eta-squared values with a small effect size having an η^2 = .01, .a medium effect size having an η^2 = .06, and a large effect size having an η^2 = .14 or greater. Each model was first created as a main effects only model. If this model explained 94% of the total variability then no further complex models were investigated. However, if less than 94% of the total variability was explained then interactions were included in the model. Two-way interactions were added to the model first followed by three-way and then four-way interactions until at least 94% of the variability was explained. Finally, line graphs were created to show bias estimates of the fixed effects as a function of the independent variables (both main effects and interactions) that had eta-squared values of .06 or higher.

Research Question Two

Research Question Two, evaluation of the confidence interval coverage of the fixed effect estimates from a three level meta-analytic single-case model, were addressed by examining box and whisker plots to illustrate the distribution of the confidence interval coverage estimates of the fixed effects. In addition, GLM was used to examine variability of each of the confidence interval coverage estimates of the fixed effects as a function of the independent variables. Models were built with the purpose of finding effects whose eta-squared values .06 or greater. The effects size, eta-squared (η^2), was calculated to determine the proportion of variability associated with each effect. Those values were compared to Cohen's (1988) standards for interpreting eta-squared values with a small effect size having an η^2 = .01, .a medium effect size having an η^2 = .06, and

69

a large effect size having an $\eta^2 = .14$ or greater. Each model was first created as a main

effects only model. If this model explained 94% of the total variability then no further

complex models were investigated. However, if less than 94% of the total variability was

explained then interactions were included in the model. Two-way interactions were

added to the model first followed by three-way and then four-way interactions until at

least 94% of the variability was explained. Finally, line graphs were created to show

confidence interval coverage estimates of the fixed effects as a function of the

independent variables (both main effects and interactions) that had eta-squared values of

.06 or higher.

Research Question Three

Research Question Three, evaluation of the confidence interval width of the fixed

effect estimates from a three level meta-analytic single-case model, were addressed by

examining box and whisker plots to illustrate the distribution of the confidence interval

width estimates of the fixed effects. In addition, GLM was used to examine variability of

each of the confidence interval precision estimates of the fixed effects as a function of the

independent variables. Models were built with the purpose of finding effects whose eta-

squared values .06 or greater. The effects size, eta-squared (η^2), was calculated to

determine the proportion of variability associated with each effect. Those values were

compared to Cohen's (1988) standards for interpreting eta-squared values with a small

effect size having an $\eta^2 = .01$, .a medium effect size having an $\eta^2 = .06$, and a large effect

size having an $\eta^2 = .14$ or greater. Each model was first created as a main effects only

model. If this model explained 94% of the total variability then no further complex

models were investigated. However, if less than 94% of the total variability was

70

explained then interactions were included in the model. Two-way interactions were added to the model first followed by three-way and then four-way interactions until at least 94% of the variability was explained. Finally, line graphs were created to show confidence interval width estimates of the fixed effects as a function of the independent variables (both main effects and interactions) that had eta-squared values of .06 or higher.

Research Question Four

Research Question Four, evaluation of the bias of the variance components from a three level meta-analytic single-case model, were addressed by examining box and whisker plots to illustrate the distribution of the bias estimates of the variance components. In addition, GLM was used to examine variability of each of the bias estimates of the variance components as a function of the independent variables. Models were built with the purpose of finding effects whose eta-squared values .06 or greater. The effects size, eta-squared (η^2), was calculated to determine the proportion of variability associated with each effect. Those values were compared to Cohen's (1988) standards for interpreting eta-squared values with a small effect size having an $\eta^2 = .01$, .a medium effect size having an $\eta^2 = .06$, and a large effect size having an $\eta^2 = .14$ or greater. Each model was first created as a main effects only model. If this model explained 94% of the total variability then no further complex models were investigated. However, if less than 94% of the total variability was explained then interactions were included in the model. Two-way interactions were added to the model first followed by three-way and then four-way interactions until at least 94% of the variability was explained. Finally, line graphs were created to show bias estimates of the variance

71

components as a function of the independent variables (both main effects and interactions) that had eta-squared values of .06 or higher.

Research Question Five

Research Question Five, evaluation of the confidence interval coverage of the variance components from a three level meta-analytic single-case model, was addressed by examining box and whisker plots to illustrate the distribution of the confidence interval coverage estimates of the variance components. In addition, GLM was used to examine variability of each of the confidence interval coverage estimates of the variance components as a a function of the independent variables. Models were built with the purpose of finding effects whose eta-squared values .06 or greater. The effects size, eta-squared (η^2), was calculated to determine the proportion of variability associated with each effect. Those values were compared to Cohen's (1988) standards for interpreting eta-squared values with a small effect size having an $\eta^2 = .01$, .a medium effect size having an $\eta^2 = .06$, and a large effect size having an $\eta^2 = .14$ or greater. Each model was first created as a main effects only model. If this model explained 94% of the total variability then no further complex models were investigated. However, if less than 94% of the total variability was explained then interactions were included in the model. Two-way interactions were added to the model first followed by three-way and then four-way interactions until at least 94% of the variability was explained. Finally, line graphs were created to show confidence interval coverage estimates of the variance components as a function of the independent variables (both main effects and interactions) that had eta-squared values of .06 or higher.

72

Research Question Six

Research Question Six, evaluation of the confidence interval width of the variance components from a three level meta-analytic single-case model, was addressed by examining box and whisker plots to illustrate the distribution of the confidence interval width estimates of the variance components. In addition, GLM was used to examine variability of each of the confidence interval width estimates of the variance components as a a function of the independent variables. Models were built with the purpose of finding effects whose eta-squared values .06 or greater. The effects size, eta-squared (η^2), was calculated to determine the proportion of variability associated with each effect. Those values were compared to Cohen's (1988) standards for interpreting eta-squared values with a small effect size having an $\eta^2 = .01$, .a medium effect size having an $\eta^2 = .06$, and a large effect size having an $\eta^2 = .14$ or greater. Each model was first created as a main effects only model. If this model explained 94% of the total variability then no further complex models were investigated. However, if less than 94% of the total variability was explained then interactions were included in the model. Two-way interactions were added to the model first followed by three-way and then four-way interactions until at least 94% of the variability was explained. Finally, line graphs were created to show confidence interval width estimates of the variance components as a function of the independent variables (both main effects and interactions) that had eta-squared values of .06 or higher.

CHAPTER FOUR: RESULTS

This chapter presents the results organized in the order of the research questions. This chapter begins by describing how the results were examined and then presents the results in two sections, the fixed effects and the variance components. Each section presents each outcome measure (bias of the point estimate, confidence interval coverage, and confidence interval width) with the first section comprising the first three research questions and the second section comprising the last three research questions. The following research questions were addressed:

1. To what extent are the fixed effect estimates from a three level meta-analytic single-case model biased as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

2. To what extent does the confidence interval coverage of the fixed effect estimates from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

3. To what extent does the confidence interval width of the fixed effect estimates from a three level meta-analytic single-case model vary as a function of specific

design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

4. To what extent are the variance components from a three level meta-analytic single-case model biased as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

5. To what extent does the confidence interval coverage of the variance components from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

6. To what extent does the confidence interval width of the variance components from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

There were 108 conditions simulated using the five factors of this Monte Carlo study. The five factors were the number of primary studies per meta-analysis (10, 30, and 80), modal number of participants per primary study (small [mode = 4] and large [mode = 8]), modal series length per primary study (small [mode = 10], medium [mode = 20], and large [mode = 30]), level of autocorrelation (0, .2, and .4), and variances of the error

terms (most of the variance at level-1 [$\sigma^2 = 1$, $\tau_{\pi00} = \tau_{\pi10} = .2$, and $\tau_{\beta00} = \tau_{\beta10} = .05$] and

most of the variance at level-2 [$\sigma^2 = 1$, $\tau_{\pi00} = \tau_{\pi10} = 2$, and $\tau_{\beta00} = \tau_{\beta10} = .5$]). This yielded

3 (number of primary studies per meta-analysis) X 2 (modal number of participants per

primary study) X 3 (modal series length per primary study) X 3 (level of autocorrelation)

X 2 (variances of the error terms) = 108 conditions.

First, the dependent variables, bias of the point estimates, confidence interval

coverage, and confidence interval width were evaluated for both the fixed effects and the

variance components. In addition, an index of relative bias was calculated for all

parameter estimates whose known value was anything other than a value of 1.0. This

was accomplished by creating box plots, across all conditions, for each dependent

variable. Then, the results of the simulation were analyzed using PROC GLM in SAS for

both the fixed effects and the variance components such that the dependent variables

were bias, relative bias (where appropriate), confidence interval coverage, and confidence

interval width and the independent variables were the five factors. Models were built

with the purpose of finding effects whose eta-squared values were .06 or greater. The

effect size, eta-squared (η^2), was calculated to measure the degree of association

between the independent variables main effects and the dependent variables along with

the two-way, three-way, and four-way interaction effects between independent variables

and the dependent variables. Eta-squared is the proportion of variability in each of the

outcome measures associated with each effect in this simulation study. It is calculated as

the ratio of the effect variance (SS_{effect}) to the total variance (SS_{total}).

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \tag{30}$$

The calculated eta-squared values were compared to Cohen's (1988) standards for interpreting eta-squared values with a small effect size having an $\eta^2 = .01$, .a medium effect size having an $\eta^2 = .06$, and a large effect size having an $\eta^2 = .14$ or greater. Each model was first created as a main effects only model. If this model explained 94% of the total variability then no further complex models were investigated. However, if less than 94% of the total variability was explained then interactions were included in the model. Two-way interactions were added to the model first followed by three-way and then four-way interactions until at least 94% of the variability was explained. Finally, line graphs were created to show bias and/or relative bias of the point estimates, confidence interval coverage, and confidence interval width estimates of the fixed effects and variance components as a function of the independent variables (both main effects and interactions) that had eta-squared values of .06 or higher.

Fixed Effects

The fixed effects are comprised of γ_{000}, the overall average baseline level, and γ_{100}, the overall average treatment effect. The first research question involves the extent to which the fixed effects are biased as a function of the five factors used in this simulation study. The second research question involves the extent to which the confidence interval coverage of the fixed effects varied as a function of the five factors used in this simulation study. The third research question involves the extent to which the confidence interval width of the fixed effects varied as function of the five factors used in this simulation study.

Bias

The distribution of bias values for each fixed effect is illustrated in box plots in Figure 6. The overall average baseline level (γ_{000}) had bias values close to zero with an average bias value of 0.000 ($SD = 0.002$) and a range of values from -0.005 to 0.005. In addition, the overall average treatment effect (γ_{100}) had bias values close to zero with a mean of 0.000 ($SD = 0.001$) and values ranging from -0.003 to 0.009. As indicated by the results, there was limited variation in both of the fixed effects and none of the bias estimates exceeded 1% of the known parameter values (recall that all known fixed effect parameter values were set to 1.0). Therefore, any further exploration was unwarranted.
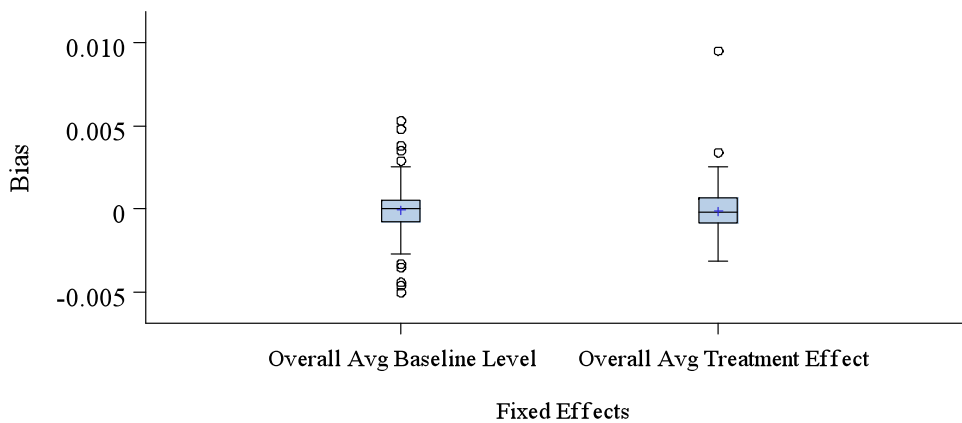
Figure 6. Box plots showing the distribution of bias estimates for each fixed effect in the three level model.

Confidence Interval Coverage

The distribution of confidence interval coverage rates for each fixed effect is illustrated in box plots in Figure 7. The overall average baseline level (γ_{000}) had confidence interval coverage rates that tended to slightly overcover with values that ranged from a high of .973 to a low of .951, with a mean of .961 ($SD = 0.005$). Similarly,

the overall average treatment effect (γ_{100}) had confidence interval coverage rates that ranged from a high of .971 to a low of .951, with a mean of .960 ($SD = 0.005$).

Figure 7. Box plots showing the distribution of confidence interval coverage rates for each fixed effect in the three level model

Overall average baseline level. Variation in confidence interval coverage rates of the overall average baseline level was explored by modeling confidence interval coverage with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms), all possible two-way interactions involving the number of primary studies per meta-analysis, all possible three-way interactions involving the number of primary studies per meta-analysis and three four-way interactions. One of the four-way interactions involved the number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, and level of autocorrelation. The next four-way interaction involved the number of primary studies per meta-analysis, modal number of participants per primary study, level of autocorrelation, and variances of the error terms, and the final

four-way interaction involved the number of primary studies per meta-analysis, modal

series length per primary study, level of autocorrelation, and variances of the error terms.

This model explained 96% of the variability in the confidence interval coverage rates of

the overall average baseline level. Eta-squared (η^2) values for each of the main effects

and interactions are in Table 3.

Table 3

Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval Coverage of the Overall Average Baseline Level

	η^2
Number of Primary Studies	.762
Modal Number of Participants	.005
Modal Series Length	.001
Autocorrelation	.012
Variances of the Error Terms	.004
Number of Primary Studies*Modal Number of Participants	.000
Number of Primary Studies*Modal Series Length	.013
Number of Primary Studies*Autocorrelation	.005
Number of Primary Studies*Variances of the Error Terms	.001
Number of Primary Studies*Modal Series Length*Autocorrelation	.026
Number of Primary Studies*Modal Series Length*Variances of the Error Terms	.005
Number of Primary Studies*Modal Series Length*Modal Number of Participants	.012
Number of Primary Studies*Modal Number of Participants*Autocorrelation	.008
Number of Primary Studies*Autocorrelation*Variances of the Error Terms	.028
Number of Primary Studies*Variances of the Error Terms*Modal Number of Participants	.000
Number of Primary Studies*Modal Number of Participants*Modal Series Length*Autocorrelation	.036
Number of Primary Studies*Modal Number of Participants* Autocorrelation * Variances of the Error Terms	.014
Number of Primary Studies*Modal Series Length*Autocorrelation * Variances of the Error Terms	.027
Total Explained	.959

In order to explore these effects further line graphs were created for those main effects and/or interactions whose eta-squared values exceeded the pre-established standard of Cohen's (1988) medium effect size criteria of $\eta^2 = .06$ or greater. Therefore, the only effect that met this standard was the main effect of the number of primary studies per meta-analysis with an $\eta^2 = .76$. The 95% confidence interval coverage rates of the overall average baseline level as a function of the number of primary studies per meta-analysis (see Figure 8) illustrated that as the number of primary studies per meta-analysis increased from 10 to 30 to 80 the closer the coverage rates came to .95, with means of .968 ($SD = 0.002$), .960 ($SD = 0.003$), and .956 ($SD = 0.003$), respectively.
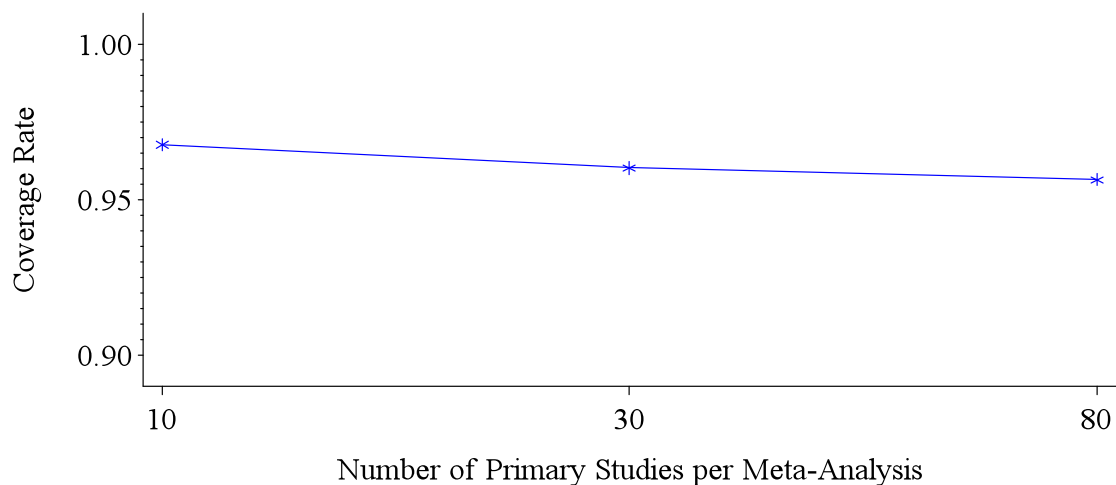
Figure 8. Line graph showing the estimated confidence interval coverage rates for the overall average baseline level as a function of the number of primary studies per meta-analysis.

Overall average treatment effect. Variation in confidence interval coverage rates of the overall average treatment effect was explored by modeling confidence interval coverage with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per

primary study, level of autocorrelation, and variances of the error terms), all possible

two-way interactions involving the number of primary studies per meta-analysis, all

possible three-way interactions involving the number of primary studies per meta-

analysis and three four way interactions. One of the four-way interactions involved the

number of primary studies per meta-analysis, modal number of participants per primary

study, modal series length per primary study, and level of autocorrelation. The next four-

way interaction involved the number of primary studies per meta-analysis, modal number

of participants per primary study, level of autocorrelation, and the variances of the error

terms, and the final four-way interaction included the number of primary studies per

meta-analysis, modal series length per primary study, level of autocorrelation, and

variances of the error terms. This model explained 97% of the variability in the

confidence interval coverage rates of the overall average treatment effect. Eta-squared

(η^2) values for each of the main effects and interactions are in Table 4.

Table 4

Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval Coverage of the Overall Average Treatment Effect

	η^2
Number of Primary Studies	.701
Modal Number of Participants	.009
Modal Series Length	.013
Autocorrelation	.000
Variances of the Error Terms	.000
Number of Primary Studies*Modal Number of Participants	.001
Number of Primary Studies*Modal Series Length	.008
Number of Primary Studies*Autocorrelation	.015
Number of Primary Studies*Variances of the Error Terms	.018
Number of Primary Studies*Modal Series Length*Autocorrelation	.044
Number of Primary Studies*Modal Series Length*Variances of the Error Terms	.022
Number of Primary Studies*Modal Series Length*Modal Number of Participants	.009
Number of Primary Studies*Modal Number of Participants*Autocorrelation	.009
Number of Primary Studies*Autocorrelation*Variances of the Error Terms	.024
Number of Primary Studies*Variances of the Error Terms*Modal Number of Participants	.003
Number of Primary Studies*Modal Number of Participants*Modal Series Length*Autocorrelation	.041
Number of Primary Studies*Modal Number of Participants*Autocorrelation*Variances of the Error Terms	.030
Number of Primary Studies*Modal Series Length*Autocorrelation*Variances of the Error Terms	.021
Total Explained	.968

Similar to the results for the overall average baseline level, the only effect that met the standard of a medium effect size or greater was the main effect of the number of primary studies per meta-analysis with an $\eta^2 = .70$. The 95% confidence interval coverage rates of the overall average treatment effect as a function of the number of primary studies per meta-analysis (see Figure 9) illustrated that as the number of primary studies per meta-analysis increased the closer the confidence interval coverage rates came to .95 with means of .966 ($SD = 0.002$), .960 ($SD = .003$), and .956 ($SD = .002$), respectively.
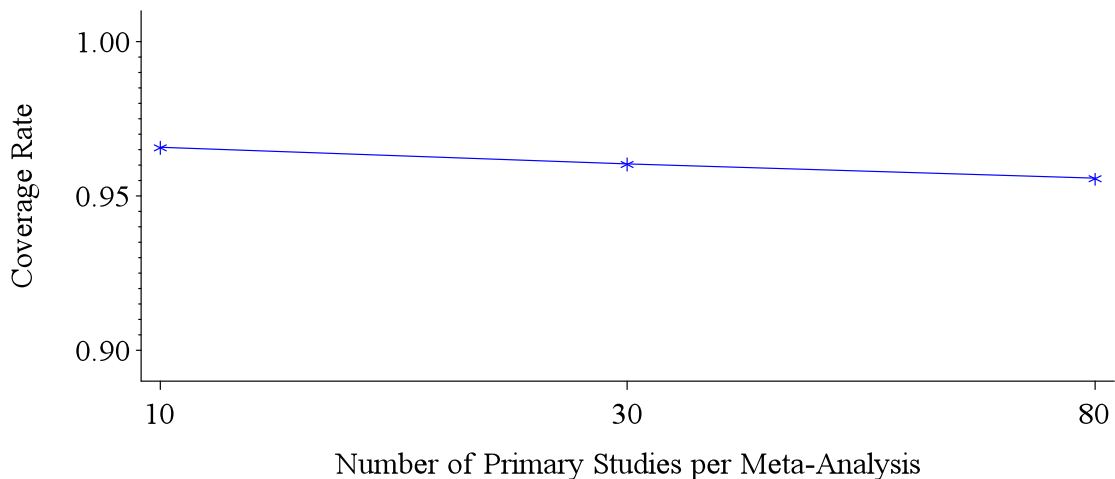
Figure 9. Line graph showing the estimated confidence interval coverage rates for the overall average treatment effect as a function of the number of primary studies per meta-analysis.

Confidence Interval Width

The box plot illustrating the distribution of the confidence interval width estimates for each fixed effect is presented in Figure 10. The confidence interval width estimates for the overall average baseline level (γ_{000}) ranged from a low of 0.099 to a high of 1.132, with a mean of 0.428 ($SD = 0.291$). Similarly, the confidence interval

width estimates for the overall average treatment effect (γ_{100}) ranged from a low of 0.114

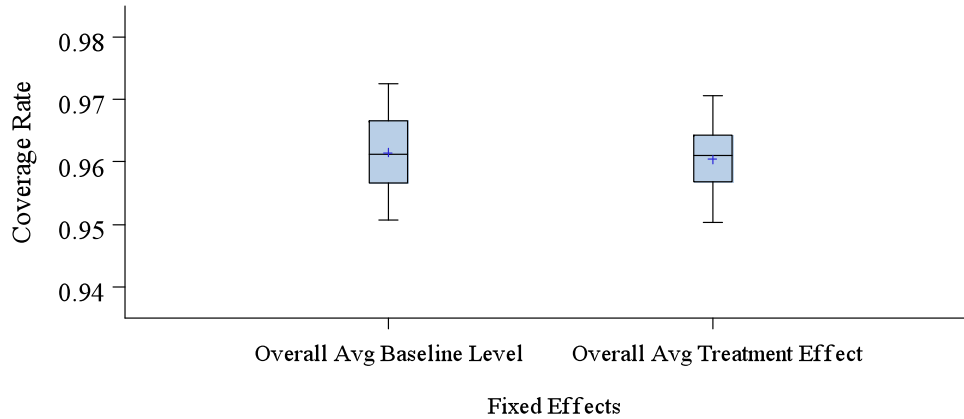to a high of 1.174, with a mean of 0.459 (*SD* = 0.293).
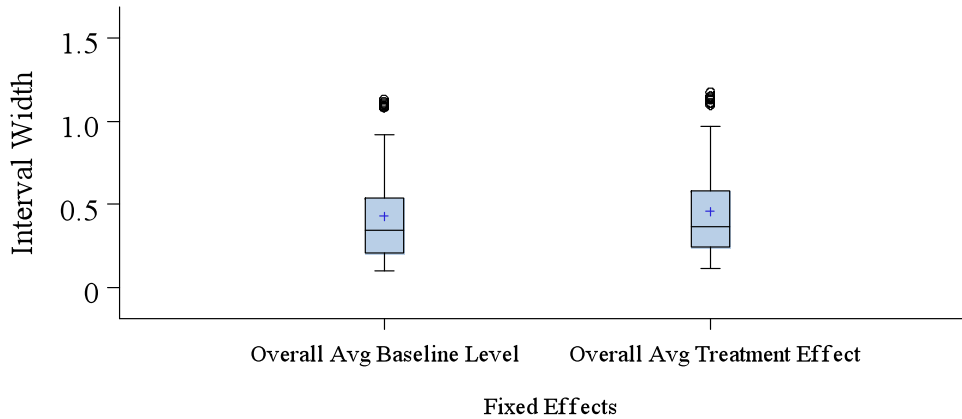
Figure 10. Box plots showing the distribution of confidence interval width estimates for each fixed effect in the three level model.

Overall average baseline level. Variation in confidence interval width estimates

of the overall average baseline level was explored by modeling confidence interval width

with the five main effects (number of primary studies per meta-analysis, modal number

of participants per primary study, modal series length per primary study, level of

autocorrelation, and variances of the error terms) and a two-way interaction involving the

number of primary studies per meta-analysis and the variances of the error terms. This

model explained 99% of the variability in the confidence interval width estimates of the

overall average baseline level. Eta-squared (η^2) values for each of the main effects and

interaction are in Table 5.

Table 5

Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval Width of the Overall Average Baseline Level

	η^2
Number of Primary Studies	.472
Modal Number of Participants	.023
Modal Series Length	.000
Autocorrelation	.001
Variances of the Error Terms	.402
Number of Participants * Variances of the Error Terms	.089
Total Explained	.987

To further examine these effects a line graph was created for those effects whose eta-squared values exceeded the pre-established standard of a medium effect size or greater. Therefore, confidence interval width was modeled as a function of the number of primary studies per meta-analysis ($\eta^2 = .47$), variances of the error terms ($\eta^2 = .40$), and their interaction ($\eta^2 = .09$). This model explained 96% of the variance in confidence interval width estimates of the overall average baseline level. As the number of primary studies per meta-analysis increased from 10 to 30 to 80 the confidence interval width decreased, with means of 0.696 (*SD* = 0.316), 0.368 (*SD* = 0.165), and 0.220 (*SD* = 0.098), respectively. Conversely, as the variances of the error terms shifted from most of the variance at level-1 (or less variance at level-2 and level-3) to most of the variance at level-2 the confidence interval widths increased, with means of 0.244 (*SD* = 0.119) and 0.612 (*SD* = 0.297), respectively. The graph (see Figure 11) indicates that when the number of primary studies per meta-analysis was 30 or 80 the confidence interval widths slightly increased when the variance of the error terms shifted from most of the variance

at level-1 to most of the variance at level-2. However, when the number of primary studies was 10 there was a greater increase of the estimated confidence interval widths when the variances of the error terms shifted from most of the variance at level-1 ($M = 0.396$, $SD = 0.053$) to most of the variance at level-2 ($M = 0.997$, $SD = 0.107$). In addition, confidence interval widths were smallest when the number of primary studies per meta-analysis was 80 and most of the variance of the error terms was at level-1 (or less variance at level-2 and level-3).
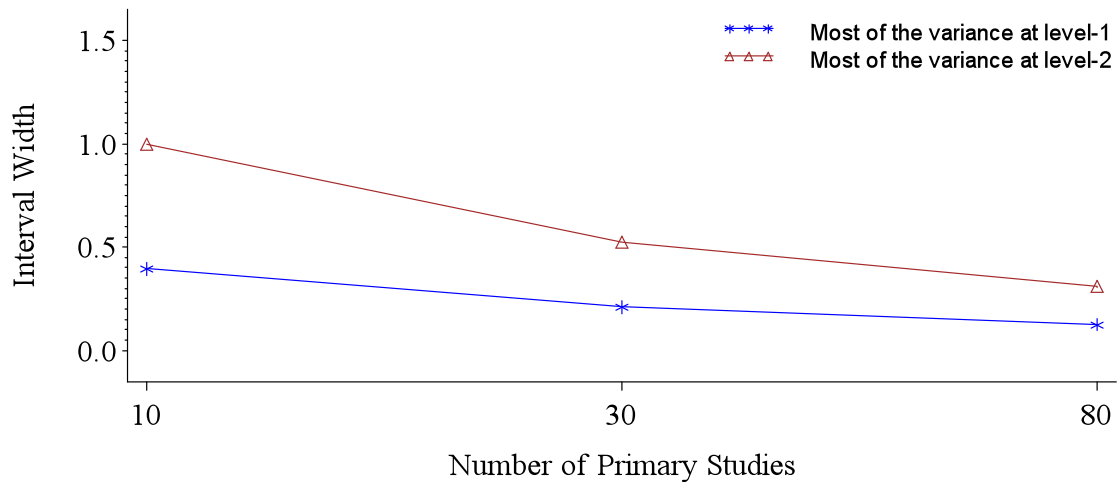
Figure 11. Line graph showing the confidence interval width estimates of the overall average baseline level as a function of the variances of the error terms for each level of the number of primary studies per meta-analysis.

Overall average treatment effect. Variation in confidence interval width estimates of the overall average treatment effect was explored by modeling confidence interval width with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms) and a two-way interaction involving the number of primary studies per meta-analysis and the variances of the error terms. This model explained 99% of the variability in the confidence interval width

estimates of the overall average treatment effect. Eta-squared (η^2) values for each of the

main effects and interaction are in Table 6.

Table 6

Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval Width of the Overall Average Treatment Effect

	η^2
Number of Primary Studies	.534
Modal Number of Participants	.029
Modal Series Length	.002
Autocorrelation	.003
Variances of the Error Terms	.350
Number of Participants * Variances of the Error Terms	.078
Total Explained	.996

To explore the variation between confidence interval width estimates of the

overall average treatment effect a line graph was created that modeled confidence interval

width as a function of the number of primary studies per meta-analysis ($\eta^2 = .53$),

variances of the error terms ($\eta^2 = .35$), and their interaction ($\eta^2 = .08$). This model

explained 96% of the variance in confidence interval width estimates of the overall

average treatment effect. Similar to the confidence interval width estimates of the overall

average baseline level, the results indicated that as the number of primary studies

increased from 10 to 30 to 80, the average width decreased from 0.746 (*SD* = 0.299) to

0.236 (*SD* = 0.093), respectively. Conversely, as the variances of the error terms shifted

from most of the variance at level-1 (or less variance at level-2 and level-3) to most of the

variance at level-2 the confidence interval widths increased, with means of 0.287 (*SD* =

0.140) and 0.632 (*SD* = 0.306), respectively. The graph (see Figure 12) indicates that

89

when the number of primary studies per meta-analysis was 30 or 80 the confidence interval widths slightly increased when the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. However, when the number of primary studies was 10 there was a greater increase of the estimated confidence interval widths when the variances of the error terms shifted from most of the variance at level-1 ($M = 0.465$, $SD = 0.063$) to most of the variance at level-2 ($M = 1.028$, $SD = 0.109$). In addition, confidence interval widths were smallest when the number of primary studies per meta-analysis was 80 and most of the variance of the error terms was at level-1 (or less variance at level-2 and level-3).
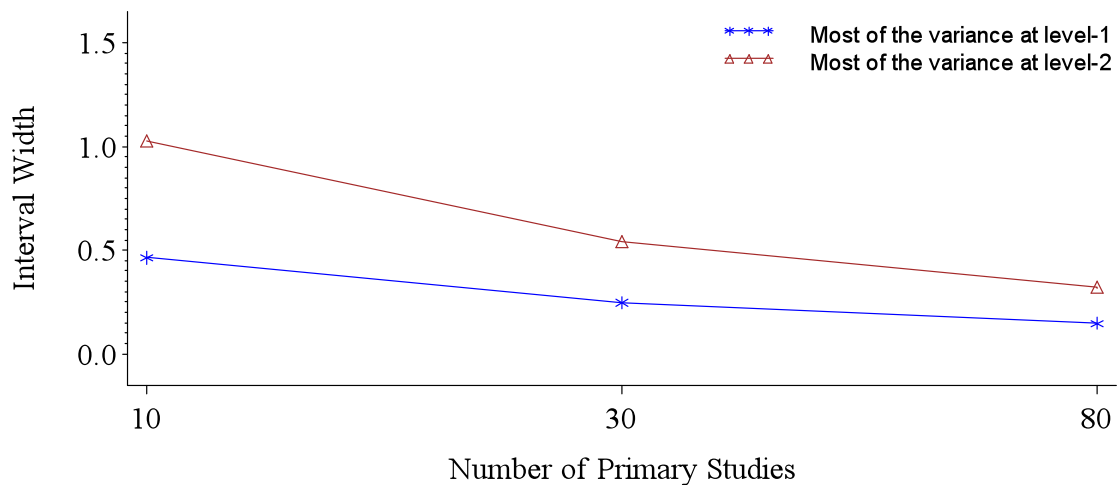
Figure 12. Line graph showing the estimated confidence interval width of the overall average treatment effect as a function of the variances of the error terms for each level of the number of primary studies per meta-analysis.

Variance Components

The variance components are comprised of $\tau_{\beta00}$, between study variance in the overall average baseline level, $\tau_{\beta10}$, between study variance in the overall average treatment effect, $\tau_{\pi00}$ between person within study variance in the average baseline level,

$\tau_{\pi 10}$, between person within study variance in the average treatment effect, σ^2, within person residual variance, and $\hat{\rho}$, amount of estimated autocorrelation. The fourth research question involves the extent to which the variance components are biased as a function of the five factors used in this simulation study. The fifth research question involves the extent to which the confidence interval coverage of the variance components varied as a function of the five factors used in this simulation study. The sixth and final research question involves the extent to which the confidence interval width of the variance components varied as function of the five factors used in this simulation study.

Bias

The distribution of bias values for each variance component is illustrated in box plots in Figures 13 - 15. Both level-3 variance components ($\tau_{\beta 00}$ and $\tau_{\beta 10}$) tended to be underestimated (see Figure 13). Between study variance in the overall average baseline level ($\tau_{\beta 00}$) was biased with negative bias values ranging from -0.477 to -0.031 and a mean of -0.241 (*SD* = 0.201). In addition, between study variance in the overall average treatment effect ($\tau_{\beta 10}$) was biased with negative bias values ranging from -0.474 to -0.024 and a mean of -0.237 (*SD* = 0.201).
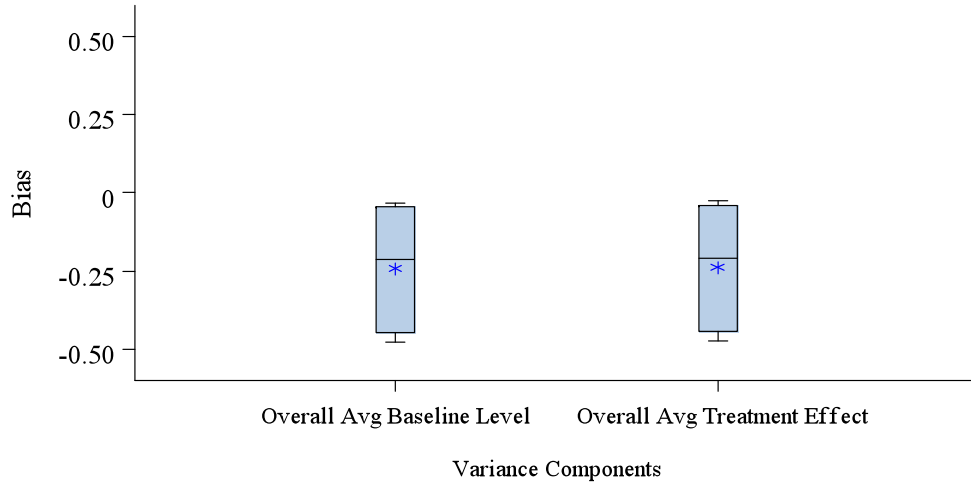
Figure 13. Box plots showing the distribution of bias estimates for each level-3 variance component in the three level model.

Conversely, the level-2 variance components ($\tau_{\pi00}$ and $\tau_{\pi10}$) both tended to be overestimated (see Figure 14). Between person within study variance in the average baseline level ($\tau_{\pi00}$) was biased with positive bias values ranging from 0.033 to 0.479 and an average bias value of 0.243 (*SD* = 0.202). Similarly, between person within study variance in the average treatment effect ($\tau_{\pi10}$) had positive bias values with a mean of 0.238 (*SD* = 0.201) and values ranging from 0.027 to 0.476.
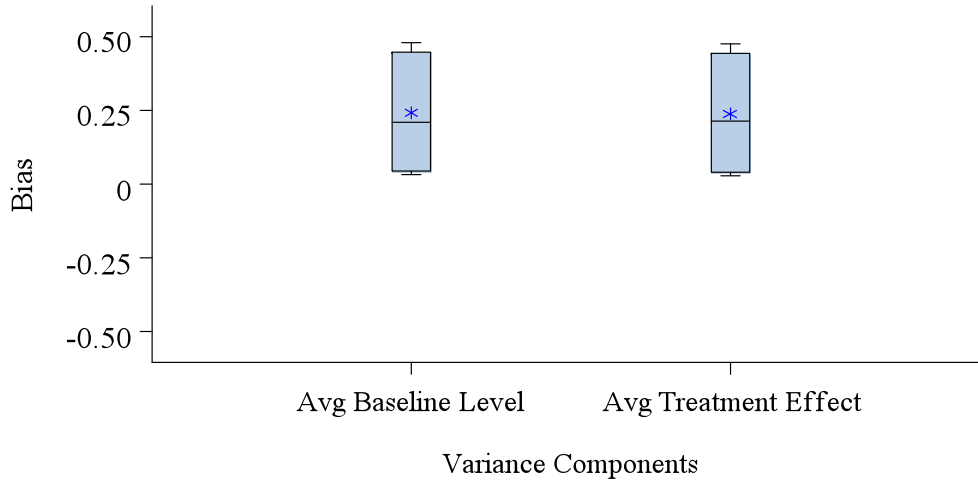
Figure 14. Box plots showing the distribution of bias estimates for each level-2 variance component in the three level model.

The within person residual variance (σ^2) also tended to be slightly overestimated with an average bias value of 0.078 (*SD* = 0.082) and values ranging from -0.001 to 0.194. Recall that the population value for the within person residual variance was 1.0, thus an average bias estimate of 0.078 represents 8% of the average parameter value. The amount of estimated autocorrelation ($\hat{\rho}$) had bias values close to zero with a mean of 0 (*SD* = 0.001) and values ranging from -0.002 to 0.001. Figure 15 illustrates the distribution of bias values for the residual variance and the amount of estimated autocorrelation. As indicated by the results, there was limited variation in the amount of estimated autocorrelation and the bias estimate did not exceed 1% of the known parameter value (see Relative Bias results section). Therefore, any further exploration of the amount of estimated autocorrelation was unwarranted.
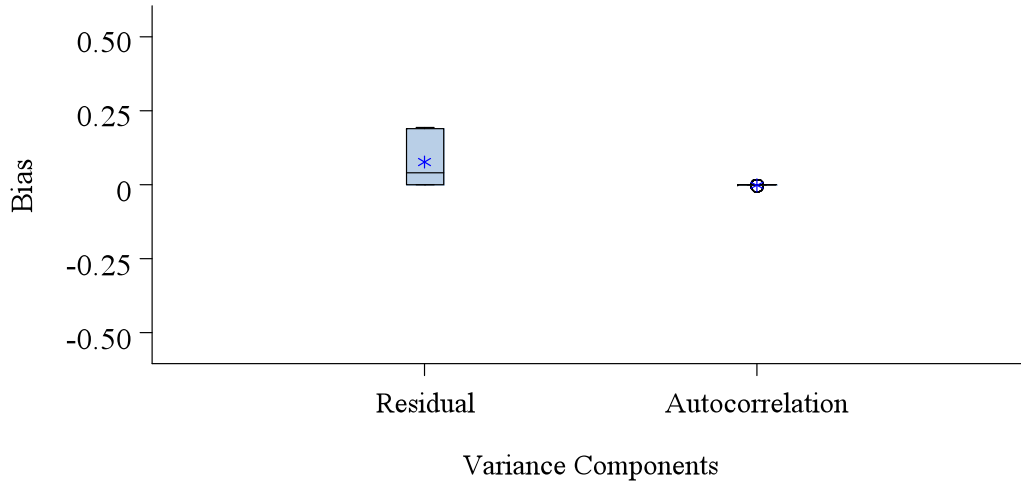
Figure 15. Box plots showing the distribution of bias estimates for the within person residual variance and amount of estimated autocorrelation in the three level model.

Between study variance in the overall average baseline level. Variation in the bias estimates of the between study variance in the overall average baseline level was explored by modeling bias with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms). This model explained 99% of the variability in the bias estimates of the between study variance in the overall average baseline level. Eta-squared (η^2) values for each of the main effects are in Table 7.

94

Table 7

Eta-squared Values (η^2 *) for Association of Design Factors with Bias of the Between Study Variance in the Overall Average Baseline Level*

	η^2
Number of Primary Studies	.005
Modal Number of Participants	.001
Modal Series Length	.000
Autocorrelation	.000
Variances of the Error Terms	.989
Total Explained	.995

In order to further explore these effects a line graph was created for the main effect of variances in the error terms ($\eta^2 = .99$), as it was the only effect that met the pre-established standard of a medium effect size or greater. The bias estimates of the between study variance in the overall average baseline level as a function of the variances of the error terms (see Figure 16) illustrated that as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the parameter estimates decreased to become increasingly underestimated and progressively more biased with mean bias estimates of -0.042 (*SD* = 0.004) to -0.440 (*SD* = 0.030), respectively.
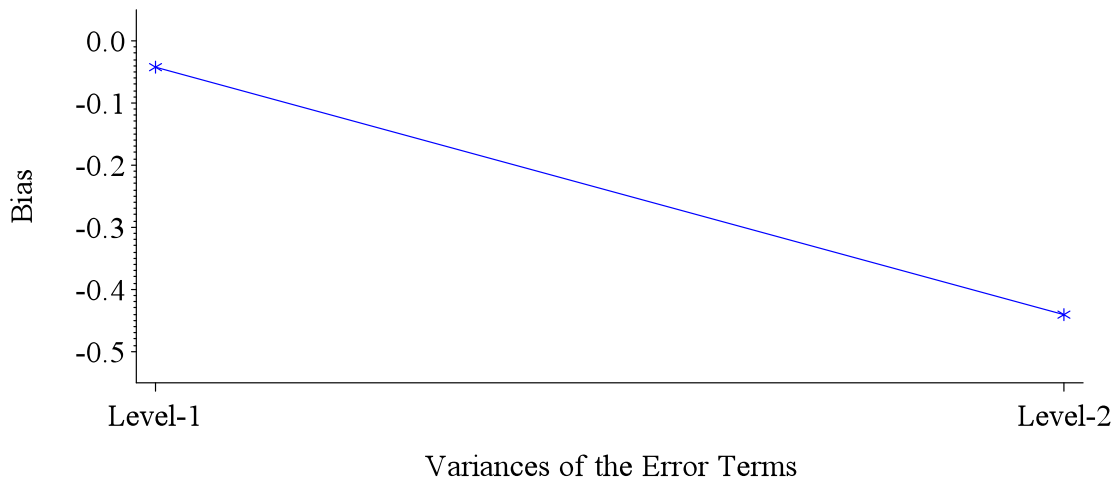
Figure 16. Line graph showing the bias estimates for the between study variance in the overall average baseline level as a function of the variances of the error terms. Level-1 = most of the variance at level-1; Level-2 = most of the variance at level-2.

Between study variance in the overall average treatment effect. Variation in the bias estimates of the between study variance in the overall average treatment effect was explored by modeling bias with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms). This model explained 99% of the variability in the bias estimates of the between study variance in the overall average treatment effect. Eta-squared (η^2) values for each of the main effects are in Table 8.

Table 8

Eta-squared Values (η^2) for Association of Design Factors with Bias of the Between Study Variance in the Overall Average Treatment Effect

	η^2
Number of Primary Studies	.006
Modal Number of Participants	.001
Modal Series Length	.000
Autocorrelation	.000
Variances of the Error Terms	.988
Total Explained	.995

Similar to the bias estimates of the between study variance in the overall average baseline level, between study variance in the overall average treatment effect was further explored with a line graph for the main effect of variances in the error terms ($\eta^2 = .99$). The bias estimates of the between study variance in the overall average treatment effect as a function of the variances of the error terms (see Figure 17) illustrated that as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the parameter estimates decreased to become increasingly underestimated and progressively more biased with mean bias estimates of -0.039 (*SD* = 0.006) to -0.436 (*SD* = 0.031), respectively.
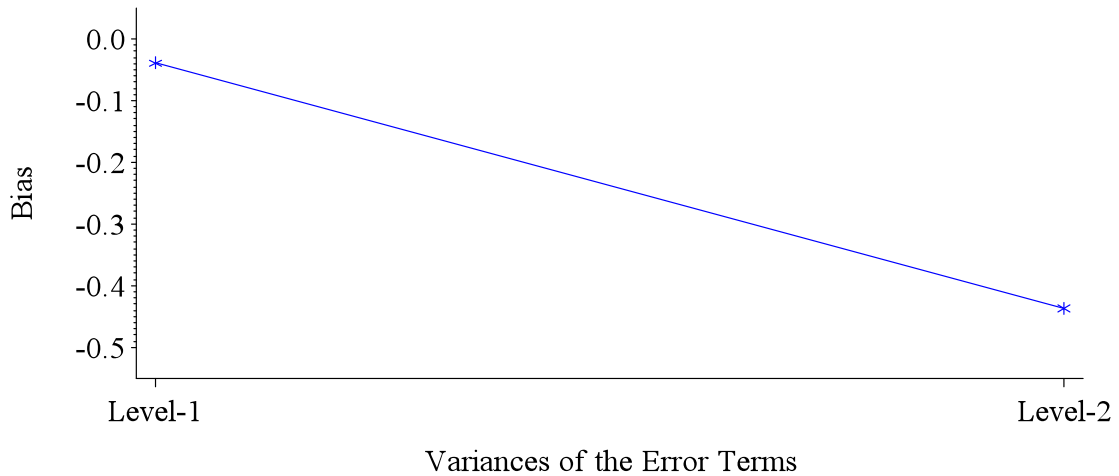
Figure 17. Line graph showing the bias estimates for the between study variance in the overall average treatment effect as a function of the variances of the error terms. Level-1 = most of the variance at level-1; Level-2 = most of the variance at level-2.

Between person within study variance in the average baseline level. Variation in the bias estimates of the between person within study variance in the average baseline level was explored by modeling bias with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms). This model explained 99% of the variability in the bias estimates of the between person within study variance in the average baseline level. Eta-squared (η^2) values for each of the main effects are in Table 9.

Table 9

Eta-squared Values (η^2) for Association of Design Factors with Bias of the Between Person Within Study Variance in the Average Baseline Level

	η^2
Number of Primary Studies	.004
Modal Number of Participants	.001
Modal Series Length	.000
Autocorrelation	.000
Variances of the Error Terms	.991
Total Explained	.996

To further explore these effects a line graph was created for the main effect variances of the error terms ($\eta^2 = .99$). The bias estimates of the between person within study variance in the average baseline level as a function of the variances of the error terms (see Figure 18) illustrated that as the variances of the error terms shifts from most of the variance at level-1 to most of the variance at level-2 the parameter estimates increased to become progressively overestimated and more biased with mean bias estimates of 0.042 (*SD* = 0.004) to 0.444 (*SD* = 0.026), respectively.

Figure 18. Line graph showing the bias estimates for the between person within study variance in the average baseline level as a function of the variances of the error terms. Level-1 = most of the variance at level-1; Level-2 = most of the variance at level-2.

Between person within study variance in the average treatment effect.

Variation in the bias estimates of the between person within study variance in the average treatment effect was explored by modeling bias with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms). This model explained 99% of the variability in the bias estimates of the between person within study variance in the average treatment effect. Eta-squared (η^2) values for each of the main effects are in Table 10.

Table 10

Eta-squared Values (η^2) for Association of Design Factors with Bias of the Between Person Within Study Variance in the Average Treatment Effect

	η^2
Number of Primary Studies	.006
Modal Number of Participants	.001
Modal Series Length	.000
Autocorrelation	.000
Variances of the Error Terms	.989
Total Explained	.995

Similar to the bias results for the between person within study variance in the average treatment effect, the only effect that met the pre-established standard of a medium effect size or greater was the variances of the error terms ($\eta^2 = .99$). The bias estimates of the between person within study variance in the average treatment effect as a function of the variances of the error terms (see Figure 19) illustrated that as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the parameter estimates increased to become progressively overestimated and more biased with mean bias estimates of 0.039 (*SD* = 0.005) to 0.438 (*SD* = 0.029), respectively.
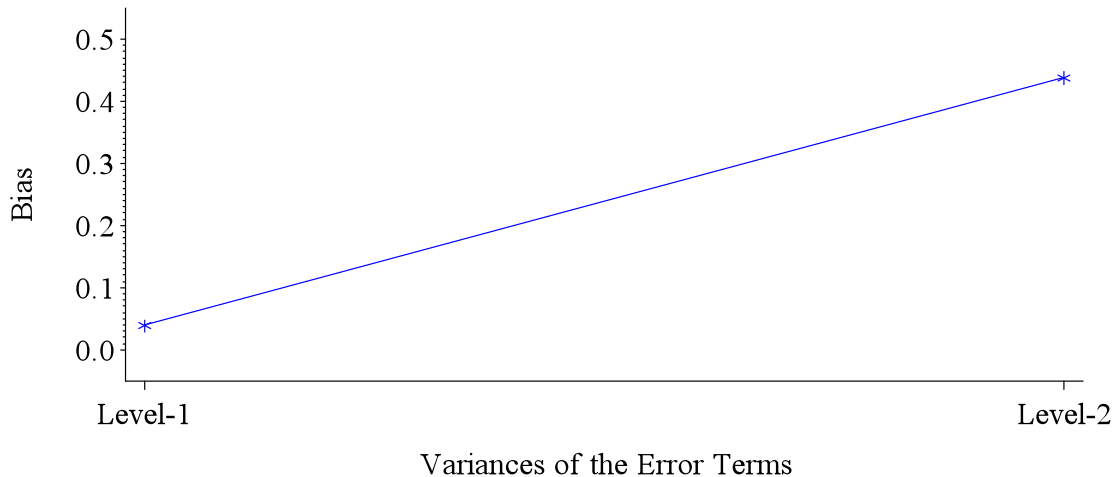
Figure 19. Line graph showing the bias estimates for the between person within study variance in the average treatment effect as a function of the variances of the error terms. Level-1 = most of the variance at level-1; Level-2 = most of the variance at level-2.

Within person residual variance. Variation in the bias estimates of the within person residual variance was explored by modeling bias with the five main effects (number or primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms). This model explained 99% of the variability in the bias estimates of the within person residual variance. Eta-squared (η^2) values for each of the main effects are in Table 11.

Table 11

Eta-squared Values (η²) for Association of Design Factors with Bias of the Within Person Residual Variance

	η²
Number of Primary Studies	.000
Modal Number of Participants	.000
Modal Series Length	.000
Autocorrelation	.999
Variances of the Error Terms	.000
Total Explained	.999

In order to explore these effects further a line graph was created for the main effect of level of autocorrelation ($\eta^2 = .99$). The bias estimates of the within person residual variance as a function of the level of autocorrelation (see Figure 20) illustrated that as the level of autocorrelation increased from 0 to .200 to .400 so did the amount of bias, with values from 0.000 (*SD* = 0.001) to 0.042 (*SD* = 0.001) to 0.191 (*SD* = 0.001), respectively.
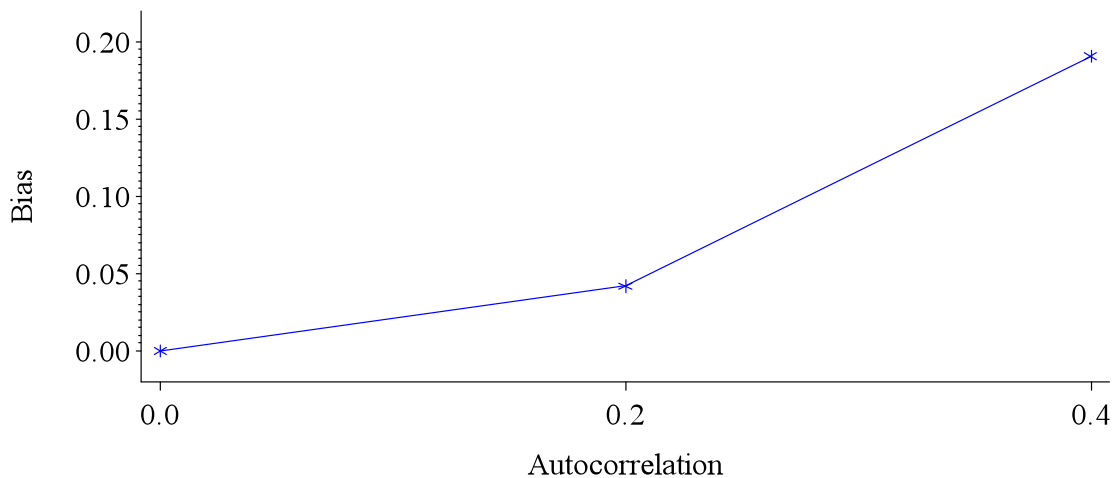
Figure 20. Line graph showing the bias estimates for the within person residual variance as a function of the level of autocorrelation.

Relative Bias

   The distribution of relative bias estimates is illustrated in box plots (see Figure 21

- 23) for all variance components with the exception of the within person residual

variance parameter as its' known parameter value was set to 1.0. The level-3 variance

components ($\tau_{\beta 00}$ and $\tau_{\beta 10}$) were the most biased of all the variance components and

tended to be underestimated (see Figure 21). The relative bias estimates for the between

study variance in the overall average baseline level ($\tau_{\beta 00}$) had values ranging from -.954

to -.624 with a mean of -.858 ($SD = 0.077$). This average relative bias estimate

represented an absolute value of 86% of the average parameter value, which is

substantial. Similarly, the between study variance in the overall average treatment effect

($\tau_{\beta 10}$) had relative bias estimates that ranged from -.948 to -.474 with an average of -.822

($SD = 0.106$). This average relative bias estimate represented an absolute value of 82%

of the average parameter value, which is also substantial.
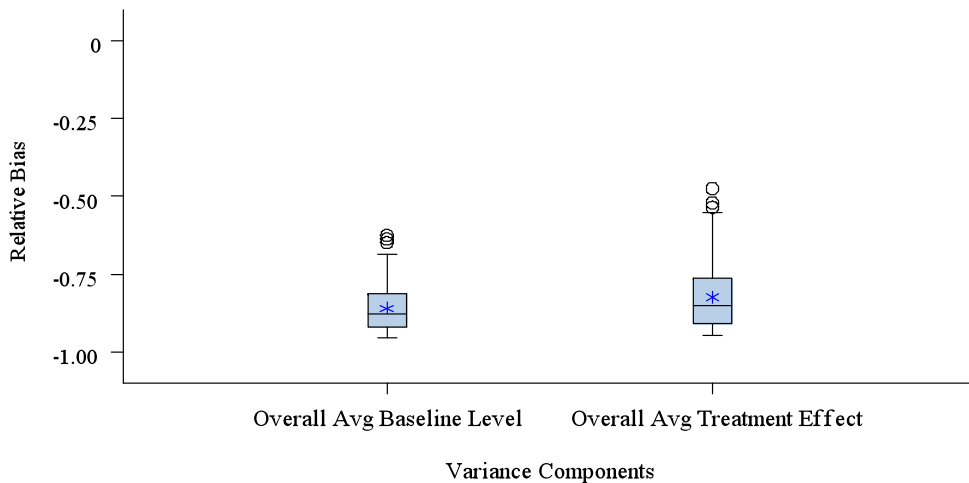
Figure 21. Box plots showing the distribution of relative bias estimates for each level-3
variance component in the three level model

The level-2 variance components ($\tau_{\pi 00}$ and $\tau_{\pi 10}$) were also biased but instead tended to be overestimated (see Figure 22) with the between person within study variance in the average baseline level ($\tau_{\pi 00}$) having relative bias values ranging from .166 to .240 with a mean of .217 ($SD = 0.016$). This average relative bias estimate represented an absolute value of 22% of the average parameter value, which is substantial. The between person within study variance in the average treatment effect ($\tau_{\pi 10}$) had relative bias values ranging from .136 to .238 with an average of .208 ($SD = 0.023$). This average relative bias estimate represented an absolute value of 21% of the average parameter value, which is also substantial.
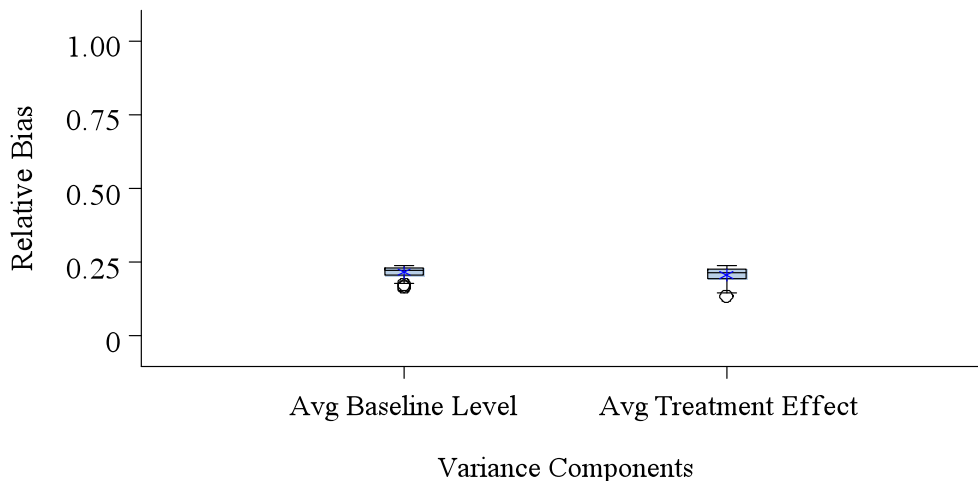
Figure 22. Box plots showing the distribution of relative bias estimates for each level-2 variance component in the three level model

The amount of estimated autocorrelation ($\hat{\rho}$) had relative bias values close to zero (see Figure 23) with a mean of -.001 ($SD = 0.002$) and values ranging from -.007 to .002. As indicated by the results, there was limited variation in the amount of estimated autocorrelation and the relative bias estimate did not exceed 1% of the known parameter value. Therefore, any further exploration was unwarranted.
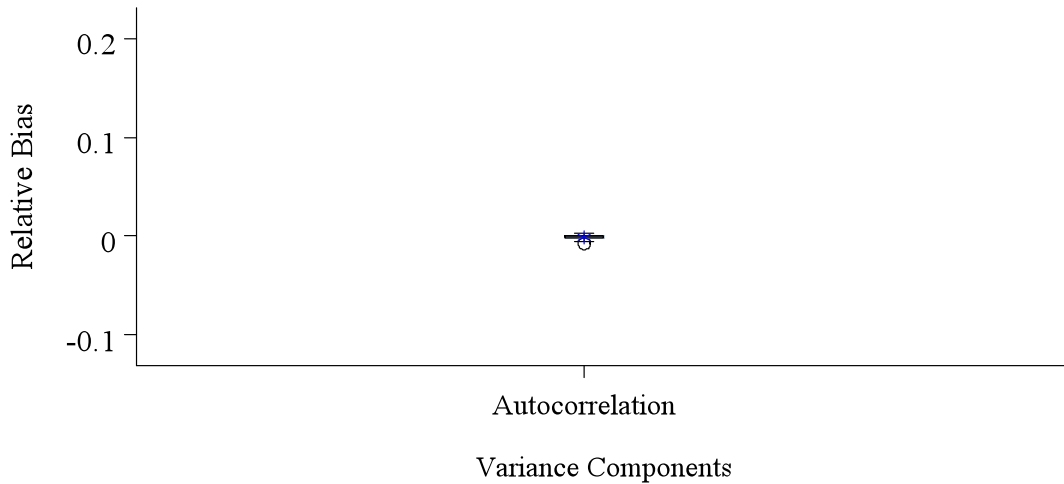
Figure 23. Box plots showing the distribution of relative bias estimates for amount of estimated autocorrelation in the three level model

Between study variance in the overall average baseline level. Variation in the relative bias estimates of the between study variance in the overall average baseline level was explored by modeling relative bias with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms) and a two-way interaction involving the number of primary studies per meta-analysis and the variances of the error terms. This model explained 94% of the variability in the relative bias estimates of the between study variance in the overall average baseline level. Eta-squared (η^2) values for each of the main effects and interaction are in Table 12.

Table 12

Eta-squared Values (η^2) for Association of Design Factors with Relative Bias of the Between Study Variance in the Overall Average Baseline Level

	η^2
Number of Primary Studies	.664
Modal Number of Participants	.156
Modal Series Length	.003
Autocorrelation	.012
Variances of the Error Terms	.084
Number of Primary Studies * Variances of the Error Terms	.019
Total Explained	.938

In order to further explore these effects line graphs were created for the main effects of number of primary studies per meta-analysis ($\eta^2 = .66$), modal number of participants ($\eta^2 = .16$), and variances in the error terms ($\eta^2 = .08$). This model explained 90% of the variance in relative bias of the between study overall average baseline level. These main effects were chosen as they were the only effects that met the pre-established standard of a medium effect size or greater. The relative bias estimates of the variance in the overall average baseline level as a function of the number of primary studies per meta-analysis (see Figure 24) illustrated that as the number of primary studies per meta-analysis increased from 10 to 30 to 80 the parameter estimates became increasingly underestimated and progressively more biased moving from an average relative bias estimate of - .774 (*SD* = 0.067) to -.876 (*SD* = 0.035) to -.925 (*SD* = 0.021), respectively.

Figure 24. Line graph showing the relative bias estimates for the between study variance in the overall average baseline level as a function of the number of primary studies per meta-analysis.

The relative bias estimates of the variance in the overall average baseline level as a function of modal number of participants per primary study (see Figure 25) illustrates that as the modal number of participants per primary study increased from small, with a mode of 4, to large, with a mode of 8, the parameter estimates became increasingly underestimated and progressively more biased moving from an average relative bias estimate of - .828 (*SD* = 0.085) to -.889 (*SD* = 0.054), respectively.

Figure 25. Line graph showing the relative bias estimates for the between study variance in the overall average baseline level as a function of the modal number of participants per primary study.

The relative bias estimates of the variance in the overall average baseline level as a

function of the variances of the error terms (see Figure 26) illustrated that as the

variances of the error terms shifted from most of the variance at level-1 to most of the

variance at level-2 the average parameter became increasingly underestimated and

progressively more biased with the relative bias estimate moving from – 0.836 (*SD* =

0.086) to -0.881 (*SD* = 0.060), respectively.

Figure 26. Line graph showing the relative bias estimates for the between study variance in the overall average baseline level as a function of the variances of the error terms. Level-1 = most of the variance at level-1; Level-2 = most of the variance at level-2.

Between study variance in the overall average treatment effect. Variation in the relative bias estimates of the between study variance in the overall average treatment effect was explored by modeling relative bias with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms) and a two-way interaction between the number of primary studies per meta-analysis and the variances of the error terms. This model explained 94% of the variability in the relative bias estimates of the between study variance in the overall average treatment effect. Eta-squared (η^2) values for each of the main effects are in Table 13.

Table 13

Eta-squared Values (η^2) for Association of Design Factors with Relative Bias of the Between Study Variance in the Overall Average Treatment Effect

	η^2
Number of Primary Studies	.538
Modal Number of Participants	.108
Modal Series Length	.006
Autocorrelation	.018
Variances of the Error Terms	.228
Number of Primary Studies * Variances of the Error Terms	.046
Total Explained	.944

Similar to the relative bias estimates of the between study variance in the overall average baseline level, these effects were further explored with line graphs for the main effects of number of primary studies per meta-analysis ($\eta^2 = .54$), modal number of participants ($\eta^2 = .11$), and variances in the error terms ($\eta^2 = .23$). This model explained 87% of the variability in the relative bias of the between study variance in the overall average treatment effect. The relative bias estimates of the between study variance in the overall average treatment effect as a function of the number of primary studies per meta-analysis (see Figure 27) illustrated that as the number of primary studies per meta-analysis increased from 10 to 30 to 80 the average parameter became increasingly underestimated and progressively more biased with the relative bias estimate moving from -.719 (*SD* = 0.106) to -.843 (*SD* = 0.058) to -.904 (*SD* = 0.035), respectively.

Figure 27. Line graph showing the relative bias estimates for the between study variance in the overall average treatment effect as a function of the number of primary studies per meta-analysis.

The relative bias estimates of the between study variance in the overall average treatment effect as a function of the modal number of participants per primary study (see Figure 28) illustrated that as the modal number of participants per primary study increased from small, with a mode of 4, to large, with a mode of 8, the average parameter became increasingly underestimated and progressively more biased with the relative bias estimate moving from -.788 ($SD = 0.118$) to -.857 ($SD = 0.081$), respectively.

Figure 28. Line graph showing the relative bias estimates for the between study variance in the overall average treatment effect as a function of the modal number of participants per primary study.

The relative bias estimates of the between study variance in the overall average treatment effect as a function of the variances of the error terms (see Figure 29) illustrates that as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the average parameter estimate became increasingly underestimated and progressively more biased with the relative bias estimate moving from -.772 (*SD* = 0.117) to -.872 (*SD* = 0.062), respectively.

Figure 29. Line graph showing the relative bias estimates for the between study variance in the overall average treatment effect as a function of the variances of the error terms. Level-1 = most of the variance at level-1; Level-2 = most of the variance at level-2.

Between person within study variance in the average baseline level. Variation in the relative bias estimates of the between person within study variance in the average baseline level was explored by modeling relative bias with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms), all possible two-way interactions involving the number of primary studies per meta-analysis, and all possible two-way interactions involving the modal number of participants per primary study. This model explained 94% of the variability in the relative bias estimates of the between person within study variance in the average baseline level. Eta-squared (η^2) values for each of the main effects are in Table 14.

Table 14

Eta-squared Values (η^2) for Association of Design Factors with Relative Bias of the Between Person Within Study Variance in the Average Baseline Level

	η^2
Number of Primary Studies	.590
Modal Number of Participants	.166
Modal Series Length	.007
Autocorrelation	.021
Variances of the Error Terms	.094
Number of Primary Studies * Modal Number of Participants	.024
Number of Primary Studies * Modal Series Length	.005
Number of Primary Studies * Autocorrelation	.006
Number of Primary Studies * Variances of the Error Terms	.018
Modal Number of Participants * Modal Series Length	.005
Modal Number of Participants * Autocorrelation	.001
Modal Number of Participants * Variances of the Error Terms	.002
Total Explained	.939

To further explore these effects line graphs were created for the main effects of number of primary studies per meta-analysis ($\eta^2 = .59$), modal number of participants per primary study ($\eta^2 = .17$), and variances of the error terms ($\eta^2 = .09$). This model explained 85% of the variability in the relative bias estimates of the between person within study variance in the average baseline level. The relative bias estimates of the between person within study variance in the average baseline level as a function of the number of primary studies per meta-analysis (see Figure 30) illustrated that as the number of primary studies per meta-analysis increased from 10 to 20 to 30 the parameter estimates increased to become progressively overestimated and slightly more biased, with

mean relative bias estimates of .200 (*SD* = 0.015) to .220 (*SD* = 0.009) to .230 (*SD* =
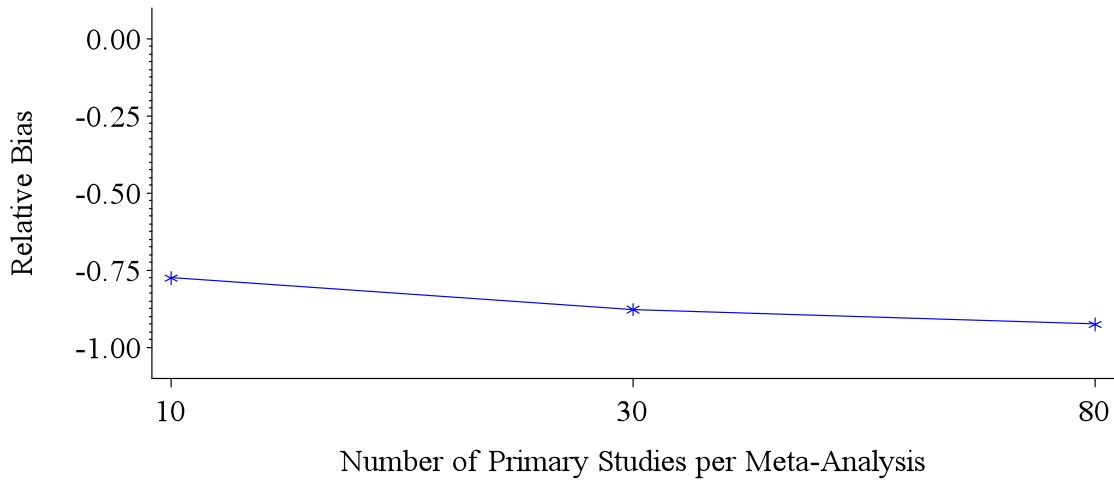
0.005), respectively.
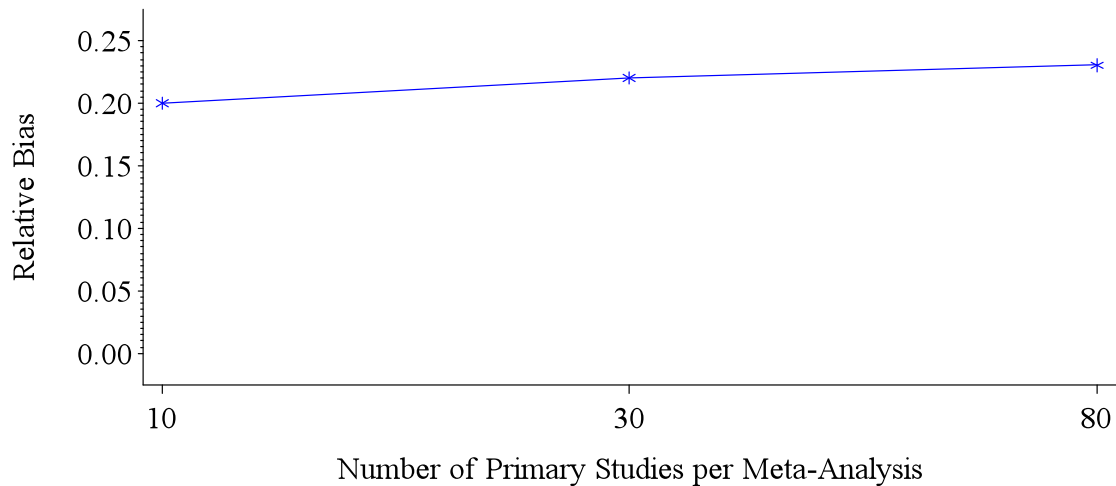
Figure 30. Line graph showing the relative bias estimates for the between person within study variance in the average baseline level as a function of the number of primary studies per meta-analysis.

The relative bias estimates of the between person within study variance in the average

baseline level as a function of the modal number of participants per primary studies (see

Figure 31) illustrates that as the modal number of participants per primary study

increased from small, with a mode of 4, to large, with a mode of 8, the parameter

estimates increased to become progressively overestimated and slightly more biased, with

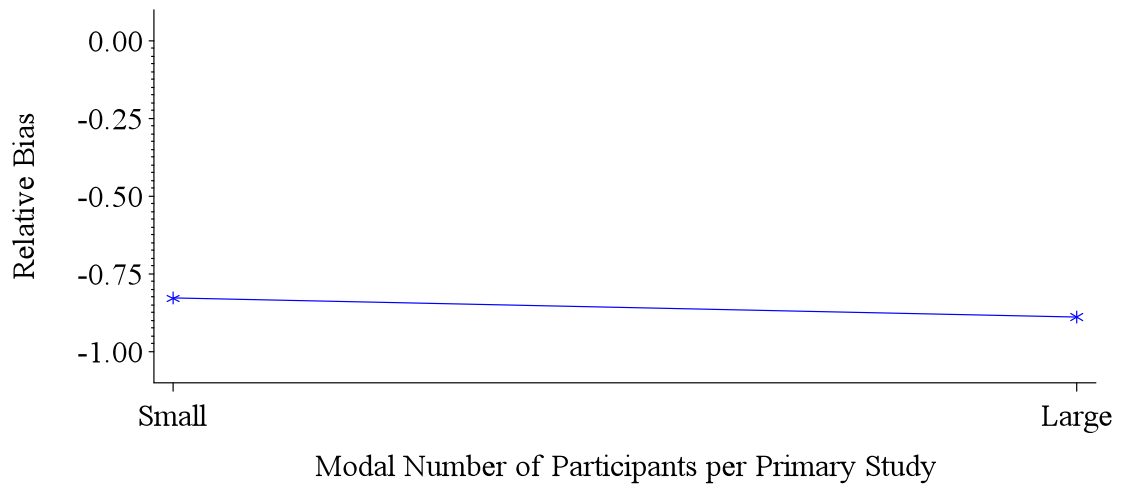mean relative bias estimates of .210 (*SD* = 0.018) to .223 (*SD* = 0.012), respectively.
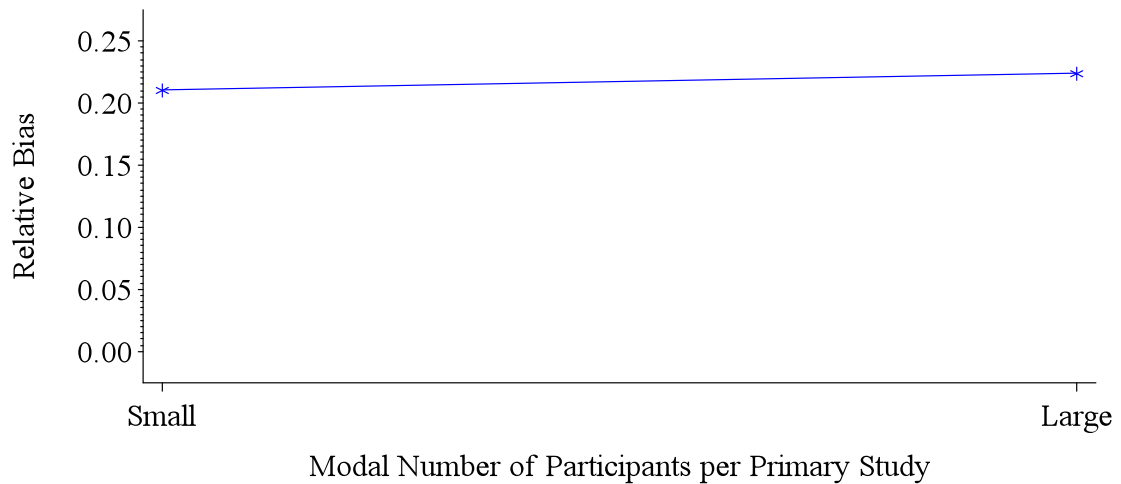
Figure 31. Line graph showing the relative bias estimates for the between person within study variance in the average baseline level as a function of the modal number of participants per primary study.

The relative bias estimates of the between person within study variance in the average baseline level as a function of the variances of the error terms (see Figure 32) illustrated that as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the parameter estimates increased to become progressively overestimated and slightly more biased, with mean relative bias estimates of .212 (*SD* = 0.018) to .222 (*SD* = 0.013), respectively.
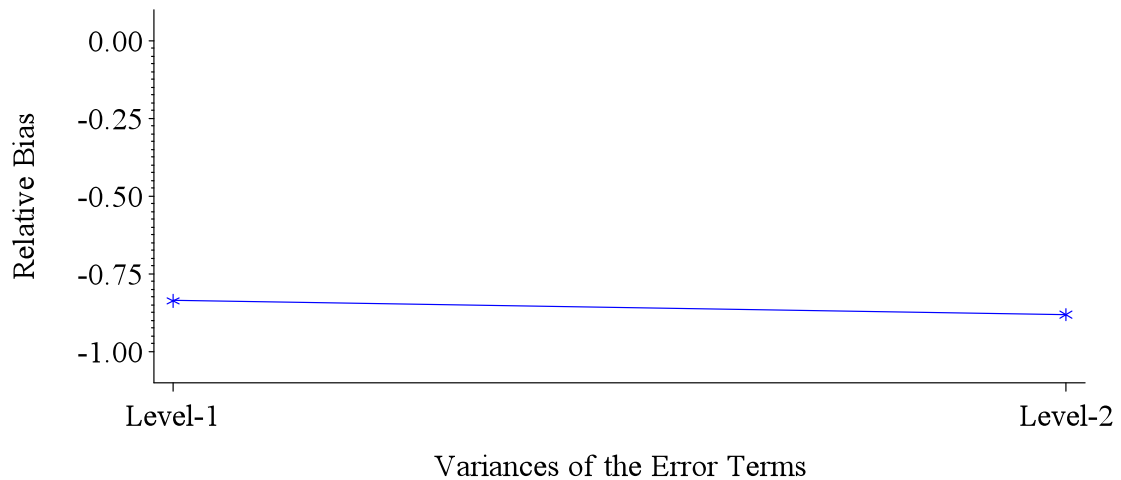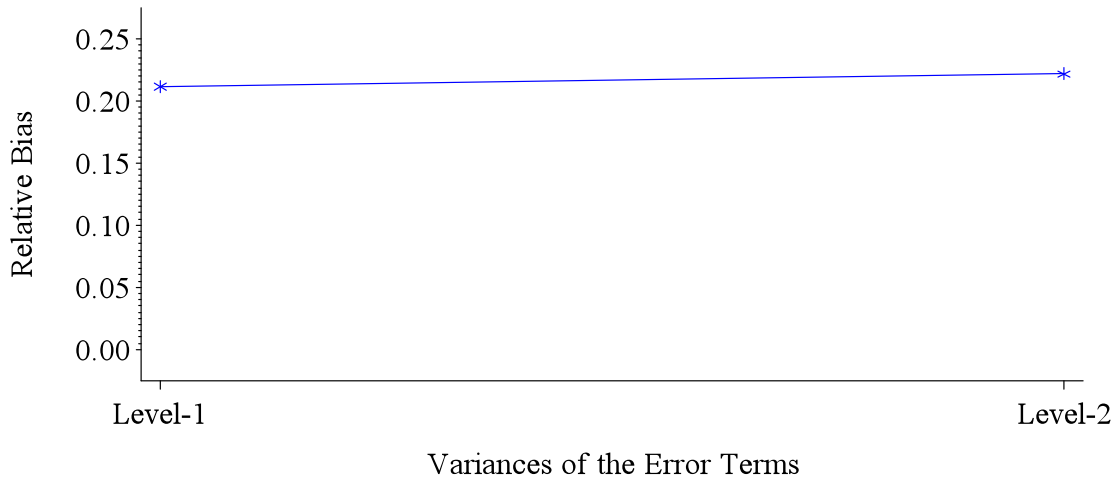
Figure 32. Line graph showing the relative bias estimates for the between person within study variance in the average baseline level as a function of the variances of the error terms. Level-1 = most of the variance at level-1; Level-2 = most of the variance at level-2.

Between person within study variance in the average treatment effect.

Variation in the relative bias estimates of the between person within study variance in the average treatment effect was explored by modeling relative bias with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms), a two-way interaction between the number of primary studies per meta-analysis and the modal number of participants per primary study, and a two-way interaction between the number of primary studies per meta-analysis and the variances of the error terms. This model explained 94% of the variability in the relative bias estimates of the between person within study variance in the average treatment effect. Eta-squared (η^2) values for each of the main effects are in Table 15.

Table 15

Eta-squared Values (η^2) for Association of Design Factors with Relative Bias of the Between Person Within Study Variance in the Average Treatment Effect

	η^2
Number of Primary Studies	.537
Modal Number of Participants	.100
Modal Series Length	.005
Autocorrelation	.019
Variances of the Error Terms	.237
Number of Primary Studies * Modal Number of Participants	.034
Number of Primary Studies * Variances of the Error Terms	.013
Total Explained	.945

Similar to the relative bias results for the between person within study variance in the average treatment effect, the main effects that met the pre-established standard of a medium effect size or greater were the number of primary studies per meta-analysis ($\eta^2 =$ .54), modal number of participants ($\eta^2 = .10$), and variances of the error terms ($\eta^2 = .24$). This model explained 88% of the variability in the between person within study variance of the average treatment effect. The relative bias estimates of the between person within study variance in the average treatment effect as a function of the number of primary studies per meta-analysis (see Figure 33) illustrated that as the number of primary studies per meta-analysis increased from 10 to 30 to 80 the parameter estimates increased to become progressively overestimated and slightly more biased with mean relative bias estimates of .185 (*SD* = 0.022) to .211 (*SD* = 0.015) to .226 (*SD* = 0.008), respectively.
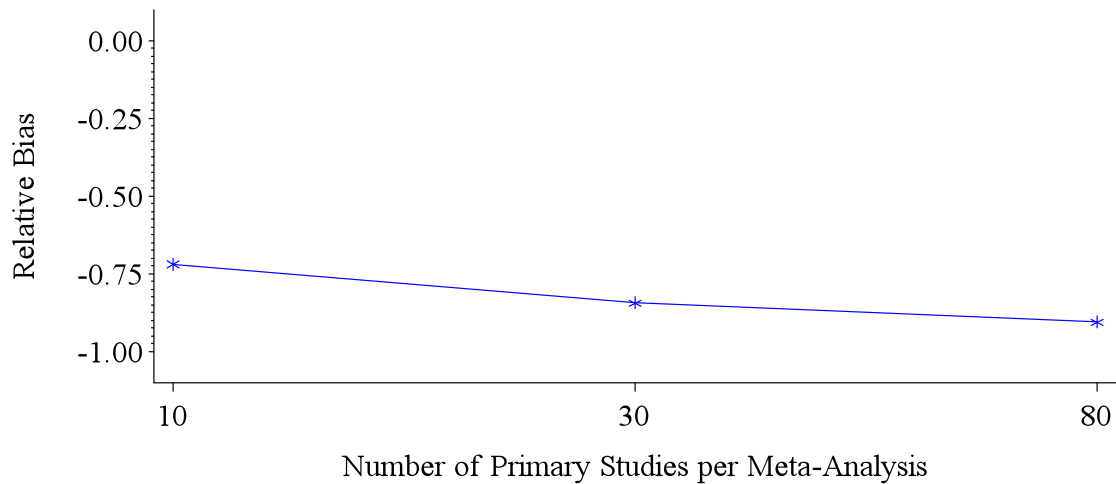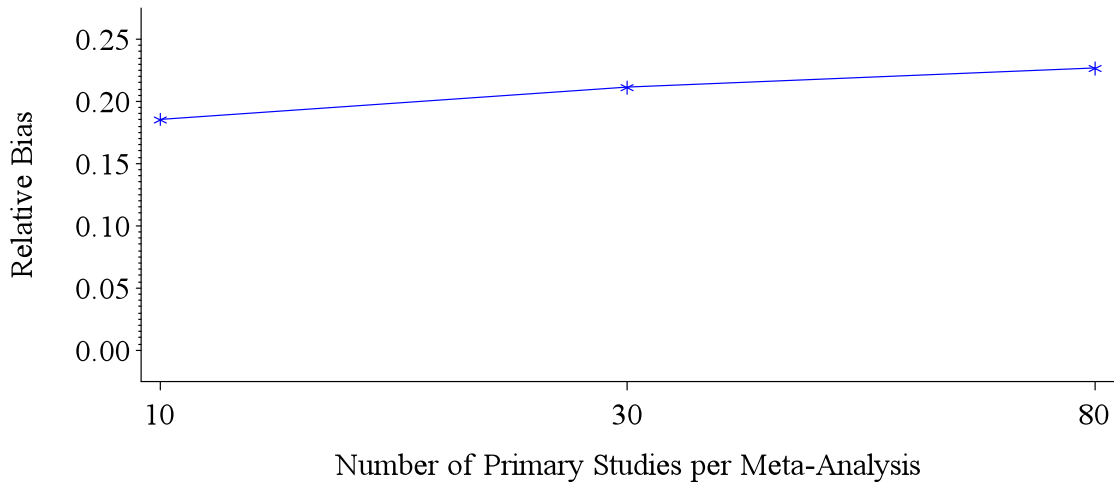
Figure 33. Line graph showing the relative bias estimates for the between person within study variance in the average treatment effect as a function of the number of primary studies per meta-analysis.

The relative bias estimates of the between person within study variance in the average treatment effect as a function of the modal number of participants per primary study (see Figure 34) illustrated that as the modal number of participants per primary study increased from small, with a mode of 4, to large, with a mode of 8, the parameter estimates increased to become progressively overestimated and slightly more biased with mean relative bias estimates of .200 (*SD* = 0.025) to .215 (*SD* = 0.019), respectively.
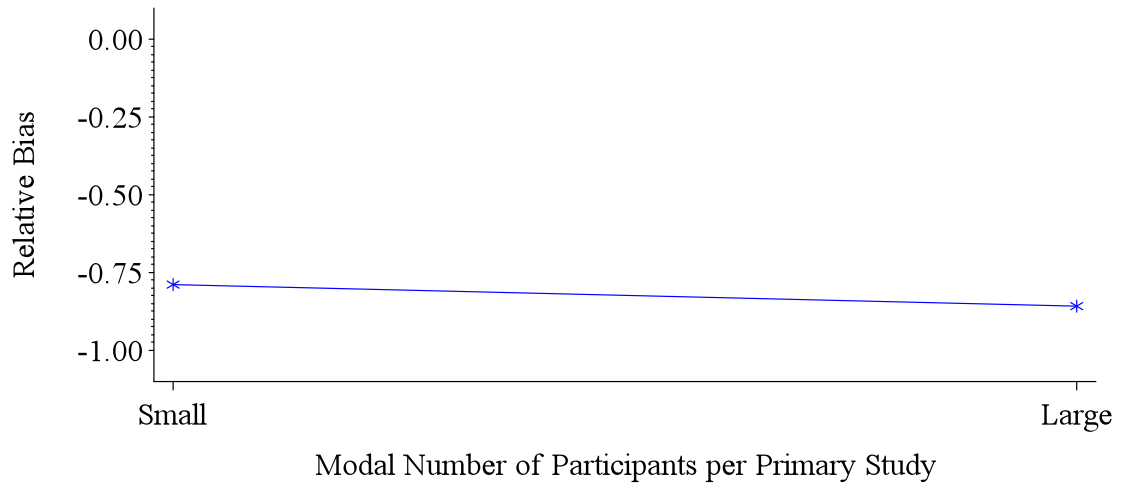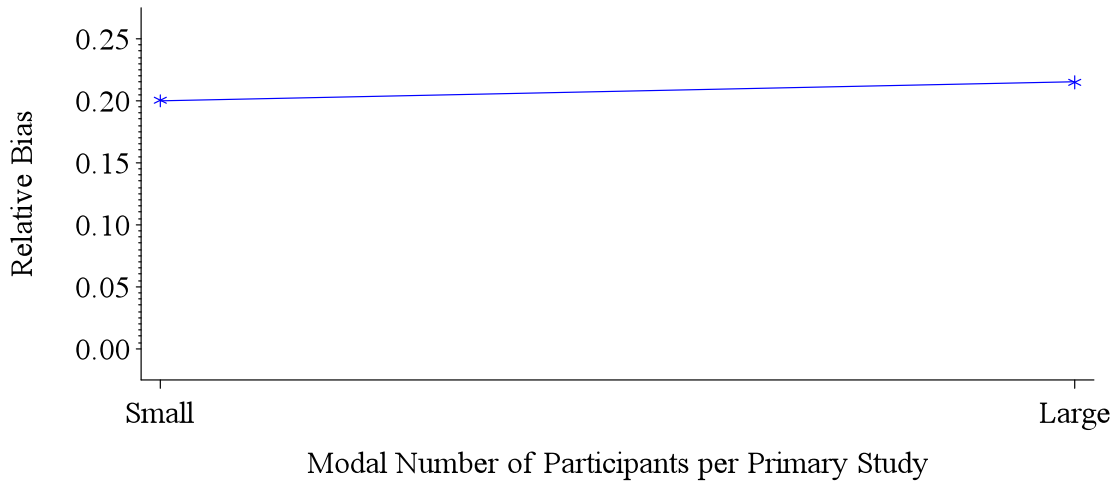
Figure 34. Line graph showing the relative bias estimates for the between person within study variance in the average treatment effect as a function of the modal number of participants per primary study.

The relative bias estimates of the between person within study variance in the average treatment effect as a function of the variances of the error terms (see Figure 35) illustrated that as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the parameter estimates increased to become progressively overestimated and slightly more biased with mean relative bias estimates of .196 ($SD = 0.025$) to .219 ($SD = 0.015$), respectively.
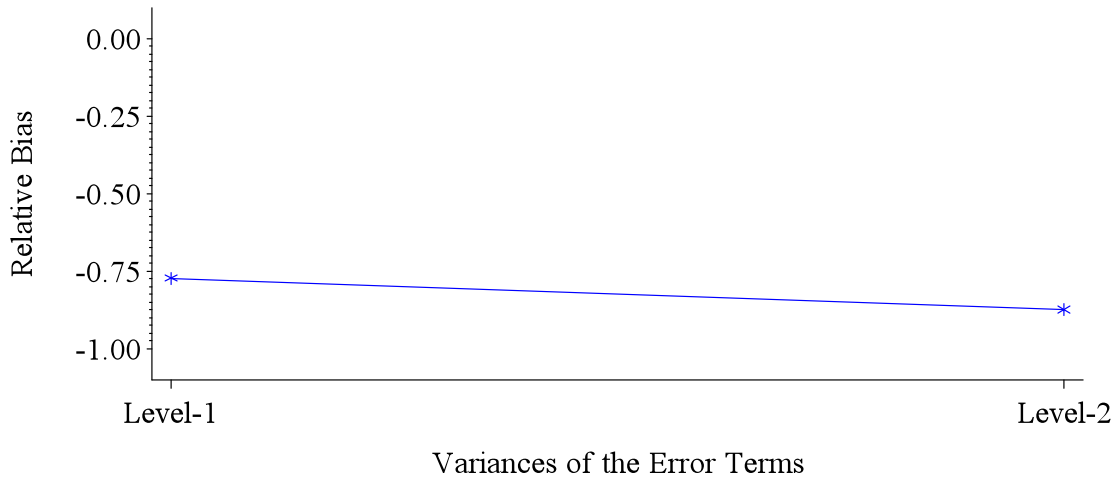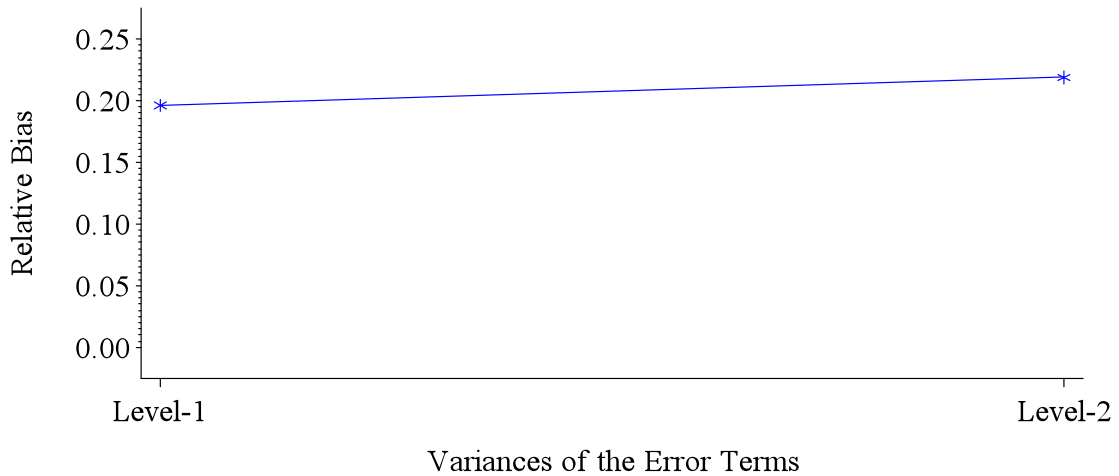
Figure 35. Line graph showing the relative bias estimates for the between person within study variance in the average treatment effect as a function of the variances of the error terms. Level-1 = most of the variance at level-1; Level-2 = most of the variance at level-2.

Confidence Interval Coverage

The distribution of confidence interval coverage rates for each variance component is illustrated in box plots in Figures 36 - 38. The level-3 variance components ($\tau_{\beta00}$ and $\tau_{\beta10}$) tended to overcover (see Figure 36). The between study variance in the overall average baseline level ($\tau_{\beta00}$) had confidence interval coverage rates with values that ranged from a high of 1.000 to a low of .978, with a mean of .998 ($SD = 0.004$). Similarly, the between study variance in the overall average treatment effect ($\tau_{\beta10}$) had confidence interval coverage rates that ranged from a high of 1.000 to a low of .934, with a mean of .995 ($SD = 0.012$).
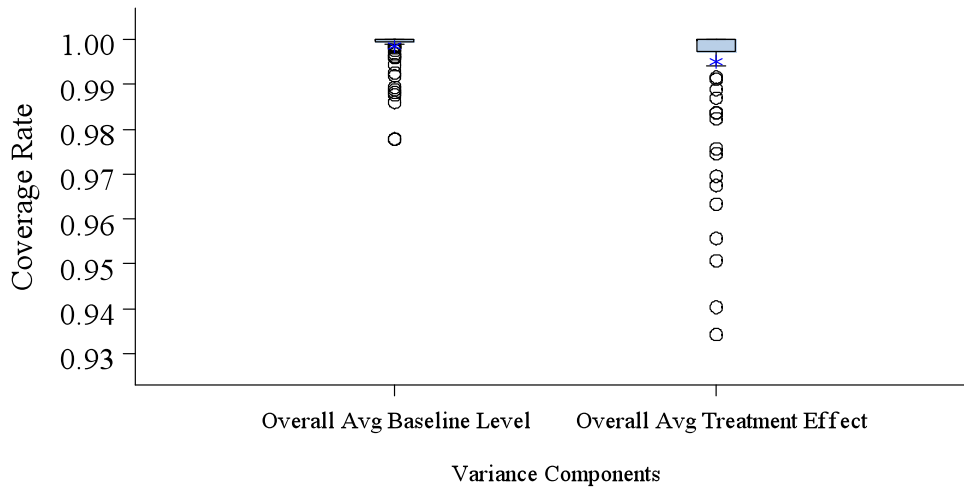
Figure 36. Box plots showing the distribution of confidence interval coverage rates for the level-3 variance components in the three level model.

The level-2 variance components ($\tau_{\pi00}$ and $\tau_{\pi10}$) tended to undercover (see Figure 37).

The between person within study variance in the average baseline level ($\tau_{\pi00}$) had

confidence interval coverage rates with values that ranged from a high of .895 to a low of

.083 with an average coverage rate of .612 ($SD = 0.241$). Likewise, the between person

within study variance in the average treatment effect ($\tau_{\pi10}$) had confidence interval

coverage rates with values ranging from a high of .892 to a low of .109 with a mean of

.675 ($SD = 0.222$).
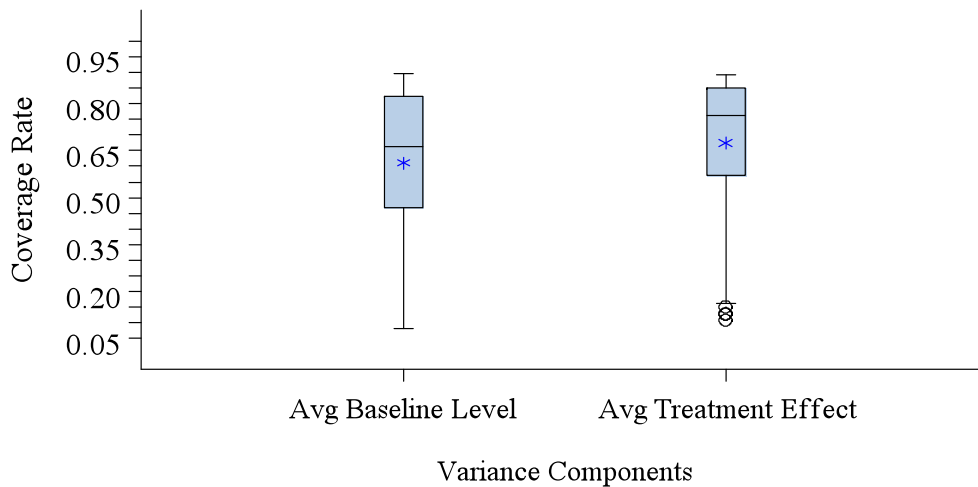
Figure 37. Box plots showing the distribution of confidence interval coverage rates for the level-2 variance components in the three level model.

The within person residual variance (σ^2) tended to undercover with values of confidence interval coverage rates ranging from a low of 0 to a high of .958 and a mean of .550 (*SD* = 0.398). Conversely, the amount of estimated autocorrelation ($\hat{\rho}$) had confidence interval coverage rates that came close to the .95 coverage rate corresponding to a nominal level of .05 with values ranging from a low of .943 to a high of .956 and an average confidence interval coverage rate of .950 (*SD* = 0.003). Figure 38 illustrates the box plots of confidence interval coverage rates for both the within person residual variance and the amount of estimated autocorrelation. As indicated by the results, there was limited variation in the confidence interval coverage rates when estimating the amount of estimated autocorrelation and the 95% confidence interval coverage rates were on average close to the .95 coverage rate. Therefore, any further exploration of confidence interval coverage rates for the amount of estimated autocorrelation is unwarranted.
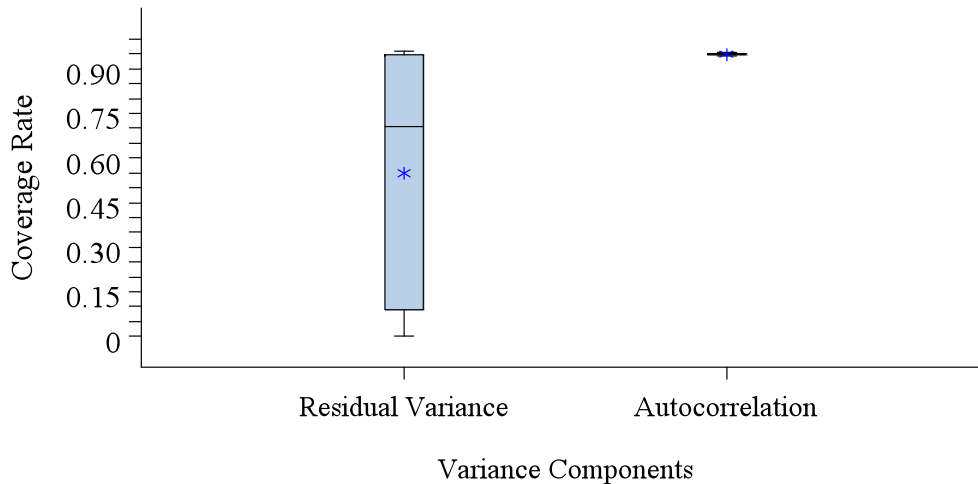
Figure 38. Box plots showing the distribution of confidence interval coverage rates for the within person residual variance and amount of estimated autocorrelation in the three level model.

Between study variance in the overall average baseline level. Variation in confidence interval coverage rates of the between study variance in the overall average baseline level was explored by modeling confidence interval coverage with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms), all possible two-way interactions involving the number of primary studies per meta-analysis, all possible two-way interactions involving modal number of participants per primary study , all possible two-way interactions involving variances of the error terms, a three-way interaction involving the number of primary studies per meta-analysis, modal number of participants per primary study and variances of the error terms, and one four-way interaction involving the number of primary studies per meta-analysis, modal number of participants per primary study, level of autocorrelation, and variances of the error terms. This model explained 96% of the variability in the confidence interval coverage rates of the between study variance in the

overall average baseline level. Eta-squared (η^2) values for each of the main effects and

interactions are in Table 16.

Table 16

Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval
Coverage of the Between Study Variance in the Overall Average Baseline Level

	η^2
Number of Primary Studies	.244
Modal Number of Participants	.088
Modal Series Length	.005
Autocorrelation	.013
Variances of the Error Terms	.081
Number of Primary Studies*Modal Number of Participants	.156
Number of Primary Studies*Modal Series Length	.007
Number of Primary Studies*Autocorrelation	.025
Number of Primary Studies*Variances of the Error Terms	.140
Modal Number of Participants * Variances of the Error Terms	.046
Modal Number of Participants * Modal Series Length	.003
Modal Number of Participants * Autocorrelation	.001
Variances of the Error Terms * Modal Series Length	.003
Variances of the Error Terms * Autocorrelation	.018
Number of Primary Studies * Modal Number of Participants * Variances of the Error Terms	.079
Number of Primary Studies * Modal Number of Participants * Autocorrelation * Variances of the Error Terms	.046
Total Explained	.955

In order to explore these effects further line graphs were created for those main

effects and interactions whose eta-squared values exceeded the pre-established standard

of a medium effect size or greater. Therefore, the effects that met this standard were the

main effects of the number of primary studies per meta-analysis ($\eta^2 = .24$), modal number of participants per primary study ($\eta^2 = .09$), variances of the error terms ($\eta^2 = .08$), the two-way interaction of number of primary studies per meta-analysis with modal number of participants per primary study ($\eta^2 = .16$), the two-way interaction of number of primary studies per meta-analysis with variances of the error terms ($\eta^2 = .14$), and the three-way interaction involving number of primary studies per meta-analysis, modal number of participants, and variances of the error terms ($\eta^2 = .08$). This explained 79% of the variability of the confidence interval coverage rates for the between study variance in the overall average baseline level. As the number of primary studies per meta-analysis increased from 10 to 30 to 80 the confidence interval coverage rates increased, with means of .996 ($SD = 0.006$), .999 ($SD = 0.000$), and 1.000 ($SD = 0.000$), respectively. Also, as the modal number of participants increased from small, with a mode of 4, to large, with a mode of 8, the confidence interval coverage rates increased from an average of .997 ($SD = 0.005$) to 1 ($SD = 0.001$), respectively. Likewise, as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the confidence interval coverage rates increased, with means of .997 ($SD = 0.005$) and 1 ($SD = 0.001$), respectively. The graph (see Figure 39) indicates that when the number of primary studies per meta-analysis was 10, the modal number of participants per primary study was small, and most of the variance in the error terms was at level-1 the confidence interval coverage rates were closest to the nominal level with a mean of .986 ($SD = 0.005$). In addition, all confidence interval coverage rates of the between study variance in the overall average baseline level increased when the variance of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. However,

when the number of primary studies was greater than 10 and/or the modal number of participants was large the increase in confidence interval coverage rates of the between study variance in the overall average baseline level was less noticeable.

Figure 39. Line graph showing the estimated confidence interval coverage rates for the between study variance in the overall average baseline level as a function of the three-way interaction between number of primary studies per meta-analysis, modal number of participants per primary study, and the variances of the error terms.

Between study variance in the overall average treatment effect. Variation in confidence interval coverage rates of the between study variance in the overall average treatment effect was explored by modeling confidence interval coverage with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms), all possible two-way interactions involving the number of primary studies per meta-analysis, all possible two-way interactions involving modal number of participants per primary study , all possible two-way interactions involving variances of the error terms, a three-way interaction involving the number of primary studies per meta-analysis, modal number of participants per primary study and variances

128

of the error terms, and a four-way interaction involving the number of primary studies per meta-analysis, modal number of participants, level of autocorrelation, and variances of the error terms. This model explained 98% of the variability in the confidence interval coverage rates of the between study variance in the overall average treatment effect. Eta-squared (η^2) values for each of the main effects and interactions are in Table 17.

Table 17
Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval Coverage of the Between Study Variance in the Overall Average Treatment Effect

	η^2
Number of Primary Studies	.250
Modal Number of Participants	.063
Modal Series Length	.007
Autocorrelation	.025
Variances of the Error Terms	.132
Number of Primary Studies*Modal Number of Participants	.082
Number of Primary Studies*Modal Series Length	.012
Number of Primary Studies*Autocorrelation	.032
Number of Primary Studies*Variances of the Error Terms	.195
Modal Number of Participants * Variances of the Error Terms	.047
Modal Number of Participants * Modal Series Length	.001
Modal Number of Participants * Autocorrelation	.006
Variances of the Error Terms * Modal Series Length	.006
Variances of the Error Terms * Autocorrelation	.023
Number of Primary Studies * Modal Number of Participants * Variances of the Error Terms	.058
Number of Primary Studies * Modal Number of Participants * Autocorrelation * Variances of the Error Terms	.042
Total Explained	.975

Similar to the results for the between study variance in the overall average baseline level, a line graph was created for the number of primary studies per meta-analysis ($\eta^2 = .25$), modal number of participants per primary study ($\eta^2 = .06$), variances of the error terms ($\eta^2 = .13$), the two-way interaction of number of primary studies per meta-analysis with modal number of participants per primary study ($\eta^2 = .08$), the two-way interaction of number of primary studies per meta-analysis with variances of the error terms ($\eta^2 = .19$), and the three-way interaction involving number of primary studies per meta-analysis, modal number of participants, and variances of the error terms ($\eta^2 = .06$). This model explained 77% of the variability in the confidence interval coverage of the between study variance in the overall average treatment effect. As the number of primary studies per meta-analysis increased from 10 to 30 to 80 the confidence interval coverage rates of the between study variance of the overall average treatment effect increased, with means of .986 ($SD = 0.018$), .999 ($SD = 0.003$), and 1.000 ($SD = 0.000$), respectively. Also, as the modal number of participants increased from small, with a mode of 4, to large, with a mode of 8, the confidence interval coverage rates of the between study variance of the overall average treatment effect increased with an average of .992 ($SD = 0.016$) to .998 ($SD = 0.005$), respectively. Likewise, as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the confidence interval coverage rates of the between study variance of the overall average treatment effect increased, with means of .991 ($SD = 0.016$) and .999 ($SD = 0.001$), respectively. The graph (see Figure 40) indicates that when the number of primary studies per meta-analysis was 10, the modal number of participants per primary study was small, and most of the variance of the error terms was at level-1 the confidence

130

interval coverage rates of the between study variance of the overall average treatment effect were closest to the nominal level with a mean of .960 ($SD = 0.016$). In addition, all confidence interval coverage rates of the between study variance of the overall average treatment effect increased when the variance of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. However, when the number of primary studies was 10 and the modal number of participants was small the increase in confidence interval coverage rates of the between study variance of the overall average treatment effect was the most noticeable.
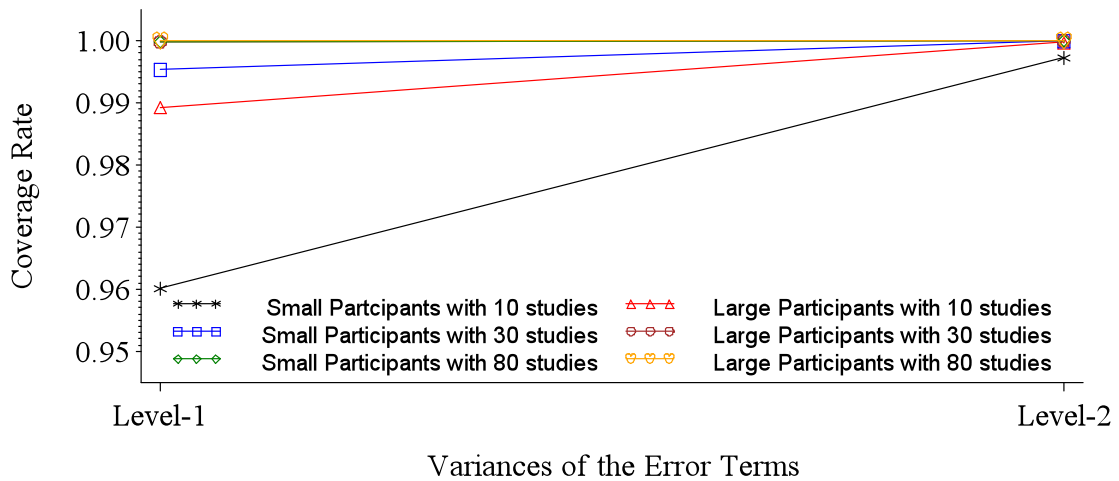
Figure 40. Line graph showing the estimated confidence interval coverage rates for the between study variance in the overall average treatment effect as a function of the three-way interaction between number of primary studies per meta-analysis, modal number of participants per primary study, and the variances of the error terms.

Between person within study variance in the average baseline level. Variation in the confidence interval coverage rates of the between person within study variance in the average baseline level was explored by modeling confidence interval coverage with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of

131

autocorrelation, and variances of the error terms) and a two-way interaction involving

number of primary studies per meta-analysis and variances of the error terms. This

model explained 97% of the variability in the confidence interval coverage of the

between person within study variance in the average baseline level. Eta-squared (η^2)

values for each of the main effects and interaction are in Table 18.

Table 18
Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval
Coverage of the Between Person Within Study Variance in the Average Baseline Level

	η^2
Number of Primary Studies	.736
Modal Number of Participants	.055
Modal Series Length	.006
Autocorrelation	.011
Variances of the Error Terms	.119
Number of Primary Studies * Variances of the Error Terms	.041
Total Explained	.968

To further explore these effects line graphs were created for the main effects of

number of primary studies per meta-analysis ($\eta^2 = .74$), modal number of participants per

primary study ($\eta^2 = .06$), and variances of the error terms ($\eta^2 = 12$). The confidence

interval coverage rates of the between person within study variance in the average

baseline level as a function of the number of primary studies per meta-analysis (see

Figure 41) illustrated that as the number of primary studies per meta-analysis increased

from 10 to 30 to 80 confidence interval coverage rates of the between person within study

variance in the average baseline level decreased with means of .840 (*SD* = 0.036) to .656

(*SD* = 0.112) to .340 (*SD* = 0.182), respectively.
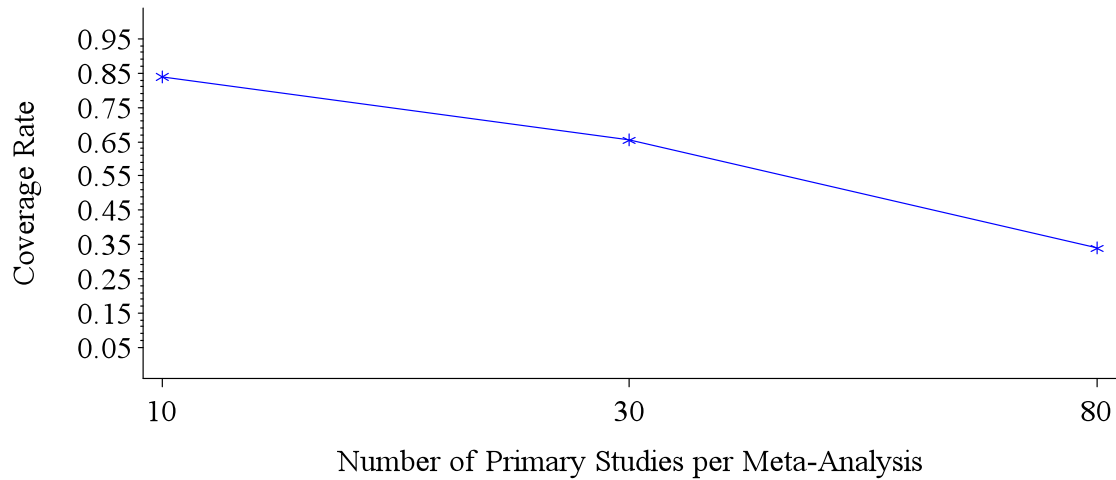
Figure 41. Line graph showing the estimated confidence interval coverage rates for the between person within study variance in the average baseline level as a function of the number of primary studies per meta-analysis.

The confidence interval coverage rates of the between person within study variance in the average baseline level as a function of the modal number of participants per primary study (see Figure 42) illustrated that as modal number of participants per primary study increased from small, with a mode of 4, to large, with a mode of 8, the confidence interval coverage rates of the between person within study variance in the average baseline level decreased with means of .668 (*SD* = 0.208) to .556 (*SD* = 0.261), respectively.
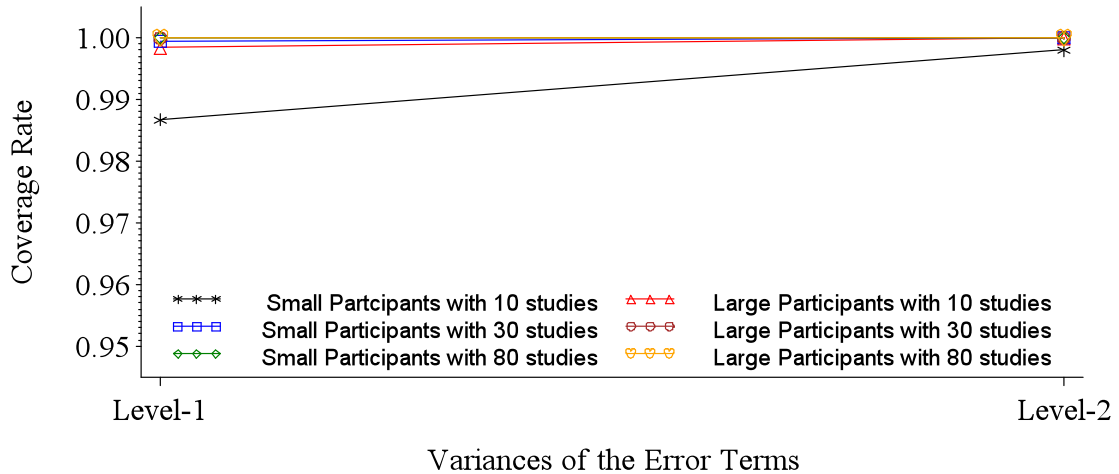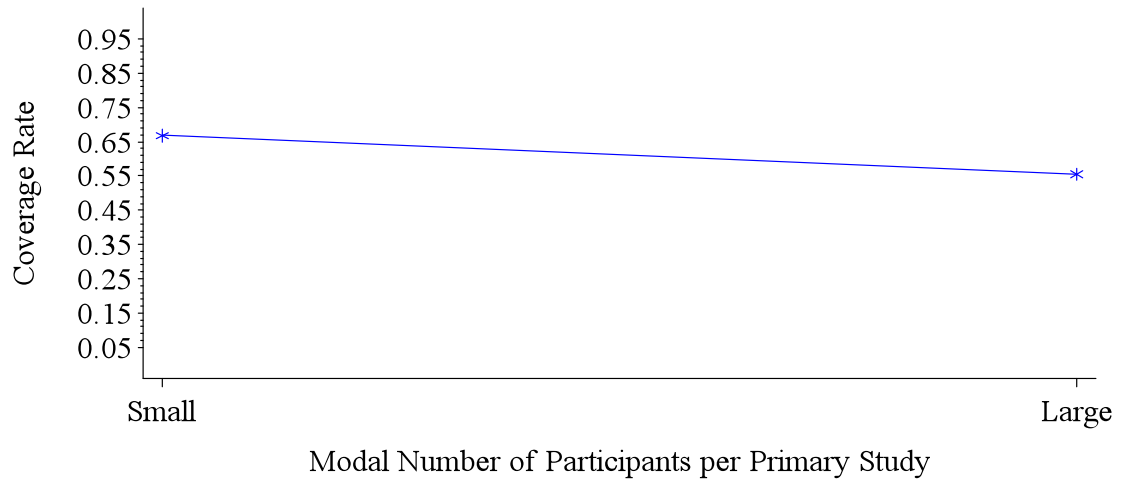
Figure 42. Line graph showing the estimated confidence interval coverage rates for the between person within study variance in the average baseline level as a function of the modal number of participants per primary study.

The confidence interval coverage rates of the between person within study variance in the average baseline level as a function of the variances of the error terms (see Figure 43) illustrated that as variances in the error terms shifted from most of the variance in level-1 to most of the variance in level-2 the confidence interval coverage rates of the between person within study variance in the average baseline level decreased with means of .695 ($SD = 0.178$) to .529 ($SD = 0.267$), respectively.
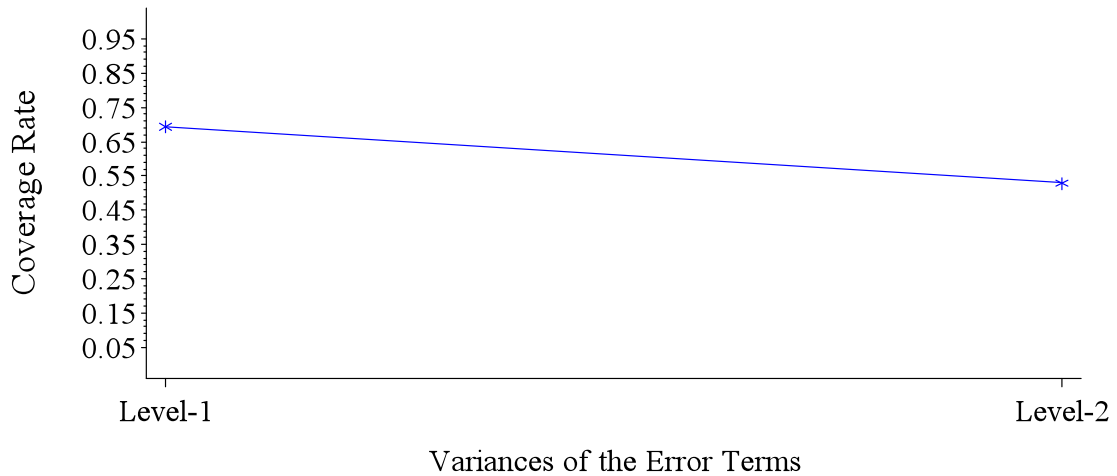
Figure 43. Line graph showing the estimated confidence interval coverage rates for the between person within study variance in the average baseline level as a function of the variances of the error terms.

Between person within study variance in the average treatment effect.

Variation in the confidence interval coverage rates of the between person within study variance in the average treatment effect was explored by modeling confidence interval coverage with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms) and a two-way interaction involving number of primary studies per meta-analysis and variances of the error terms. This model explained 97% of the variability in the confidence interval coverage rates of the between person within study variance in the average treatment effect. Eta-squared (η^2) values for each of the main effects and interaction are in Table 19.

135

Table 19

Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval Coverage of the Between Person Within Study Variance in the Average Treatment Effect

	η^2
Number of Primary Studies	.554
Modal Number of Participants	.038
Modal Series Length	.004
Autocorrelation	.008
Variances of the Error Terms	.248
Number of Primary Studies * Variances of the Error Terms	.120
Total Explained	.972

To further explore these effects a line graph was created that modeled the confidence interval coverage rates of the between person within study variance in the average treatment effect as a function of the number of primary studies per meta-analysis ($\eta^2 = .55$), variances of the error terms ($\eta^2 = .25$) and their interaction ($\eta^2 = .12$) and thus explaining 92% of the variability in the confidence interval coverage rates of the between person within study variance in the average treatment effect. As the number of primary studies per meta-analysis increased from 10 to 30 to 80 the confidence interval coverage rates of the between person within study variance in the average treatment effect decreased, with means of .854 ($SD = 0.030$), .715 ($SD = 0.116$), and .457 ($SD = 0.230$), respectively. Likewise, as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the confidence interval coverage rates of the between person within study variance in the average treatment effect decreased, with means of .785 ($SD = 0.103$) and .565 ($SD = 0.254$), respectively. In addition, the graph (see Figure 44) indicates that when the number of primary studies per

meta-analysis was 10 and most of the variance was at level-1 the confidence interval

coverage rates of the between person within study variance in the average treatment

effect were closest to the nominal level with a mean of .876 ($SD = 0.010$). The graph

also illustrates a steep decline in confidence interval coverage rates of the between person

within study variance in the average treatment effect as the number of primary studies per

meta-analysis increased and when most of the variance in the error terms was at level-2

with the worst coverage rates occurring when the number of primary studies per meta-

analysis was 80 ($M= .248$, $SD = 0.960$).
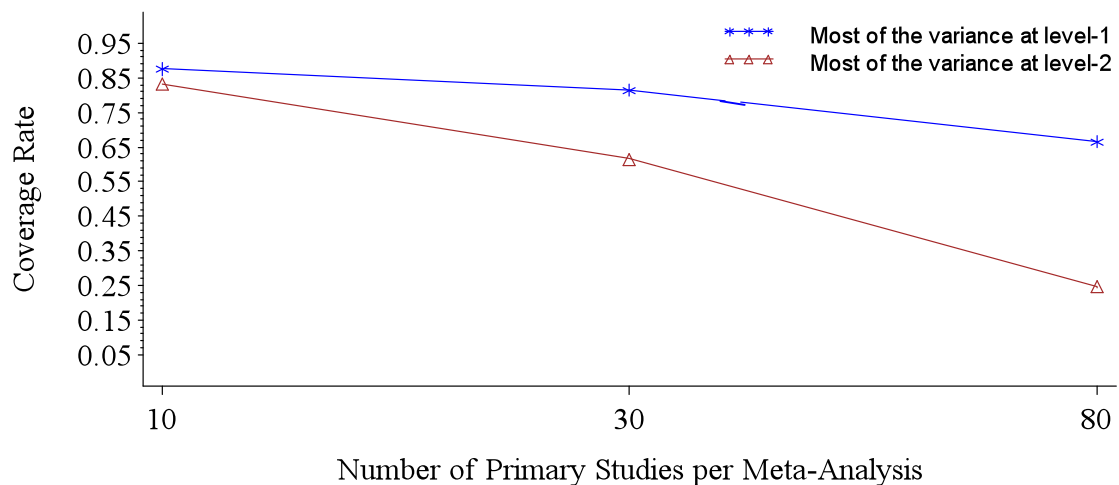
Figure 44. Line graph showing the estimated coverage rates for the between person
within study variance in the average treatment effect as a function of the variance of the
error terms for each level of the number of primary studies per meta-analysis.

Within person residual variance. Variation of the confidence interval coverage

rates of the within person residual variance was explored by modeling confidence interval

coverage with the five main effects (number of primary studies per meta-analysis, modal

number of participants per primary study, modal series length per primary study, level of

autocorrelation, and variances of the error terms) and a two-way interaction involving the

number of primary studies per meta-analysis and the level of autocorrelation. This model

137

explained 98% of the variability in the confidence interval coverage rates of the within

person residual variance. Eta-squared (η^2) values for each of the main effects and

interactions are in Table 20.

Table 20

Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval Coverage of the Within Person Residual Variance

	η^2
Number of Primary Studies	.055
Modal Number of Participants	.003
Modal Series Length	.008
Autocorrelation	.863
Variances of the Error Terms	.000
Number of Primary Studies * Level of Autocorrelation	.052
Total Explained	.981

To further explore these effects line graphs were created that modeled the

confidence interval coverage rates of the within person residual variance as a function of

the number of primary studies per meta-analysis ($\eta^2 = .06$), and the level of

autocorrelation ($\eta^2 = .86$). The graph (see Figure 45) illustrates that as the number of

primary studies per meta-analysis increased from 10 to 30 to 80 the confidence interval

coverage rates of the within person residual variance decreased, with means of .665 (*SD*

$= 0.356$), .547 (*SD* $= 0.407$), and .437 (*SD* $= 0.406$), respectively. In addition, as the level

of autocorrelation increased from 0 to .200 to .400 the confidence interval coverage rates

of the within person residual variance decreased with means of .951 (*SD* $= 0.003$), .636

(*SD* $= 0.233$), and .062 (*SD* $= 0.109$), respectively (see Figure 46).
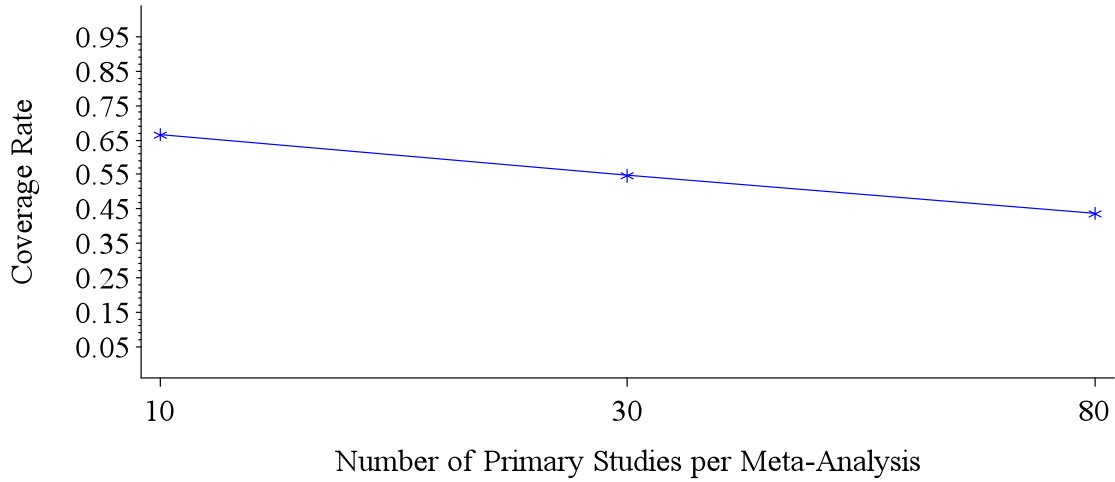
Figure 45. Line graph showing the estimated confidence interval coverage rates of the within person residual variance as a function of the number of primary studies per meta-analysis.
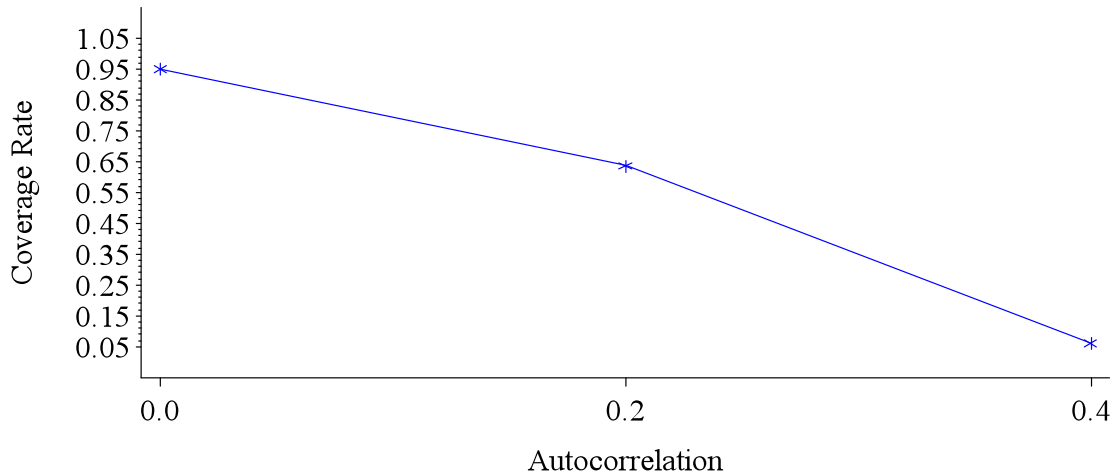
Figure 46. Line graph showing the estimated confidence coverage rates of the within person residual variance as a function of the level of autocorrelation.

Confidence Interval Width

Interval widths were so large for the level-3 ($\tau_{\beta00}$ and $\tau_{\beta10}$) and level-2 ($\tau_{\pi00}$ and

$\tau_{\pi10}$) variance components that they provided no information. Specifically, the

confidence interval width estimates of between study variance in the overall average

139

baseline level ($\tau_{\beta 00}$) ranged from a low of 3.890 X 10^{269} to a high of 3.197 X 10^{286}, with

a mean of 1.568 X 10^{285}. The confidence interval width estimates for the between study

variance in the overall average treatment effect ($\tau_{\beta 10}$) ranged from a low of 4.136 X 10^{269}

to a high of 5.919 X 10^{286}, with a mean of 2.449 X 10^{285}. In addition, the confidence

interval width estimates of the between person within study variance in the average

baseline level ($\tau_{\pi 00}$) ranged from a low of 0.077 to a high of 5.429 X 10^{282}, with a mean

of 5.027 X 10^{280}. The confidence interval width estimates for the between person within

study variance in the average treatment effect ($\tau_{\pi 10}$) ranged from a low of 0.103 to a high

of 9.691 X 10^{286}, with a mean of 9.419 X 10^{282}. As illustrated by the results, the

confidence interval widths were so large for the level-3 and level-2 variance components

that further investigation was unwarranted.

The within person residual variance (σ^2) had an average confidence interval

width estimate of 0.146 ($SD = 0.075$) and values ranging from 0.047 to 0.368. The

amount of estimated autocorrelation ($\hat{\rho}$) had a mean confidence interval width estimate

of 0.090 ($SD = 0.041$) and values ranging from 0.034 to 0.177. Figure 47 illustrates the

distribution of confidence interval width estimates for the within person residual variance

and the amount of estimated autocorrelation.
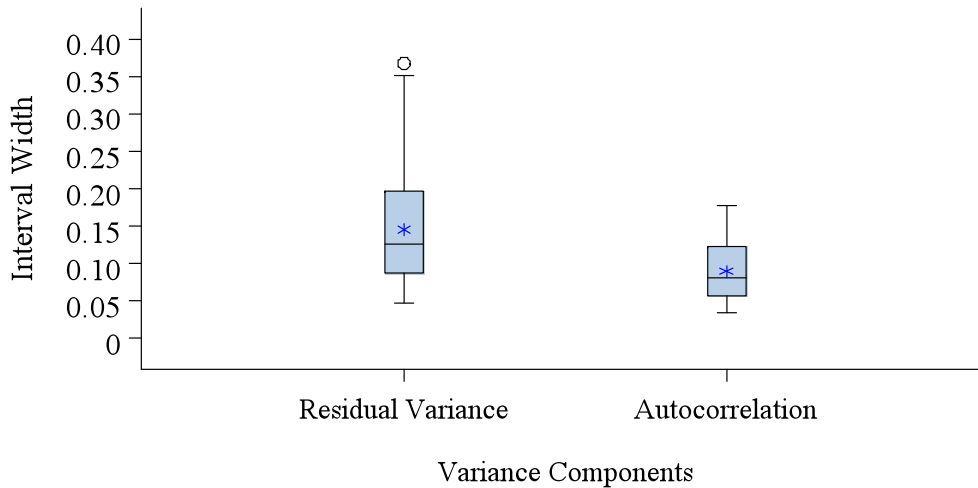
Figure 47. Box plots showing the distribution of confidence interval width estimates for the within person residual variance and the amount of estimated autocorrelation in the three level model.

Within person residual variance. Variation of the confidence interval widths in the within person residual variance was explored by modeling confidence interval width with the five main effects (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms) and a two-way interaction involving the number of primary studies per meta-analysis and the level of autocorrelation. This model explained 97% of the variability in the confidence interval width estimates of the within person residual variance. Eta-squared (η^2) values for each of the main effects and interactions are in Table 21.

Table 21

Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval Width of the Within Person Residual Variance

	η^2
Number of Primary Studies	.674
Modal Number of Participants	.026
Modal Series Length	.078
Autocorrelation	.160
Variances of the Error Terms	.000
Number of Primary Studies * Level of Autocorrelation	.029
Total Explained	.967

To further explore these effects line graphs were created that modeled the confidence interval width estimates as a function of the number of primary studies per meta-analysis ($\eta^2 = .67$), modal series length per primary study ($\eta^2 = .08$), and the level of autocorrelation ($\eta^2 = .16$) as they were the only effects that met the pre-established standard of a medium effect size or greater. The confidence interval width estimates of the within person residual variance as a function of the number of primary studies per meta-analysis (see Figure 48) illustrated that as the number of primary studies per meta-analysis increased from 10 to 30 to 80 confidence interval width estimates decreased with means of .227 (*SD* = 0.062) to .130 (*SD* = 0.035) to .080 (*SD* = 0.022), respectively.
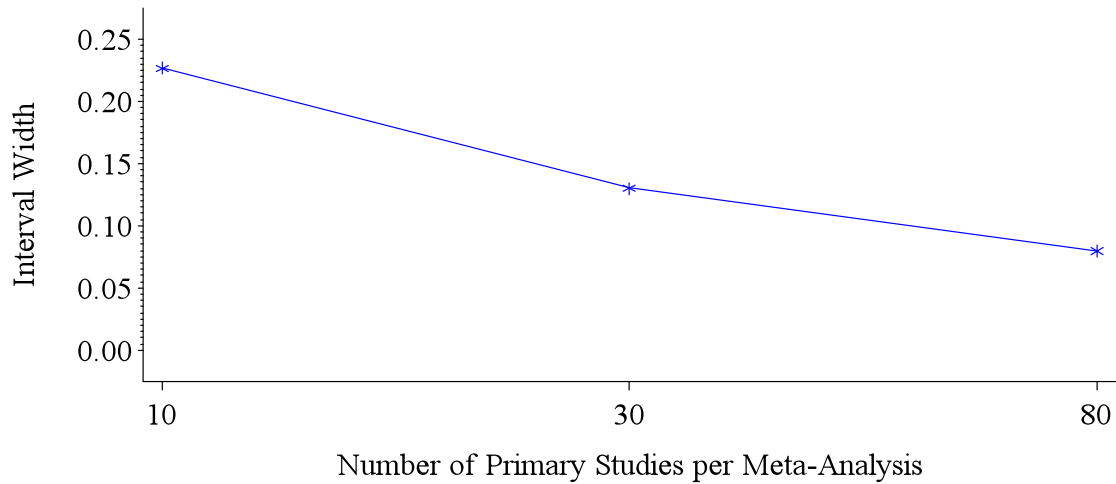
Figure 48. Line graph showing the estimated confidence interval widths for the within person residual variance as a function of the number of primary studies per meta-analysis.

The confidence interval width estimates of the within person residual variance as a function of the modal series length per primary study (see Figure 49) illustrates that as the modal series length increased from small, with a mode of 10, to medium, with a mode of 20, to large, with a mode of 30, the confidence interval width estimates of the within person residual variance decreased with means of .172 ($SD = 0.084$) to .143 ($SD = 0.070$) to .122 ($SD = 0.061$), respectively.
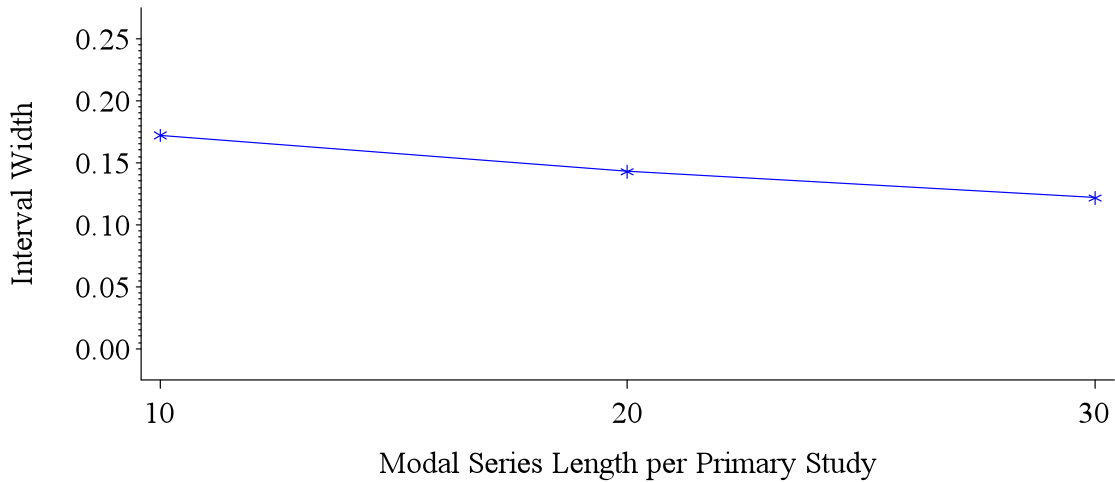
Figure 49. Line graph showing the estimated confidence interval widths for the with person residual variance as a function of the modal series length per primary study.

The confidence interval width estimates of the within person residual variance as a function of the level of autocorrelation (see Figure 50) illustrated that as the level of autocorrelation increased from 0 to .200 to .400 the confidence interval width estimates of the within person residual variance increased with means of .117 (*SD* = 0.053) to .134 (*SD* = 0.062) to .186 (*SD* = 0.087), respectively.
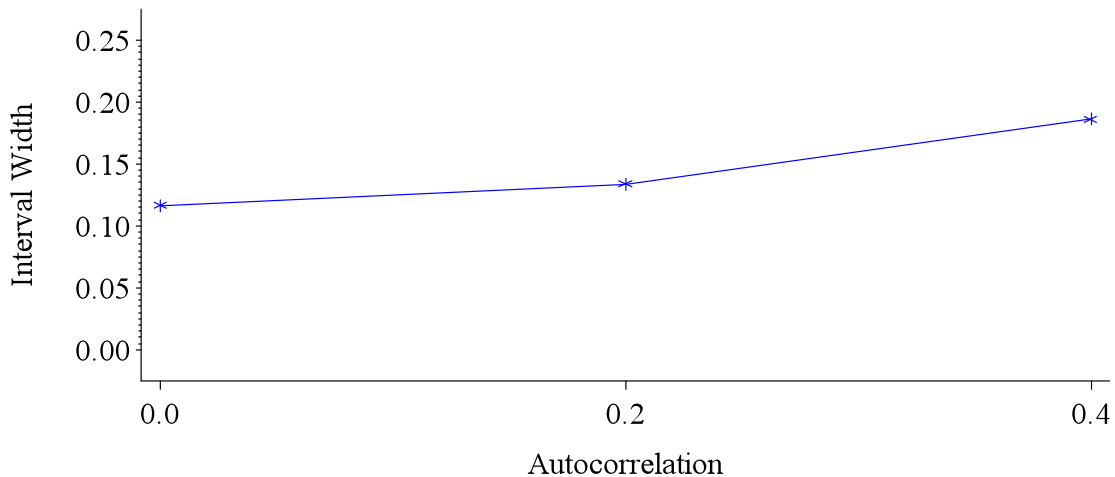
Figure 50. Line graph showing the estimated confidence interval widths for the within person residual variance as a function of the level of autocorrelation.

Amount of estimated autocorrelation. Variation in the confidence interval

width estimates of the amount of autocorrelation was explored by modeling confidence

interval width with the five main effects (number of primary studies per meta-analysis,

modal number of participants per primary study, modal series length per primary study,

level of autocorrelation, and variances of the error terms). This model explained 97% of

the variability in the confidence interval width estimates of the amount of autocorrelation.

Eta-squared (η^2) values for each of the main effects are in Table 22.

Table 22
*Eta-squared Values (η^2) for Association of Design Factors with Confidence Interval
Width of the Amount of Estimated Autocorrelation*

	η^2
Number of Primary Studies	.827
Modal Number of Participants	.031
Modal Series Length	.107
Autocorrelation	.000
Variances of the Error Terms	.002
Total Explained	.967

To further explore these effects line graphs were created that modeled the

confidence interval width estimates as a function of the number of primary studies per

meta-analysis ($\eta^2 = .83$), and modal series length per primary study ($\eta^2 = .11$). The

confidence interval width estimates of the amount of estimated autocorrelation as a

function of the number of primary studies per meta-analysis (see Figure 51) illustrate that

as the number of primary studies per meta-analysis increased from 10 to 30 to 80

confidence interval width estimates of the amount of estimated autocorrelation decreased

with means of .139 (*SD* = 0.025) to .080 (*SD* = 0.014) to .050 (*SD* = 0.009), respectively.

145
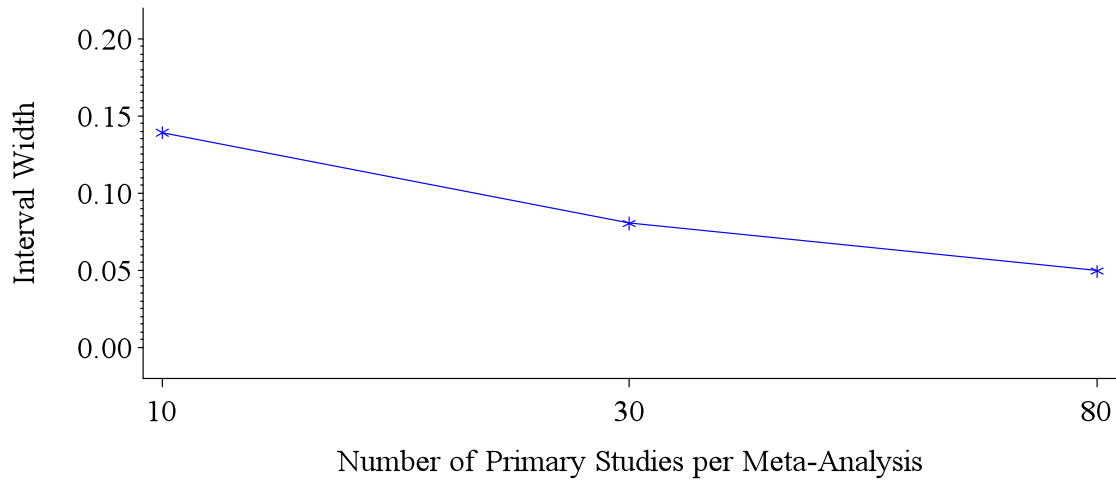
Figure 51. Line graph showing the estimated confidence interval widths for the amount of estimated autocorrelation as a function of the number of primary studies per meta-analysis.

The confidence interval width estimates of the amount of estimated autocorrelation as a function of the modal series length per primary study (see Figure 52) illustrated that as the modal series length increased from small, with a mode of 10, to medium, with a mode of 20, to large, with a mode of 30, the confidence interval width estimates of the amount of estimated autocorrelation decreased with means of .107 (*SD* = 0.045) to .088 (*SD* = 0.038) to .074 (*SD* = 0.033), respectively.
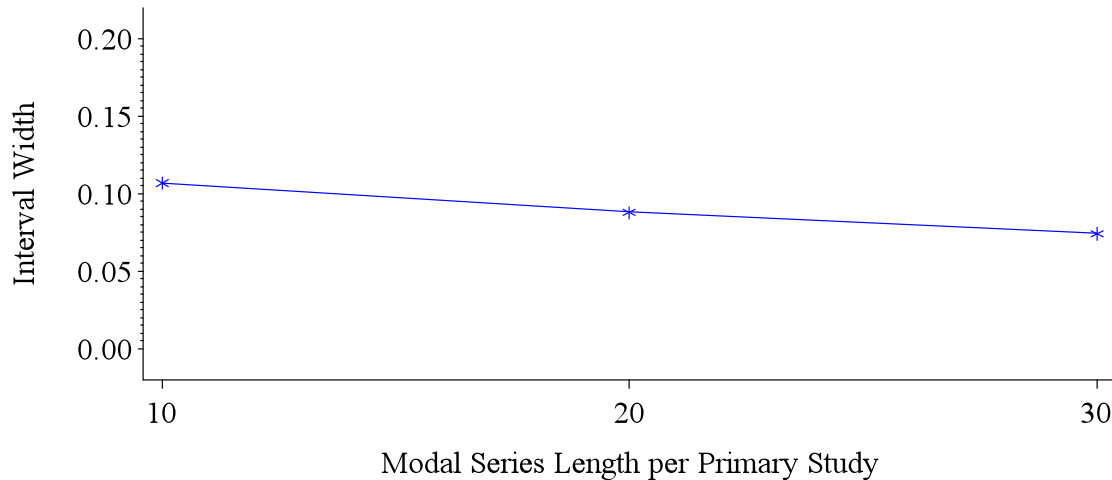
Figure 52. Line graph showing the estimated confidence interval widths for the amount of estimated autocorrelation as a function of the modal series length per primary study.

Relationships Among Dependent Variables

One may have anticipated that relationships existed between certain dependent variables for example confidence interval coverage and confidence interval width. In addition, when reviewing the results of this study it became apparent that relationships existed among the dependent variables per estimated effect. Therefore, correlations among the dependent variables were examined for each fixed effect and variance component. Correlation coefficients were compared to Cohen's (1988) standards for interpreting correlation coefficients with a weak relationship having an $r = .1$, a moderate relationship having an $r = .3$, and a strong relationship having an $r = .5$.

Fixed Effects

Overall average baseline level. The relationships among the overall average baseline level dependent variables are summarized in Table 23. A perusal of Table 23 indicates a strong and positive relationship between confidence interval coverage and width of the overall average baseline level with a correlation of .671. Specifically, as

147

confidence interval coverage of the overall average baseline level increased so did the

width of the overall average baseline level. In addition, the relationships between bias

and confidence interval coverage and width of the overall average baseline level were

weak.

Table 23
*Summary of Correlations, Means, and Standard Deviations for the Overall Average
Baseline Level Dependent Variables*

Dependent Variable	1	2	3
1. Bias	-	.063	.056
2. Confidence Interval Coverage		-	.671
3. Confidence Interval Width			-
Mean	0.000	.961	0.427
Standard Deviation	0.002	0.005	0.291

Note. Values in the table are based on 108 conditions in the simulation

Overall average treatment effect. Similar to the correlation results for the

overall average baseline level, a strong and positive relationship existed between

confidence interval coverage and width of the overall average treatment effect with a

correlation of .612. Specifically, as confidence interval coverage of the overall average

treatment effect increased so did the confidence interval width of the overall average

treatment effect. However, relationships involving the bias estimates of the overall

average treatment effect were weak. Relationships among the overall average treatment

effect dependent variables are summarized in Table 24.

Table 24
*Summary of Correlations, Means, and Standard Deviations for the Overall Average
Treatment Effect Dependent Variables*

Dependent Variable	1	2	3
1. Bias	-	.149	.012
2. Confidence Interval Coverage		-	.612
3. Confidence Interval Width			-
Mean	0.000	.960	0.459
Standard Deviation	0.001	0.005	0.293

Note. Values in the table are based on 108 conditions in the simulation

Variance Components

Between study variance in the overall average baseline level. The

relationships among the between study variance in the overall average baseline level

dependent variables are summarized in Table 25. It should be noted that due to

extremely large confidence interval width estimates of the between study variance in the

overall average baseline level correlations between the confidence interval widths and the

other dependent variables was not calculated. An examination of Table 25 indicates

several noteworthy relationships among the between study variance in the overall average

baseline level dependent variables. Specifically, the relationship between relative bias

and confidence interval coverage of the between study variance in the overall average

baseline level is strong and negative with a correlation of -.781. This relationship

indicates that as relative bias increased confidence interval coverage decreased. A

moderate, positive relationship existed between bias and relative bias of the between

study variance in the overall average baseline level ($r = .357$) indicating that as bias

increased so did relative bias. Conversely, a moderate, negative relationship existed

between bias and confidence interval coverage of the between study variance in the

overall average baseline level ($r = -.306$) indicating that as bias increased confidence

interval coverage decreased.

Table 25

Summary of Correlations, Means, and Standard Deviations for the Between Study Variance in the Overall Average Baseline Level Dependent Variables

Dependent Variable	1	2	3
1. Bias	-	.357	-.306
2. Relative Bias		-	-.781
3. Confidence Interval Coverage			-
Mean	-0.241	-.858	.999
Standard Deviation	0.201	0.078	0.004

Note. Values in the table are based on 108 conditions in the simulation

Between study variance in the overall average treatment effect. It should be noted that due to extremely large confidence interval width estimates of the between study variance in the overall average treatment effect correlations between the confidence interval widths and the other dependent variables was unwarranted. Similar to the correlation results for the between study variance in the overall average baseline level, several noteworthy relationships among the between study variance in the overall average treatment effect dependent variables existed. Specifically, the relationship between relative bias and confidence interval coverage of the between study variance in the overall average treatment effect is strong and negative with a correlation of -.850. This relationship indicates that as relative bias increased confidence interval coverage decreased. A strong, positive relationship existed between bias and relative bias of the between study variance in the overall average treatment effect (r = .536) indicating that as bias increased so did relative bias. Conversely, a moderate, negative relationship existed between bias and confidence interval coverage of the between study variance in the overall average baseline level (r = -.384) indicating that as bias increased confidence interval coverage decreased. Relationships among the dependent variables are summarized in Table 26.

Table 26

Summary of Correlations, Means, and Standard Deviations for the Between Study
Variance in the Overall Average Treatment Effect Dependent Variables

Dependent Variable	1	2	3
1. Bias	-	.536	-.384
2. Relative Bias		-	-.851
3. Confidence Interval Coverage			-
Mean	-0.237	-.822	.995
Standard Deviation	0.201	0.106	0.012

Note. Values in the table are based on 108 conditions in the simulation

Between person within study variance in the average baseline level. The

relationships among the between person within study variance in the average baseline

level dependent variables are summarized in Table 27. It should be noted that due to

extremely large confidence interval width estimates of the between person within study

variance in the average baseline level correlations between the confidence interval widths

and the other dependent variables was unwarranted. A perusal of Table 27 indicates

several notable relationships among the between person within study variance in the

average baseline level dependent variables. Specifically, the relationship between

relative bias and confidence interval coverage of the between person within study

variance in the average baseline level was strong and negative with a correlation of -.802.

This relationship indicates that as relative bias increased confidence interval coverage

decreased. A moderate, positive relationship existed between bias and relative bias of the

between person within study variance in the average baseline level (r = .364) indicating

that as bias increased so did relative bias. Conversely, a moderate, negative relationship

existed between bias and confidence interval coverage of the between person within

study variance in the average baseline level (r = -.409) indicating that as bias increased

confidence interval coverage decreased.

Table 27

Summary of Correlations, Means, and Standard Deviations for the Between Person Within Study Variance in the Average Baseline Level Dependent Variables

Dependent Variable	1	2	3
1. Bias	-	.364	-.409
2. Relative Bias		-	-.802
3. Confidence Interval Coverage			-
Mean	0.243	.217	.612
Standard Deviation	0.202	0.017	0.241

Note. Values in the table are based on 108 conditions in the simulation

Between person within study variance in the average treatment effect. It should be noted that due to extremely large confidence interval width estimates of the between person within study variance in the average treatment effect correlations between the confidence interval widths and the other dependent variables was unwarranted. Similar to the correlation results for the between participants within study variance in the average baseline level, several notable relationships among the between person within study variance in the average treatment effect dependent variables existed. Specifically, the relationship between relative bias and confidence interval coverage of the between person within study variance in the average treatment effect was strong and negative with a correlation of -.760. This relationship indicated that as relative bias increased confidence interval coverage decreased. A strong, positive relationship existed between bias and relative bias of the between person within study variance in the average treatment effect ($r = .542$) indicating that as bias increased so did relative bias. Conversely, a moderate, negative relationship existed between bias and confidence interval coverage of the between person within study variance in the average treatment effect ($r = -.573$) indicating that as bias increased confidence interval coverage decreased. Relationships among the dependent variables are summarized in Table 28.

Table 28
Summary of Correlations, Means, and Standard Deviations for the Between Participant Within Study Variance in the Average Treatment Effect Dependent Variables

Dependent Variable	1	2	3
1. Bias	-	.542	-.573
2. Relative Bias		-	-.760
3. Confidence Interval Coverage			-
Mean	0.238	.208	.675
Standard Deviation	0.201	0.023	0.222

Note. Values in the table are based on 108 conditions in the simulation

Within person residual variance. The relationships among the within person residual variance dependent variables are summarized in Table 29. A perusal of Table 29 indicates a strong and negative relationship between bias and confidence interval coverage of the within person residual variance with a correlation of -.919. Specifically, as bias of the within person residual variance increased so did the confidence interval coverage. In addition, a moderate, positive relationship existed between bias and confidence interval width of the within person residual variance with a correlation of .402. Conversely, the relationship between confidence interval coverage and width of the within person residual variance were weak.

Table 29
Summary of Correlations, Means, and Standard Deviations for the Within Person Residual Variance Dependent Variables

Dependent Variable	1	2	3
1. Bias	-	-.919	.402
2. Confidence Interval Coverage		-	-.133
3. Confidence Interval Width			-
Mean	0.078	.550	0.146
Standard Deviation	0.082	0.398	0.075

Note. Values in the table are based on 108 conditions in the simulation

Amount of estimated autocorrelation. The relationships among the amount of estimated autocorrelation dependent variables are summarized in Table 30. A perusal of

153

Table 30 indicates several notable relationships among the amount of estimated

autocorrelation dependent variables. Specifically, the relationship between bias and

relative bias of the estimated autocorrelation was strong and positive with a correlation of

.931. This relationship indicated that as bias increased so did relative bias. A strong,

negative relationship existed between relative bias and confidence interval width of the

amount of estimated autocorrelation ($r = -.625$) indicating that as relative bias increased

confidence interval width decreased. Similarly, a moderate, negative relationship existed

between bias and confidence interval width ($r = -.444$) of the amount of estimated

autocorrelation indicating that as bias increased confidence interval width decreased.

Table 30
Summary of Correlations, Means, and Standard Deviations for the Amount of Estimated
Autocorrelation Dependent Variables

Dependent Variable	1	2	3	4
1. Bias	-	.931	.100	-.444
2. Relative Bias		-	.164	-.615
3. Confidence Interval Coverage			-	-.146
4. Confidence Interval Width				-
Mean	0.000	-.001	.950	0.090
Standard Deviation	0.000	0.002	0.003	0.042

Note. Values in the table are based on 108 conditions in the simulation

Summary

Please see Table 31 for a summary of the results. The table is comprised of a

column for each dependent variable (i.e., bias and/or relative bias of the point estimates,

confidence interval coverage, and confidence interval width) and a row for each effect

(i.e., fixed effects and variance components) with a brief summary provided in each cell.

Table 31
Summary of Results

Parameter Estimate	Bias	Relative Bias	Confidence Interval Coverage	Confidence Interval Width
Fixed Effects				
Overall Average Baseline Level	- Unbiased (M = 0.00)	N/A	-Tended to overcover (M = .961) - Came close to nominal level as level-3 sample size increased	- Relatively small (M = 0.428) - Intervals smallest when level-3 sample size was largest and less variance was at level-2 and level-3
Overall Average Treatment Effect	- Unbiased (M = 0.00)	N/A	-Tended to overcover (M = .960) - Came close to nominal level as level-3 sample size increased	- Relatively small (M = 0.459) -Intervals smallest when level-3 sample size was largest and less variance was at level-2 and level-3
Variance Components				
Level-3				
Between Study Variance in the Overall Average Baseline Level	- Biased (M = - 0.241) - Tended to be underestimated - Less biased when most of the variances in the error terms was at level-1	-Biased (M = -.858) - Less biased as level-3, level-2 sample sizes decreased and when most of the variances in the error terms was at level-1	- Tended to overcover (M = .998) - Moved towards the nominal level when the level-3 and level-2 sample sizes were smallest and most of the variances in the error terms was at level-1	- Too large to provide any information (M = 1.568 X 10^{285})
Between Study Variance in the Overall Average Treatment Effect	- Biased (M = - 0.237) - Tended to be underestimated - Less biased when most of the variances in the error terms was at level-1	-Biased (M = -.822) - Less biased as level-3, level-2 sample sizes decreased and when most of the variances in the error terms was at level-1	- Tended to overcover (M = .995) - Moved towards the nominal level when the level-3 and level-2 sample sizes were smallest and most of the variance in the error terms was at level-1	- Too large to provide any information (M = 2.449 X 10^{285})

Table 31 (Continued)

Parameter	Bias	Relative Bias	Confidence Interval Coverage	Confidence Interval Width
Level-2				
Between Person Within Study Variance in the Average Baseline Level	- Biased (M = 0.243) - Tended to be overestimated - Less biased when most of the variance in the error terms was at level-1	-Biased (M = .217) - Less biased as level-3, level-2 sample sizes decreased and when most of the variance was at level-1	-Tended to undercover (M = .612) - Moved towards the nominal level when the level-3 and level-2 sample sizes were smallest and when most of the variance was at level-1	- Too large to provide any information (M = 5.027 X 10^{280})
Between Person Within Study Variance in the Average Treatment Effect	- Biased (M = 0.238) - Tended to be overestimated - Less biased when most of the variance in the error terms was at level-1	-Biased (M = .208) - Less biased as level-3, level-2 sample sizes decreased and when most of the variance in the error terms was at level-1	- Tended to undercover (M = .675) - Moved towards the nominal level when the level-3 sample size was smallest and when most of the variance in the error terms was at level-1	- Too large to provide any information (M = 9.419 X 10^{282})
Level-1				
Within Person Residual Variance	- Biased (M = 0.078) - Tended to be slightly overestimated - Least biased when autocorrelation was 0	N/A	- Tended to undercover (M = .550) - Close to the nominal level when autocorrelation was 0	- Relatively small intervals (M = 0.146) - Smallest intervals when level-3 and level-1 sample sizes were largest and level of autocorrelation was 0
Amount of Estimated Autocorrelation	- Unbiased (M = 0.001)	-Unbiased (M = -.001)	- Close to the nominal level, regardless of condition (M = .950)	- Relatively small intervals (M = 0.090) - Smallest intervals when level-3 and level-1 sample sizes were largest

Fixed Effects

Results indicated that the fixed effects, both the overall average baseline level (γ_{000}) and the overall average treatment effect (γ_{100}), were unbiased regardless of condition, with average bias values of zero. However, confidence interval coverage rates of the fixed effects tended to overcover. Variation in the confidence interval coverage rates of the overall average baseline level was explored by examining the only factor with a medium or larger effect size, the number of primary studies per meta-analysis (level-3 sample size). Further examination of this effect illustrated that as the level-3 sample size became larger the closer the confidence interval coverage rates came to a .95 coverage rate representing an alpha level of .05. Likewise, the number of primary studies per meta-analysis had the same impact on the confidence interval coverage of the overall average treatment effect indicating that as the level-3 sample size increased the closer the confidence interval coverage rates came to a .95 coverage rate.

Confidence interval widths of both fixed effects, the overall average baseline level and the overall treatment effect, were relatively small. To gain a better understanding for widths of this size, it is helpful to recall that the level-1 variance was set to 1.0 and both fixed effects were set to 1.0. Therefore, average confidence interval widths of 0.459 for the overall average treatment effect would produce an overall average treatment effect interval that ranged from around 0.770 to 1.230. Variation in the confidence interval width estimates of the overall average baseline level were explored by creating a line graph showing the confidence interval width estimates as a function of the interaction between the number of primary studies per meta-analysis and the variances of the error terms. The results indicated that confidence interval widths were smallest when the

number of primary studies per meta-analysis was 80 and most of the variance was at level-1 or less variance was at level-2 and level-3. Similar results were found for the overall average treatment effect.

Variance Components

Level-3 and level-2 variance components tended to be biased, with level-3 variance components tending to be underestimated and level-2 variance components tending to be overestimated. Parameter estimates of the between study variance in the overall average baseline level ($\tau_{\beta00}$) and the between study variance in the overall average treatment effect ($\tau_{\beta10}$) tended to become increasingly underestimated and progressively more biased when the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. Parameter estimates of the between person within study variance in the average baseline level ($\tau_{\pi00}$) and in the average treatment effect ($\tau_{\pi10}$) tended to become increasingly overestimated and progressively more biased when the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. The within person residual variance (σ^2) was also biased with an average bias value being slightly above zero. Parameter estimates of the within person residual variance tended to become increasingly overestimated and slightly more biased when the level of autocorrelation increased. However, the amount of estimated autocorrelation ($\hat{\rho}$) in the three level model was on average unbiased with the bias estimate not exceeding 1% of the known parameter value.

Relative bias was also evaluated for any parameter whose known value was different from one. Results indicated that the parameter estimates of the level-3 variance components (between study variance in the overall average baseline level and between

158

study variance in the overall average treatment effect) tended to be underestimated and became progressively more underestimated and biased when the number of primary studies per meta-analysis increased, the modal number of participants per primary study increased, and the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. In addition, the parameter estimates of the level-2 variance components (between person within study variance in the average baseline level and between person within study variance in the average treatment effect) tended to be overestimated and became progressively more overestimated and biased when the number of primary studies per meta-analysis increased, the modal number of participants per primary study increased, and the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2.

Confidence intervals of the level-3 variance components tended to overcover but were closest to a .95 coverage rate when the number of primary studies per meta-analysis was 10, the modal number of participants per primary study was small, and most of the variance was at level-1. In addition, as the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 confidence interval coverage rates for the level-3 variance components increased regardless of the number of primary studies per meta-analysis or modal number of participants per primary study. Confidence intervals of the level-2 variance components and the residual variance tended to undercover. The confidence interval coverage rates of the between person within study variance in the average baseline level tended to decrease when the number of primary studies per meta-analysis increased, the modal number or participants per primary study increased and the variances of the error terms shifted from most of the variance at level-1

to most of the variance at level-2. The confidence interval coverage rates of the between person within study variance in the average treatment effect tended to decrease as the number of primary studies per meta-analysis increased and the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. In addition, the confidence interval coverage rates of the within person residual variance decreased as the number of primary studies per meta-analysis and the level of autocorrelation increased. Conversely, the confidence interval coverage rates of the amount of estimated autocorrelation were close to a .95 coverage rate regardless of condition.

Confidence interval widths were so large for the level-3 and level-2 variance components that they provided no information. However, the confidence interval width estimates for the within person residual variance produced relatively small intervals ($M =$ 0.146) and tended to decrease as the number of primary studies per meta-analysis, modal series length per primary study, and the level of autocorrelation increased. For example, consider the fact that the within person residual variance was set to 1.0 therefore a small series length, with a mode of 10, would yield a confidence interval from about .914 to 1.086, but a medium series length, with a mode of 20, would produce a confidence interval from .929 to 1.072, and a large series length, with a mode of 30, would provide an even tighter confidence interval from .939 to 1.061. Similarly, the confidence interval width estimates for the amount of estimated autocorrelation were also relatively small (M $= 0.090$) and tended to decrease as the number of primary studies per meta-analysis and series length per primary study increased. Therefore, based on the results of the confidence interval widths of the amount of estimated autocorrelation, when the amount

160

of estimated autocorrelation was set to 0 a level-3 sample size of 10 primary studies would produce a confidence interval from about -.070 to .070, but a level-3 sample size of 30 primary studies would lead to a confidence interval from around -.040 to .040, and a level-3 sample size of 80 would yield even greater precision with a confidence interval from about -.025 to .025.

CHAPTER FIVE: DISCUSSION

This chapter outlines a summary of the study and results, along with a discussion of the findings, limitations of the study, and implications.

Summary of the Study

Quantitative integration of study results, termed meta-analysis, involves the combining of data across multiple studies to evaluate and summarize research findings. The term meta-analysis was first coined by Glass (1976) and was defined as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p.3). This type of research is an important way to determine relationships among variables and the effectiveness of interventions across studies. It also allows researchers to integrate study findings with the goal of generalization. Quantitative integration of study findings should cross research domains and include all types of quantitative research, including single-case research. However, meta-analysis of single-case research has resulted in much disagreement in the field.

Although the use of single-case designs has grown over the past decades, the majority of literature on meta-analysis focuses on group comparison studies and leaves out single-case research (Van den Noortgate & Onghena, 2008). This lack of literature related to single-case designs is often why these designs are excluded from meta-analyses. This exclusion of single-case designs is concerning when one considers the plethora of information single-case research can add to the literature. Single-case designs

not only provide information related to average treatment effects but also offers information related to how that treatment effect is related to specific cases. Meta-analyses of single-case designs offer the ability to summarize and evaluate the overall effect without the loss of that specific case information. In addition, the meta-analysis of single-case data increases the generalizabilty of research findings.

Researchers have proposed a variety of methods to meta-analyze single-case data. Van den Noortgate and Onghena's (2008) proposed method of using multilevel modeling to meta-analyze single-case data offers many advantages. The use of multilevel modeling provides the flexibility of appropriately modeling the autocorrelational nature of single-case data, can take into consideration multiple effect sizes per study, and can apply appropriate meta-analytic models, such as fixed or random effects models. Although the use of multilevel modeling offers advantages in the analysis of single-case data, there is still concern as to whether the use of multilevel modeling is appropriate for single-case data. Specifically, multilevel modeling is based on large sample theory, which is not representative of single-case data. Therefore, it was necessary to further investigate the appropriateness of inferences made from multilevel modeling when applied to single-case data.

The purpose of this study was to examine the appropriateness of Van den Noortgate and Onghena's (2008) raw data multilevel modeling approach to the meta-analysis of single-case data. Specifically, the study examined the fixed effects (i.e., the overall average baseline level and the overall average treatment effect) and the variance components (e.g., the between person within study variance in the average baseline level, the between study variance in the overall average baseline level, the between person

163

within study variance in the average treatment effect) in a three level multilevel model. More specifically, bias of point estimates, confidence interval coverage rates, and confidence interval widths were examined as a function of specific design and data factors.

Monte Carlo simulation methods were used to examine the appropriateness of multilevel modeling inferences. The use of simulation methods allowed for the control and manipulation of specific design and data factors. The Monte Carlo study included five factors in the design. These factors were (a) number of primary studies per meta-analysis (10, 30, and 80); (b) modal number of participants per primary study (small [mode = 4] and large [mode = 8]); (c) modal series length per primary study (small [mode = 10], medium [mode = 20], and large [mode = 30]); (d) level of autocorrelation (0, .2, and .4); and (e) variances of the error terms (most of the variance at level-1 [$\sigma^2 = 1$, $\tau_{\pi00} = \tau_{\pi10} = .2$, and $\tau_{\beta00} = \tau_{\beta10} = .05$] and most of the variance at level-2 [$\sigma^2 = 1$, $\tau_{\pi00} = \tau_{\pi10} = 2$, and $\tau_{\beta00} = \tau_{\beta10} = .5$]). The values chosen for each of these factors were based on previous simulation research and observed factors of actual single-case meta-analyses.

The data for this study were generated based on Van den Noortgate and Onghena's (2008) raw data, three level meta-analytic single-case model shown in Equations 24 through 28. Each data set was analyzed using the same model that was used for data generation (see Equations 24 - 28). The three level model was estimated using restricted maximum likelihood (REML) via PROC MIXED with the Kenward-Roger degrees of freedom method in SAS version 9.2 (SAS Institute Inc., 2008). In addition, a first order auto-regressive model for the level-1 errors was specified. Based

on the current model, the treatment effect was modeled as a change in level, and estimates were obtained for autocorrelation, variance within participants, variance in baseline levels across participants and studies, and variance in treatment effects across participants and studies.

The appropriateness of Van den Noortgate and Onghena's (2008) raw data multilevel modeling approach to the meta-analysis of single-case data was evaluated by examining bias and/or relative bias of the point estimates, confidence interval coverage, and confidence interval width of both the fixed effects and the variance components. This was accomplished by creating box plots, across all conditions, for each dependent variable. Then, the results of the simulation were analyzed using PROC GLM in SAS 9.2 for both the fixed effects and the variance components such that the dependent variables were bias, relative bias (where appropriate), confidence interval coverage, and confidence interval width and the independent variables were the five factors. Models were built with the purpose of finding effects whose eta-squared values were .06 or greater. The effects size, eta-squared (η^2), was calculated to determine the proportion of variability associated with each effect. Those values were compared to Cohen's (1988) standards for interpreting eta-squared values with a small effect size having an $\eta^2 = .01$, .a medium effect size having an $\eta^2 = .06$, and a large effect size having an $\eta^2 = .14$ or greater. Each model was first created as a main effects only model. If this model explained 94% of the total variability then no further complex models were investigated. However, if less than 94% of the total variability was explained then interactions were included in the model. Two-way interactions were added to the model first followed by three-way and then four-way interactions until at least 94% of the variability was explained.

Research Questions

1. To what extent are the fixed effect estimates from a three level meta-analytic single-case model biased as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

2. To what extent does the confidence interval coverage of the fixed effect estimates from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

3. To what extent does the confidence interval width of the fixed effect estimates from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

4. To what extent are the variance components from a three level meta-analytic single-case model biased as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

5. To what extent does the confidence interval coverage of the variance components from a three level meta-analytic single-case model vary as a function of specific

design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

6. To what extent does the confidence interval width of the variance components from a three level meta-analytic single-case model vary as a function of specific design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, level of autocorrelation, and variances of the error terms)?

Discussion of Study Results

Fixed Effects

The extent to which the fixed effects from a three level meta-analytic single-case model were biased, as a function of the specific design factors, was evaluated by the average amount that the estimated parameter differed from the known parameter. The results indicated that regardless of condition the fixed effects were unbiased with average bias values of zero. The unbiased fixed effect estimates revealed in this research are consistent with previous research regarding the utility of the inferences made from fixed effects in two level models (Ferron et al., 2009; Raudenbush & Bryk, 2002). Therefore, the use of fixed effects from a three level meta-analytic single-case model are likely to provide unbiased estimates of the average baseline level and average treatment effect across studies, if the model is correctly specified.

The proportion of the 95% confidence intervals that contained the parameter value estimated the confidence interval coverage of the fixed effects from a three level meta-analytic single-case model. The confidence interval coverage rates of the fixed

effects, both the overall average baseline level and the overall average treatment effect, tended to overcover with means of .961 and .960, respectively. Further examination of the extent to which the fixed effects varied as a function of the specific design factors illustrated that the 95% confidence interval coverage rates of the fixed effects came close to a .95 coverage rate as the level-3 sample size increased. These findings suggest that whenever possible researchers should increase the level-3 sample size or number of primary studies included in the meta-analysis. In addition, these findings validate previous literature related to two level models for single-case data that states larger upper level units lead to greater accuracy and precision (Ferron et al., 2009).

These findings are also consistent with general methodological research on more traditional designs of repeated measurements using multilevel models and the Kenward Roger degrees of freedom (Fouladi & Shieh, 2004; Gomez, Schaalje, & Fellingham, 2005; Kenward & Roger, 1997; Kowalchuk, Keselman, Algina, & Wolfinger, 2004; Schaalje, McBride, &Fellingham, 2001). These previous simulation studies have indicated that across a variety of conditions and sample sizes Type I error rates have been close to the nominal alpha level but variability in performance was noted. For example, Gomez, Schaalje, and Fellingham (2005) examined a three-group design with three participants per group and each participant measured at three points in time and they found that Type I error control varied based on the type of covariance structure. In particular, results indicated that when the data were generated and analyzed assuming compound symmetry the estimated Type I error rate was .052 ($\alpha = .05$). However, when the data were generated and analyzed assuming a 1st order autoregressive with random effects model the estimated Type I error rate was .1165 ($\alpha = .05$).

The average difference between the upper and lower limits of the 95% confidence intervals defined the confidence interval widths of the fixed effects from a three level meta-analytic single-case model. The confidence interval widths of the fixed effects, both the overall average baseline level and the overall average treatment were relatively small with average confidence interval width estimates of 0.428 and 0.459, respectively. To gain a better understanding for widths of this size, it is helpful to recall that the level-1 variance was set to 1.0 and both fixed effects were set to 1.0. Therefore, average confidence interval widths of 0.459 for the overall average treatment effect would produce interval estimates that ranged from around 0.770 to 1.230. Further examination of the extent to which the confidence interval widths of the fixed effects varied as a function of the specific design factors indicated that the interaction between the level-3 sample size and the variances of the error terms impacted the variability in confidence interval widths of the fixed effects. Specifically, confidence interval widths of the fixed effects were smallest when the level-3 sample sizes were largest (mode = 80) and most of the variance in the error terms was at level-1 or less variance at level-2 and level-3. Similar to previous research examining two level models for single-case data (Ferron et al., 2009; Ferron et al., 2010), which found that confidence interval widths of the treatment effect decreased with more participants, more observations per participant, and smaller variance components, this study's results would suggest that a larger number of upper level units and less variability between persons and studies would produce more precise confidence intervals of the fixed effects.

Variance Components

      The extent to which the variance components from a three level single-case meta-analytic model were biased, as a function of the specific design factors, was evaluated by the average amount that the estimated parameter differed from the known parameter. As expected, the level-3 and level-2 variance components tended to be biased. Specifically, the level-3 variance components, both in the between study variance in the overall average baseline level and the between study variance in the overall average treatment effect, tended to be underestimated with means of -0.241 and -0.237, respectively. The level-2 variance components, both the between person within study variance in the average baseline level and the between person within study variance in the average treatment effect, tended to be overestimated with means of 0.243 and 0.238, respectively. These findings are not too surprising given other research from a broader methodological perspective. Previous Monte Carlo research on growth curve models with studies having as few as 30 participants and series lengths of 4 or 8 (Kwok, West, & Green, 2007) and series length of 5 or 8 (Murphy & Pituch, 2009) have all reported substantial bias in the variance components when the model was correctly specified and the number of participants was small ($N = 30$). In the present study, bias in the level-3 variance components was mainly impacted by one factor, the variances of the error terms. As the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2, the level-3 variance components tended to become increasingly underestimated and progressively more biased. Conversely, the level-2 variance components became increasingly overestimated and progressively more biased as the

variances in the error terms shifted from most of the variance at level-1 to most of the variance at level-2.

Similar to previous research on two levels models with single-case data (Ferron et al., 2009), level-1 variance or within person residual variance was slightly biased but differing from previous research the bias in the estimates of within person residual variance remained constant at around 8% regardless of level-3 or level-2 sample size. However, results from this study did reveal the within person residual variance of the three level model became increasingly biased as the level of autocorrelation increased. This finding is not surprising given the notion that as autocorrelation increases the errors between observations within a person become more similar and therefore make it difficult to produce unbiased within person variability estimates. However, the amount of estimated autocorrelation in the three level meta-analytic single-case model was on average unbiased. Both the within person residual variance and amount of estimated autocorrelation bias results were not consistent with previous literature on two level models that found both parameters to be substantially biased (Ferron et al, 2009). However, this current study did focus on a three level model as opposed to the previously investigated two level model and therefore more information was ultimately available in the estimation of those parameters.

The extent to which the variance components from a three level meta-analytic single-case model were biased, based on specific design factors, were also evaluated by examining relative bias for any parameter whose known value was different from one so as to gain an index of bias in relation to the known parameter value. As was expected, based on previous literature (Ferron et al., 2009; Raudenbush & Bryk, 2002), the variance

components were biased; however, the trend in bias of the variance components was not expected. Previous Monte Carlo research on two level models for singe-case data has shown biased variance components at both level-2 and level-1 but with a decrease in bias of the point estimates as the upper level units increased (Ferron et al., 2009). In this study when the level-3 and level-2 sample size increased and the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2 the level-3 and level-2 variance components became increasing more biased, albeit in opposing directions. Specifically, the level-3 variance components became increasingly underestimated and the level-2 variance components became increasingly overestimated. However, the level-3 sample size only went as high as 80 and the level-2 sample size only went as high as 8 therefore there is no way of knowing if and when the variance components would have begun showing less bias with larger sample sizes. Another interesting finding of the present study that was contradictory to previous literature examining two level models for single-case data (Ferron et al., 2009), which found that the amount of autocorrelation tended to be biased and underestimated, was that on average the amount of estimated autocorrelation was unbiased with relative bias estimates not exceeding 1% of the known parameter value. This finding suggests that it is tenable to assume that estimates of the amount of estimated autocorrelation from this three level meta-analytic single-case model, under these specific design conditions, are unbiased, if the model is correctly specified.

The extent to which the confidence interval coverage estimates of the variance components from a three level meta-analytic single-case model produced accurate confidence intervals, as a function of specific design factors, was estimated by the

proportion of the 95% confidence intervals that contained the parameter value. The

level-3 variance components, both in the between study variance in the overall average

baseline level and the between study variance in the overall average treatment effect,

tended to overcover with means of .998 and .995, respectively. Further examination of

these effects indicated that the main factors that influenced the variability in confidence

interval coverage rates of the level-3 variance components were the level-3 sample size,

level-2 sample size, and the variances of the error terms. Specifically, confidence

interval coverage rates of the level-3 variance components were closest to a .95 coverage

rate when the level-3 sample size was smallest (10 primary studies), level-2 sample size

was smallest (mode = 10), and most of the variance in the error terms was at level-1.

Recall that bias of the level-3 variance components was smallest when the level-3 sample

size was smallest, level-2 sample size was smallest, and most of the variance in the error

terms was at level-1. Therefore, given the relative bias results, it was not surprising that

the confidence interval coverage was problematic for the level-3 variance components.

Similar results were found for the level-2 variance components and the within

person residual variance. The level-2 variance components, both the between person

within study variance in the average baseline level and the between person within study

variance in the average treatment effect, tended to undercover with means of .612 and

.675, respectively. Several design factors were found to have impacted the variability in

the confidence interval coverage rates of the level-2 variance components. The

confidence interval coverage rates of the between person within study variance in the

average baseline level tended to decrease and move farther away from a .95 coverage rate

when the level-3 sample size increased, the level-2 sample size increased, and the

variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. In addition, the confidence interval coverage rates of the other level-2 variance component, the between person within study variance in the average treatment effect, tended to decrease and move farther away from a .95 coverage rate as the level-3 sample size increased and the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. Recall that relative bias results of the level-2 variance components indicated that estimates of the level-2 variance components became more biased as the level-3 sample size increased, the level-2 sample size increased, and the variances of the error terms shifted from most of the variance at level-1 to most of the variance at level-2. Therefore, it was not surprising that confidence interval coverage of the level-2 variance components was troublesome. Additionally, these results are consistent with previous findings (Maas & Hox, 2004) from a broader methodological perspective on two-level organizational models, which found that coverage rates of the level-2 variance components tended to undercover with small sample sizes ($N = 30$).

Confidence interval coverage rates were the most problematic for the within person residual variance with average confidence interval coverage rates well below the nominal level ($M = .550$). However, confidence interval coverage rates of the within person residual variance were close to a .95 coverage rate when autocorrelation was zero. This finding was consistent given the bias results for the within person residual variance. Conversely, confidence interval coverage rates for the amount of estimated autocorrelation were close to a .95 coverage rate ($M = .950$), regardless of condition, which is not surprising given the bias results for the amount of estimated autocorrelation.

The average difference between the upper and lower limits of the 95% confidence intervals defined the confidence interval widths of the variance components from a three level meta-analytic single-case model. Confidence interval widths for the level-3 and level-2 variance components were so large that they provided no information. These findings are not surprising given previous research on two level models for single-case data (Ferron et al., 2009) where the results indicated that the confidence interval widths for the level-2 variance components were so large that they provided no information.

However, the confidence interval width estimates for the within person residual variance produced relatively small interval widths ($M = 0.146$) which tended to become even smaller as the level-3 and level-1 sample size, and level of autocorrelation increased. For example, consider the fact that the within person residual variance was set to 1.0 therefore a small series length, with a mode of 10, would yield a confidence interval from about .914 to 1.086, but a medium series length, with a mode of 20, would produce a confidence interval from .929 to 1.072, and a large series length, with a mode of 30, would provide an even tighter confidence interval from .939 to 1.061. These results are not too surprising considering the confidence interval coverage estimates for the within person residual variance tended to undercover.

Likewise, the confidence interval width estimates for the amount of estimated autocorrelation were small ($M = 0.090$) and tended to decrease as the level-3 and level-1 sample size increased. Therefore, based on the results of the confidence interval widths of the amount of estimated autocorrelation, when the level of autocorrelation was set to 0 a level-3 sample size of 10 primary studies would produce a confidence interval with from about -.070 to .070, but a level-3 sample size of 30 primary studies would lead to a

confidence interval from around -.040 to .040, and a level-3 sample size of 80 would yield even greater precision with a confidence interval from about -.025 to .025. These findings suggest that it is tenable to assume as the level-3 and level-1 sample sizes increase the more precise the estimates of amount of estimated autocorrelation become, if the model is correctly specified.

Limitations of the Study

Based on the design of this study, there are generalizabilty limitations to consider with regard to this research study. The Monte Carlo method used in this study provided control of specific factors to investigate the appropriateness of inferences made from a three level meta-analytic single-case model in specific situations. While this is a benefit of simulation studies it also limits the generalizabilty of the study findings. Therefore, the five design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, variances of the error terms, and level of autocorrelation) determine the types of single-case meta-analyses to which the study's findings can be generalized. In addition, another generalizabilty limitation of this study is the levels of the specific design factors. These levels were chosen to represent a range of possible values seen in single-case meta-analyses as well as previous simulation work. However, they are not exhaustive of all possible values for each design factor.

Another limitation to consider relates to the model under investigation. The specific model (see Equations 24 – 28) chosen for investigation in this research study makes several assumptions. First, Van den Noortgate and Onghena's (2008) raw data three level meta-analytic single-case model assumes that all dependent variables were

measured the same across primary studies included in the meta-analysis. Second, the model chosen for analysis was the most basic interrupted time-series model (e.g. no trends or changes in trends). The benefit of choosing this model is that it is the most basic model and therefore the most logical for an initial study into the three level meta-analytic modeling of single-case data. In addition, model and data generation assumed normality of the level-1 errors, multivariate normality of the level-2 errors, multivariate normality of the level-3 errors, and homoscedasticity of the errors at all levels. If the within person variance varied across the participants within studies or across studies, the autocorrelation varied, or a more complex time series model (e.g. 2nd order of higher) was needed then the model would be misspecified. The results don't allow for generalizations to performance when there is some degree of misspecification or there is use for more complex model specifications.

Implications

Researchers have suggested that use of multilevel modeling in meta-analyzing single-case data provides many advantages (Van den Noortgate & Onghena, 2003a, 2007, 2008). Specifically, multilevel modeling provides the ability to combine the results from multiple participants and studies to gain information about not only the overall treatment effect but also if and how the treatment effect varies across participants and studies (Van den Noortgate & Onghena, 2008). Another advantage of multilevel modeling is that it can be used to aggregate data from single-case studies that include multiple participants. This use of multiple data sources or effect sizes from the same study is typically problematic and has not been addressed by other proposed single-case meta-analytic methods. Multilevel modeling is structured to account for that "nesting" of data within

studies by allowing variation within participants, between participants of the same study and between studies (Van den Noortgate & Onghena, 2008).

The results of this study suggest that the degree to which the findings of this study are supportive of using Van den Noortgate and Onghena's (2008) raw data multilevel modeling approach to meta-analyzing single-case data depends on the particular effect of interest. This in turn leads to specific implications for those who conduct meta-analyses of single-case studies, single-case researchers, and methodologists.

Implications for Researchers Conducting Single-Case Meta-Analyses

For researchers interested in the overall average baseline level and overall average treatment effect across studies, the results of this research study are encouraging. If researchers conducting single-case meta-analyses have data that conform to the assumptions of the model examined they should feel comfortable interpreting the overall average baseline levels and overall average treatment effects across studies. Still, researchers should be advised to increase the level-3 sample size or number of primary studies per meta-analysis whenever possible. With larger level-3 sample sizes, greater accuracy and precision could be gained in estimating the overall average baseline levels and treatment effects across studies. While single-case meta-analysts are constrained by the availability of primary studies they could adjust their methods for searching (e.g., expanding their search terms) whenever possible, but are limited by what the field has generated.

On the other hand statements about the variation in treatment effects across studies, which are also valued by meta-analysts and single-case researchers, should be viewed cautiously. Even assuming the model was correctly specified, the variance

components at all levels were biased and confidence intervals for those estimates were inaccurate. Specifically, the level-3 (between study) variance components tended to overcover and the level-2 (between person within study) variance components and the within person residual variance both tended to undercover and did not show signs of improvement with larger level-3 sample sizes.

Implications for Researchers Conducting Single-Case Studies

For researchers conducting single-case studies, the results of this study provide a few recommendations. The results of this study indicated that fixed effects were more precise any time the amount of variability in the model was smaller. Specifically, this study examined shifts in variability at level-2 and level-3 but one may anticipate that paying close attention to ways of reducing variability overall would produce greater precision when estimating the overall average baseline levels and treatment effects across studies. For example, single-case researchers should pay attention to baseline variability or stability in effort to decrease variability at level-1. Specifically, single-case researchers should consider increasing the number of data points in baseline to correctly specify the model in an effort to decrease the amount of variability at level-1.

Single-case researchers should also pay attention to the extent to which the intervention is delivered as intended often termed treatment fidelity or integrity (Kazdin, 2011). For example, if a treatment or intervention was administered exactly like it was intended to be administered the associated treatment effect would be different than a treatment effect associated with a treatment or intervention administered differently than intended. This modification in implementation or lack of treatment integrity could cause

increases in between person variability and ultimately decrease precision in the overall average baseline levels and treatment effects across studies.

Measurement error can also impact variability and finding ways to decrease that measurement error could ultimately decrease variability overall. For example, single-case researchers should be consistent in their methods of measurement in an effort to decrease between person within study and between study variability. Therefore, single-case researchers should make every effort to measure outcomes at the same time of day and for the same amount of time across participants and even across studies assessing similar types of interventions.

A final recommendation to single-case researchers is to consider previous single-case research that has focused in their particular area of interest when determining the most appropriate outcome measure. Specifically, if single-case researchers from similar areas of interest (e.g., reading, math) measured their outcomes variables the same across studies then single-case meta-analysts would have a larger number of primary studies to include in this specific meta-analytic model and could feel more confident in their interpretation of overall average baseline levels and treatment effects across studies.

Implications for Methodologists

For methodologists studying the use of multilevel modeling to meta-analyze single-case data more research needs to be conducted on more complex treatment effects, such as delayed changes in level, trends in the data that change linearly or nonlinearly with time, and transitory effects. Furthermore, violations of assumptions (e.g., nonnormality of the level-1, level-2, or level-3 errors, heteroscedasticity of errors at all levels) and various level-1 error models (e.g., high order autoregressive or moving

average models) needed to be investigated as well. Investigation of these more complex models would allow for a better understanding of the applicability of the models to a variety of conditions.

Future research on other approaches to estimating variance components would also be of interest. Clearly, the results of this study have indicated that the variance components at all levels are biased and provide inaccurate confidence intervals. Therefore, it would be interesting to investigate alternative methods for estimating variance such as the Bayesian approach.

Finally, this study focused on the use of a three level model to meta-analyze only single-case data. It would be interesting to investigate ways to meta-analyze single-case and large group design data together. This would allow meta-analysts the ability to synthesize research across a variety of research designs.

REFERENCES

Algozzine, B., Browder, D., Karvonen, M., Test, D. W., & Wood, W. M. (2001). Effects of interventions to promote self-determination for individuals with disabilities. *Review of Educational Research, 71*(2), 219-277.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single-case. *Behavior, Research, and Therapy, 31*, 621-631.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single Case Experimental Designs: Strategies for Studying Behavior Change* (3rd ed.). Boston, MA: Pearson.

Beretvas, S. N., & Chung, H. (2008a). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-based Communication Assessment and Intervention, 2*, 129-141.

Beretvas, S. N., & Chung, H. (2008b). An evaluation of modified R2-change effect size indices for single-subject experimental designs. *Evidence-based Communication Assessment and Intervention, 2*(3), 120-128.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L.V. Hedges, and J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nded., pp.221-236). New York: Russell Sage Foundation.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229-242.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R.

   Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New*

   directions for psychology and education (pp. 187-212). Hillsdale, NJ: Lawrence

   Erlbaum Associates.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case

   consultation outcomes: Applications to research and practice. *Journal of School*

   Psychology, 33, 269-285.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative

   synthesis of intrasubject design research. *Journal of Special Education, 19*, 387-

   400.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York:

   Lawrence Erlbaum Associates.

Cooper, H. (2007). *Evaluating and Interpreting Research Synthesis in Adult Learning*

   and Literacy. Boston, MA: National College Transition Network, New England

   Literacy Resource Center/ World Education Inc.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H.

   Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The Handbook of Research*

   Synthesis and Meta-Analysis (2nd ed., pp. 3-16). New York: Russell Sage

   Foundation.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with

   individual participant data versus aggregated data. *Psychological Methods, 14*,

   165-176.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61*, 966-974.

Crowne, D.P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Farmer, J. L., Owens, C. M., Ferron, J. M., & Allsopp, D. H. (2010a, April).*A methodological review of single-case meta-analyses*. Paper presented at the American Educational Research Association. Denver, CO.

Farmer, J., Owens, C.M., Ferron, J., & Allsopp, D. (2010b). *A review of social science single-case meta-analyses*. Manuscript in preparation.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*(2), 372-384.

Ferron, J. & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *Journal of Experimental Education, 64*, 231-239.

Ferron, J. M., Farmer, J., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study. *Behavior Research Methods, 42,* 930 - 943.

Ferron, J., Owens, C. M., & Bell, B. A. (2010, April). *Multilevel models for combining single-case data: A Monte Carlo examination of treatment effect estimates and inferences.* Paper presented at the Annual Meeting of the American Educational Research Association. Denver, CO.

Ferron, J. & Rendina-Gobioff, G. (2005). Interrupted time series design. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Behavioral Statistics* (Vol. 2, pp. 941-945). West Sussex, UK: Wiley & Sons Ltd.

Fouladi, R.T. &Shieh, Y. (2004). A comparison of two general approaches to mixed model longitudinal analyses under small sample size conditions. *Communications in Statistics: Simulation and Computation, 33*, 807-824.

Filter, K. J., & Horner, R. H. (2009). Function-based academic interventions for problem behavior. *Education and Treatment of Children, 32*(1), 1-19.

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *Journal of Applied Behavioral Science, 20*, 71-79.

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

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Gomez, E., Schaalje, G. B., &Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communication in Statistics: Simulation and Computation, 34*, 377 – 392.

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

Hedges, L.V.& Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 4*, 486-504.

Hershberger, S.L., Wallace, D.D., Green, S.B., & Marquis, J.G. (1999). Meta-analysis of single-case deigns. In R.H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 107-132). Thousand Oaks, CA: Sage.

Hox, J. (1998). Multilevel modeling: When and why? In I. Balderjahn, R. Mathar & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). New York: Springer.

Hsieh, W., Hemmeter, M. L., McCollum, J. A., & Ostrokosky, M. M. (2009). Using coaching to increase preschool teachers' use of emergent literacy teaching strategies. *Journal of Emotional and Behavioral Disorders, 17*(4), 197-212.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7*, 107-118.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3*, 104-116.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38-58.

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform:

> Evidence-based practice, meta-analyses, and single subject designs. *Psychology in*
>
> *the Schools, 44*(5), 483-493.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on

> the agreement between visual and statistical inference. *Journal of Applied*
>
> *Behavior Analysis, 11*, 277-283.

Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied*

> *Settings* (2nd ed.). New York: Oxford University Press.

Kenward, M. G. & Roger, J. H. (1997). Small sample inference for fixed effects from

> restricted maximum likelihood. *Biometrics, 53*, 983 – 997.

Kowalchuk, R. K., Keselman, H. J., Algina, J., &Wolfinger, R. D. (2004). The analysis

> of repeated measurements with mixed-model adjusted F tests. *Educational and*
>
> *Psychological Measurement, 64*, 224 – 242.

Kratochwill, T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arnston, P., et al.

> (1974). A further consideration in the application of an analysis-of-variance
>
> model for the intrasubject replication design. *Journal of Applied Behavior*
>
> *Analysis, 7*, 629-633.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention:

> The use of effect sizes for data analysis in single-subject research. *Journal of*
>
> *Experimental Education, 65*(1), 73-93.

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

> Sage.

Lundervold, D., & Bourland, G. (1988). Quantitative analysis of treatment of aggression, self-injury, and property destruction. *Behavior Modification, 12*, 590-617.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598-617.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*, 127-137.

Matt, G.E. & Cook, T.D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. Hedges, & J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed. pp.537-560). New York: Russell Sage Foundation.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Matyas, T. A., & Greenwood, K. M. (1997). Serial dependency in single-case time series. In D. Franklin, D. B. Allison & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-243). Mahwah, N.Y.: Lawrence Erlbaum.

Mok, M. (1995). *Sample size requirements for 2-level designs in educational research*. Unpublished manuscript, Macquarie University, Sydney, Australia.

Morgan, D.L. & Morgan, R.K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist, 56,* 119 – 127.

Nugent, W. (1996). Integrating single-case and group comparison designs for evaluation research. *Journal of Applied Behavioral Science, 32*, 209-226.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of Pain, 21*, 56-68.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40*, 194-204.

Parsonson, B. S.,& Baer, D. M. (1986). The graphic analysis of data. In A. Poling & W. R. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157-186). New York: Plenum Press.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Raudenbush, S.W.& Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Riley-Tillman, T.C. & Burns, M.K. (2009). *Evaluating educational interventions: Single-case design for measuring response to intervention.* New York, NY: The Guilford Press.

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

Rusch, F. R. & Kazdin, A.E. (1981). Toward a methodology of withdrawal designs for the assessment of response maintenance. *Journal of Applied Behavioral Analysis, 14,* 131-140.

Schaalje, G. B., McBride, J. B., &Fellingham, G. W. (2001). Approximations of distributions of test statistics in complex mixed linear models using SAS Proc

MIXED. In *Proceedings of the SAS Users Group International 26th Annual Conference* (Paper 262 – 26).Available at support.sas.com/events/sasglobalforum/previous/index.html.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.

Scruggs, T. E., Mastropieri, M. A., & Castro, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24-33.

Shadish, W. R. & Haddock, C. K. (2009). Combining estimates of effect sizes. In H. Cooper, L. Hedges, & J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp.557-578). New York: Russell Sage Foundation.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113*, 95-109.

Shadish, W.R., Rindskopf, D. M., & Hedges, L.V. (2008). The state of science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 3*, 188-196.

Shine, L.C., & Bower, S.M. (1971). A one-way analysis of variance for single-subject designs. *Educational and Psychological Measurement, 31,* 105-113.

Simmonds, M.C., Higgins, J.P.T., Stewart, L.A., Tierney, J.F., Clarke, M.J., & Thompson, S.G. (2005). Meta-analysis of individual participant data from

randomized trials: A review of methods used in practice. *Clinical Trials, 2*, 209-217.

Skinner, C. H. (2004). Single-subject designs: Procedures that allow school psychologists to contribute to the intervention evaluation and validation process. *Journal of Applied School Psychology, 20*, 1-10.

Suen, H. S., & Ary, D. (1987). Application of statistical power in assessing autocorrelation. *Behavioral Assessment, 9*, 125-130.

Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). N=1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 8*, 289-309.

Van den Noortgate, W. & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*(3), 325-346.

Van den Noortgate, W. & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effects sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35*(1), 1-10.

Van den Noortgate, W. & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8*(2), 196-209.

Van den Noortgate, W. & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 3*, 142-151.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inference. *Behavioral Assessment, 3*, 71-92.

APPENDICES

Appendix A: SAS Programming Code

```sas
procprinttolog=junk;
procprinttoprint = junk2;

data j0;
input Estimate Lower Upper;
datalines;
.   .   .
.   .   .
;

data j00;
input Estimate Lower Upper;
datalines;
.   .   .
.   .   .
.   .   .
.   .   .
.   .   .
.   .   .
;

%global _print_;
%let _print_ = off;

*+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
input for the macro are:
n = 5000 (number of samples to generate)
n3= 10,30,80
avgseries = 0, 1, or 2 (0 = small, 1 = medium, 2 = large)
avgpart= 0 or 1 (0 = small and 1 = large)
varerror= 0 or 1 (0= most of the error at level-1 and 1= most of the
error at level-2)
gamma = 1 (fixed effects - intercept [gamma000], effect [gamma100])
phi = 0, -.2,or -.4 (produces positive autocorrelation)
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
;

%macrohlmsim (n, n3, avgseries, avgpart, varerror, gamma, phi);

%doi=1%to&n;

prociml;

*++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
This part of the program creates the initial data set,which contains
the following variables:
n1: 10, 20, or 30 (number of time points or level-1 units)
n2: 4, 6, or 8 (nuumber of participant or level-2 units)
tau0 = .2 or .05  (level-2 variance in the intercept and treatment
effect)
tau1 = 2 or .5 (level-3 variance in the intercept and treatment effect)
  IDlevel3: level 3 ID
  IDlevel2:  level 2 ID
time:  potential level-1 predictor
```

```
phase: dichotomous level-1 predictor (0=baseline, 1=treatment)
y:   outcome
  +++++++++++++++++++++++++++++++++++++++++++++++++++++++;

create j1 var{IDlevel3 IDlevel2 time phase y tau0 tau1};

do ID3=1 to &n3;
n1=0;
n2=0;
if&n3= 10 then do;
if&avgpart = 0& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 | ID3
= 6 | ID3 = 7) then n2 = 4;
if&avgpart = 0& (ID3 = 8 | ID3 = 9) then n2 = 6;
if&avgpart = 0& (ID3 = 10) then n2 = 8;
if&avgpart = 1& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 | ID3
= 6 | ID3 = 7) then n2 = 8;
if&avgpart = 1& (ID3 = 8 | ID3 = 9) then n2 = 6;
if&avgpart = 1& (ID3 = 10) then n2 = 4;


if&avgseries = 0& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7) then n1 = 10;
if&avgseries = 0& (ID3 = 8 | ID3 = 9) then n1 = 20;
if&avgseries = 0& (ID3 = 10) then n1 = 30;
if&avgseries = 1& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7) then n1 = 20;
if&avgseries = 1& (ID3 = 8 | ID3 = 9) then n1 = 10;
if&avgseries = 1& (ID3 = 10) then n1 = 30;
if&avgseries = 2& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7) then n1 = 30;
if&avgseries = 2& (ID3 = 8 | ID3 = 9) then n1 = 10;
if&avgseries = 2& (ID3 = 10) then n1 = 20;
end;
if&n3= 30 then do;
if&avgpart = 0& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 | ID3
= 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12 |
ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3 =
19 | ID3 = 20 | ID3 = 21) then n2 = 4;
if&avgpart = 0& (ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26 |
ID3 = 27)then n2 = 6;
if&avgpart = 0& (ID3 = 28 | ID3 = 29 | ID3 = 30) then n2 = 8;
if&avgpart = 1& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 | ID3
= 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12 |
ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3 =
19 | ID3 = 20 | ID3 = 21) then n2 = 8;
if&avgpart = 1& (ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26 |
ID3 = 27)then n2 = 6;
if&avgpart = 1& (ID3 = 28 | ID3 = 29 | ID3 = 30) then n2 = 4;

if&avgseries = 0& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12
| ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3
= 19 | ID3 = 20 | ID3 = 21) then n1 = 10;
if&avgseries = 0& (ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26
| ID3 = 27)then n1 = 20;
if&avgseries = 0& (ID3 = 28 | ID3 = 29 | ID3 = 30) then n1 = 30;
if&avgseries = 1& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12
```

194

```
| ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3
= 19 | ID3 = 20 | ID3 = 21) then n1 = 20;
if&avgseries = 1& (ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26
| ID3 = 27)then n1 = 10;
if&avgseries = 1& (ID3 = 28 | ID3 = 29 | ID3 = 30) then n1 = 30;
if&avgseries = 2& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12
| ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3
= 19 | ID3 = 20 | ID3 = 21) then n1 = 30;
if&avgseries = 2& (ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26
| ID3 = 27)then n1 = 10;
if&avgseries = 2& (ID3 = 28 | ID3 = 29 | ID3 = 30) then n1 = 20;
end;
if&n3= 80 then do;
if&avgpart = 0& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 | ID3
= 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12 |
ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3 =
19 | ID3 = 20 | ID3 = 21 |
ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26 | ID3 = 27 | ID3 =
28 | ID3 = 29 | ID3 = 30 | ID3 = 31 | ID3 = 32 | ID3 = 33 | ID3 = 34 |
ID3 = 35 | ID3 = 36 | ID3 = 37 | ID3 = 38 | ID3 = 39 | ID3 = 40 | ID3 =
41 | ID3 = 42 | ID3 = 43 |
ID3 = 44 | ID3 = 45 | ID3 = 46 | ID3 = 47 | ID3 = 48 | ID3 = 49 | ID3 =
50 | ID3 = 51 | ID3 = 52 | ID3 = 53 | ID3 = 54 | ID3 = 55 | ID3 = 56)
then n2 = 4;
if&avgpart = 0& (ID3 = 57 | ID3 = 58 | ID3 = 59 | ID3 = 60 | ID3 = 61 |
ID3 = 62 | ID3 = 63 | ID3 = 64 | ID3 = 65 | ID3 = 66 | ID3 = 67 | ID3 =
68 | ID3 = 70 | ID3 = 71 | ID3 = 72) then n2 = 6;
if&avgpart = 0& (ID3 = 73 | ID3 = 74 | ID3 = 75 | ID3 = 76 | ID3 = 77 |
ID3 = 78 | ID3 = 79 | ID3 = 80) then n2 = 8;
if&avgpart = 1& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 | ID3
= 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12 |
ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3 =
19 | ID3 = 20 | ID3 = 21 |
ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26 | ID3 = 27 | ID3 =
28 | ID3 = 29 | ID3 = 30 | ID3 = 31 | ID3 = 32 | ID3 = 33 | ID3 = 34 |
ID3 = 35 | ID3 = 36 | ID3 = 37 | ID3 = 38 | ID3 = 39 | ID3 = 40 | ID3 =
41 | ID3 = 42 | ID3 = 43 |
ID3 = 44 | ID3 = 45 | ID3 = 46 | ID3 = 47 | ID3 = 48 | ID3 = 49 | ID3 =
50 | ID3 = 51 | ID3 = 52 | ID3 = 53 | ID3 = 54 | ID3 = 55 | ID3 = 56)
then n2 = 8;
if&avgpart = 1& (ID3 = 57 | ID3 = 58 | ID3 = 59 | ID3 = 60 | ID3 = 61 |
ID3 = 62 | ID3 = 63 | ID3 = 64 | ID3 = 65 | ID3 = 66 | ID3 = 67 | ID3 =
68 | ID3 = 70 | ID3 = 71 | ID3 = 72) then n2 = 6;
if&avgpart = 1& (ID3 = 73 | ID3 = 74 | ID3 = 75 | ID3 = 76 | ID3 = 77 |
ID3 = 78 | ID3 = 79 | ID3 = 80) then n2 = 4;

if&avgseries = 0& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12
| ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3
= 19 | ID3 = 20 | ID3 = 21 |
ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26 | ID3 = 27 | ID3 =
28 | ID3 = 29 | ID3 = 30 | ID3 = 31 | ID3 = 32 | ID3 = 33 | ID3 = 34 |
ID3 = 35 | ID3 = 36 | ID3 = 37 | ID3 = 38 | ID3 = 39 | ID3 = 40 | ID3 =
41 | ID3 = 42 | ID3 = 43 |
```

```
ID3 = 44 | ID3 = 45 | ID3 = 46 | ID3 = 47 | ID3 = 48 | ID3 = 49 | ID3 =
50 | ID3 = 51 | ID3 = 52 | ID3 = 53 | ID3 = 54 | ID3 = 55 | ID3 = 56)
then n1 = 10;
if&avgseries = 0& (ID3 = 57 | ID3 = 58 | ID3 = 59 | ID3 = 60 | ID3 = 61
| ID3 = 62 | ID3 = 63 | ID3 = 64 | ID3 = 65 | ID3 = 66 | ID3 = 67 | ID3
= 68 | ID3 = 70 | ID3 = 71 | ID3 = 72) then n1 = 20;
if&avgseries = 0& (ID3 = 73 | ID3 = 74 | ID3 = 75 | ID3 = 76 | ID3 = 77
| ID3 = 78 | ID3 = 79 | ID3 = 80) then n1 = 30;
if&avgseries = 1& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12
| ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3
= 19 | ID3 = 20 | ID3 = 21 |
ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26 | ID3 = 27 | ID3 =
28 | ID3 = 29 | ID3 = 30 | ID3 = 31 | ID3 = 32 | ID3 = 33 | ID3 = 34 |
ID3 = 35 | ID3 = 36 | ID3 = 37 | ID3 = 38 | ID3 = 39 | ID3 = 40 | ID3 =
41 | ID3 = 42 | ID3 = 43 |
ID3 = 44 | ID3 = 45 | ID3 = 46 | ID3 = 47 | ID3 = 48 | ID3 = 49 | ID3 =
50 | ID3 = 51 | ID3 = 52 | ID3 = 53 | ID3 = 54 | ID3 = 55 | ID3 = 56)
then n1 = 20;
if&avgseries = 1& (ID3 = 57 | ID3 = 58 | ID3 = 59 | ID3 = 60 | ID3 = 61
| ID3 = 62 | ID3 = 63 | ID3 = 64 | ID3 = 65 | ID3 = 66 | ID3 = 67 | ID3
= 68 | ID3 = 70 | ID3 = 71 | ID3 = 72) then n1 = 10;
if&avgseries = 1& (ID3 = 73 | ID3 = 74 | ID3 = 75 | ID3 = 76 | ID3 = 77
| ID3 = 78 | ID3 = 79 | ID3 = 80) then n1 = 30;
if&avgseries = 2& (ID3 = 1 | ID3 = 2 | ID3 = 3 | ID3 = 4 | ID3 = 5 |
ID3 = 6 | ID3 = 7 | ID3 = 8 | ID3 = 9 | ID3 = 10 | ID3 = 11 | ID3 = 12
| ID3 = 13 | ID3 = 14 | ID3 = 15 | ID3 = 16 | ID3 = 17 | ID3 = 18 | ID3
= 19 | ID3 = 20 | ID3 = 21 |
ID3 = 22 | ID3 = 23 | ID3 = 24 | ID3 = 25 | ID3 = 26 | ID3 = 27 | ID3 =
28 | ID3 = 29 | ID3 = 30 | ID3 = 31 | ID3 = 32 | ID3 = 33 | ID3 = 34 |
ID3 = 35 | ID3 = 36 | ID3 = 37 | ID3 = 38 | ID3 = 39 | ID3 = 40 | ID3 =
41 | ID3 = 42 | ID3 = 43 |
ID3 = 44 | ID3 = 45 | ID3 = 46 | ID3 = 47 | ID3 = 48 | ID3 = 49 | ID3 =
50 | ID3 = 51 | ID3 = 52 | ID3 = 53 | ID3 = 54 | ID3 = 55 | ID3 = 56)
then n1 = 30;
if&avgseries = 2& (ID3 = 57 | ID3 = 58 | ID3 = 59 | ID3 = 60 | ID3 = 61
| ID3 = 62 | ID3 = 63 | ID3 = 64 | ID3 = 65 | ID3 = 66 | ID3 = 67 | ID3
= 68 | ID3 = 70 | ID3 = 71 | ID3 = 72) then n1 = 10;
if&avgseries = 2& (ID3 = 73 | ID3 = 74 | ID3 = 75 | ID3 = 76 | ID3 = 77
| ID3 = 78 | ID3 = 79 | ID3 = 80) then n1 = 20;
end;

do ID2=1 to n2;
cut=0;
if n2=4 then do;
if n1=10 then cut = 2 + ID2;
if n1=20 then cut = 5 + ID2*2;
if n1=30 then cut = 7 + ID2*3;
end;
if n2=6 then do;
if n1=10 then cut = 1 + ID2;
if n1=20 then cut = 3 + ID2*2;
if n1=30 then cut = 5 + ID2*3;
end;
if n2=8 then do;
if n1=10 then cut = 0 + ID2;
if n1=20 then cut = 1 + ID2*2;
```

```
if n1=30 then cut = 3 + ID2*3;
end;

if&varerror= 0 then tau0 = .2;
if&varerror= 0 then tau1 = .05;
if&varerror= 1 then tau0 = 2;
if&varerror= 1 then tau1 = .5;

 IDlevel3=j(n1,1,ID3);
 IDlevel2=j(n1,1,ID2);
time=j(n1,1,0);
phase=j(n1,1,0);
do ii=1 to n1;
time[ii,1]=(ii)-1;
      if ii > cut then phase[ii,1]=1;
end;
rr=armasim({1,&phi},0,0,1,n1,0);
b=1;
c=0;
d=0;
*b=.90475830311225;
*c=.14721081863342;
*d=.02386092280190;
a=-1*c;
r=a+b#rr+c#rr##2+d#rr##3;
u0=repeat(rannor(0)*sqrt(tau0),n1); *error at level-2 intercept;
u1=repeat(rannor(0)*sqrt(tau0),n1); *error at level-2 treatment effect;
u2=repeat(rannor(0)*sqrt(tau1),n1); *error at level-3 intercept;
u3=repeat(rannor(0)*sqrt(tau1),n1); *error at level-3 treatment effect;
gamma000=&gamma;
gamma100=&gamma;
intercep=gamma000+u0+u2;
effect=gamma100+u1+u3;
y=intercep+(effect#phase)+r;

append;
end;
end;
close j1;

*+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
The following set of commands used PROC MIXED to estimate the
multilevel model.  This is done to create confidence intervals for the
fixed effects and variance components. For each run, the point
estimate, upper limit, and lower limit for the fixed effects and the
variance components, are written into an output data sets.
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++;

*setting the data sets that will contain the tables for the fixed
effects and variance components*;

data j2;
set j0;


data j3;
set j00;
```

```
*model commands and data set creation;

proc mixed data =j1 covtestcl;
class idlevel2 idlevel3;
model y = phase / s cl alpha = .05ddfm = kenwardroger;
randomint phase / sub = idlevel3;
randomint phase / sub = idlevel2 (idlevel3);
repeated / type = AR(1) sub = idlevel2 (idlevel3);
ods output solutionF=j2
  (keep = estimate lower upper);
ods output covparms=j3
   (keep = estimate lower upper);

data j4;
set j2;
w = estimate; output;
w = lower; output;
w = upper; output;
drop estimate lower upper;

proc transpose data = j4
out = j6
(rename = (col1=est_int col2=low_int col3=up_int
col4=est_pha col5=low_pha col6=up_pha));


data j5;
set j3;
w = estimate; output;
w = lower; output;
w = upper; output;
drop estimate lower upper;

proc transpose data = j5
out = j7
(rename = (col1=est_vc_int_lvl3 col2=low_vc_int_lvl3
col3=up_vc_int_lvl3
col4=est_vc_pha_lvl3 col5=low_vc_pha_lvl3 col6=up_vc_pha_lvl3
col7=est_vc_int_lvl2 col8=low_vc_int_lvl2 col9=up_vc_int_lvl2
col10=est_vc_pha_lvl2 col11=low_vc_pha_lvl2 col12=up_vc_pha_lvl2
col13=est_vc_ar col14=low_vc_ar col15=up_vc_ar
col16=est_vc_r col17=low_vc_r col18=up_vc_r));

*+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
The following statements merge the output data sets resulting with one
row of data containing the point estimates, lower limit, upper limit,
for each fixed effect and variance component. The data set is then
appended with a new row for each simulated data set.
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++;

data j8;
merge j6 j7;


data j9;
set j8;
```

```
counter = &i;

%if&i = 1%then%do;


data j10;
set j9;
%end;
%else%do;

data j10;
merge j10 j9;
by counter;
%end;

*DM 'LOG;*CLEAR';
*DM 'LISTING;*CLEAR';
%end;

*+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
The following set of commands creates a series of indicator variables
based on whether the fixed effect parameter and the variance component
parameter falls between the lower and upper limit.  It then computes
the width of the confidence interval.
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++;

data j11;
set j10;
if&varerror= 0 then tau0 = .2;
if&varerror= 0 then tau1 = .05;
if&varerror= 1 then tau0 = 2;
if&varerror= 1 then tau1 = .5;

cov_int=0;
if (low_int<= &gamma) & (&gamma <= up_int) then cov_int=1;
iflow_int=.thencov_int=.;

cov_pha=0;
if (low_pha<= &gamma) & (&gamma <= up_pha) then cov_pha=1;
iflow_pha=.thencov_pha=.;

cov_vc_int_lvl2=0;
if (low_vc_int_lvl2 <= tau0) & (tau0 <= up_vc_int_lvl2) then
cov_vc_int_lvl2=1;
if low_vc_int_lvl2=.then cov_vc_int_lvl2=.;

cov_vc_pha_lvl2=0;
if (low_vc_pha_lvl2 <= tau0) & (tau0 <= up_vc_pha_lvl2) then
cov_vc_pha_lvl2=1;
if low_vc_pha_lvl2=.then cov_vc_pha_lvl2=.;

cov_vc_int_lvl3=0;
if (low_vc_int_lvl3 <= tau1) & (tau1 <= up_vc_int_lvl3) then
cov_vc_int_lvl3=1;
if low_vc_int_lvl3=.then cov_vc_int_lvl3=.;
```

199

```
cov_vc_pha_lvl3=0;
if (low_vc_pha_lvl3 <= tau1) & (tau1 <= up_vc_pha_lvl3) then
cov_vc_pha_lvl3=1;
if low_vc_pha_lvl3=.then cov_vc_pha_lvl3=.;

cov_vc_ar=0;
if (low_vc_ar<= -1*&phi) & (-1*&phi <= up_vc_ar) then cov_vc_ar=1;
iflow_vc_ar=.thencov_vc_ar=.;

cov_vc_r=0;
if (low_vc_r<= 1) & (1<= up_vc_r) then cov_vc_r=1;

wid_int=up_int-low_int;

wid_pha=up_pha-low_pha;

wid_vc_int_lvl2=up_vc_int_lvl2-low_vc_int_lvl2;

wid_vc_pha_lvl2=up_vc_pha_lvl2-low_vc_pha_lvl2;

wid_vc_int_lvl3=up_vc_int_lvl3-low_vc_int_lvl3;

wid_vc_pha_lvl3=up_vc_pha_lvl3-low_vc_pha_lvl3;

wid_vc_ar=up_vc_ar-low_vc_ar;

wid_vc_r=up_vc_r-low_vc_r;



*++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Means are then calculated, giving estimates of bias in the fixed and
variance component effect estimates, the coverage probabilities for
each effect, and the average CI width.
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++;
proc means noprint data = j11;
varest_int
cov_int
wid_int

est_pha
cov_pha
wid_pha

est_vc_int_lvl2
cov_vc_int_lvl2
wid_vc_int_lvl2

    est_vc_pha_lvl2
cov_vc_pha_lvl2
wid_vc_pha_lvl2

est_vc_int_lvl3
cov_vc_int_lvl3
wid_vc_int_lvl3

    est_vc_pha_lvl3
```

```
cov_vc_pha_lvl3
wid_vc_pha_lvl3

est_vc_ar
cov_vc_ar
wid_vc_ar

     est_vc_r
cov_vc_r
wid_vc_r;

output out=j12
mean =      est_int
cov_int
wid_int

est_pha
cov_pha
wid_pha

est_vc_int_lvl2
cov_vc_int_lvl2
wid_vc_int_lvl2

     est_vc_pha_lvl2
cov_vc_pha_lvl2
wid_vc_pha_lvl2

est_vc_int_lvl3
cov_vc_int_lvl3
wid_vc_int_lvl3

     est_vc_pha_lvl3
cov_vc_pha_lvl3
wid_vc_pha_lvl3

est_vc_ar
cov_vc_ar
wid_vc_ar

     est_vc_r
cov_vc_r
wid_vc_r

     n = n_sims;
     ods listing;

%global _print_;
%let _print_ = on;

data j13;
set j12;

reps=&n;
Average_Series=&avgseries;
Average_Part=&avgpart;
Error_Variance=&varerror;
```

```
fixed=&gamma;
phi=&phi;
conv=n_sims/reps;


data j14;
set j13;
file print;
file'Y:\Documents\Dissertation\Results\DISS.txt' mod lrecl=400;
put @1(Average_Series)(1.0) @3(Average_Part)(1.0)
@5(Error_Variance)(1.0)
     @7 (fixed)(2.0) @10 (phi)(4.1)  @15 (conv)(6.4)
     @22 (est_intcov_intwid_int

est_phacov_phawid_pha

est_vc_int_lvl2 cov_vc_int_lvl2 wid_vc_int_lvl2

     est_vc_pha_lvl2 cov_vc_pha_lvl2 wid_vc_pha_lvl2

est_vc_int_lvl3 cov_vc_int_lvl3 wid_vc_int_lvl3

est_vc_pha_lvl3 cov_vc_pha_lvl3 wid_vc_pha_lvl3

est_vc_arcov_vc_arwid_vc_ar

     est_vc_rcov_vc_rwid_vc_r

     reps) (10.4);

run;
%mend;
*++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
input for the macro are:
n = 5000 (number of samples to generate)
n3= 10,30,or 80
avgseries = 0, 1, or 2 (0 = small, 1 = medium, 2 = large)
avgpart= 0 or 1 (0 = small and 1 = large)
varerror= 0 or 1 (0= most of the error at level-1 and 1= most of the
error at level-2)
gamma = 1 (fixed effects - intercept [gamma000], effect [gamma100])
phi = 0, -.2,or -.4 (produces positive autocorrelation)
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++;

%hlmsim(5000,10,0,0,0,1,-.2);

run;
```